Statistics 500=Psychology 611=Biostatistics 550
Introduction to Regression and Anova

Paul R. Rosenbaum
Professor, Statistics Department, Wharton School

Description

Statistics 500/Psychology 611 is a second course in statistics for PhD students in the social, biological and business sciences. It covers multiple linear regression and analysis of variance. Students should have taken an undergraduate course in statistics prior to Statistics 500.

Topics

1-Review of basic statistics.
2-Simple regression.
3-Multiple regression.
4-General linear hypothesis.
5-Woes of Regression Coefficients.
6-Transformations.
7-Polynomials.
8-Coded variables.
9-Diagnostics.
10-Variable selection.
11-One-way anova.
12-Two-way and factorial anova.

How do I get R for free? http://cran.r-project.org/

Final exam date: http://www.upenn.edu/registrar/
Holidays, breaks, last class: http://www.upenn.edu/almanac/3yearcal.html

My web page: http://www-stat.wharton.upenn.edu/~rosenbap/index.html
Email: rosenbaum@wharton.upenn.edu
Phone: 215-898-3120
Office: 473 Huntsman Hall (in the tower, 4th floor)
Office Hours: Tuesday 1:30-2:30 and by appointment.

The bulk pack and course data in R are on my web page.

Overview

Review of Basic Statistics

Descriptive statistics, graphs, probability, confidence intervals, hypothesis tests.

Simple Regression

Simple regression uses a line with one predictor to predict one outcome.

Multiple Regression

Multiple regression uses several predictors in a linear way to predict one outcome.

General Linear Hypothesis

The general linear hypothesis asks whether several variables may be dropped from a multiple regression.

Woes of Regression Coefficients

Discussion of the difficulties of interpreting regression coefficients and what you can do.

Transformations

A simple way to fit curves or nonlinear models: transform the variables.

Polynomials

Another way to fit curves: include quadratics and interactions.

Coded Variables

Using nominal data (NY vs Philly vs LA) as predictors in regression.

Diagnostics

How to find problems in your regression model: residual, leverage and influence.

Variable Selection

Picking which predictors to use when many variables are available.

One-Way Anova

Simplest analysis of variance: Do several groups differ, and if so, how?

Two-Way Anova

Study two sources of variation at the same time.

Factorial Anova

Study two or more treatments at once, including their interactions.

Common Questions

Statistics Department Courses (times, rooms)
http://www.upenn.edu/registrar/roster/stat.html

Final Exams (dates, rules)
http://www.upenn.edu/registrar/finals/spring05_index.html

Computing and related help at Wharton
http://inside.wharton.upenn.edu/

Statistical Computing in the Psychology Department
http://www.psych.upenn.edu

When does the the course start? When does it end? Holidays?
http://www.upenn.edu/almanac/3yearcal.html

Does anybody have any record of this?
http://www.upenn.edu/registrar/

Huntsman Hall
http://www.facilities.upenn.edu/mapsBldgs/view_bldg.php3?id=146
http://www.facilities.upenn.edu/mapsBldgs/view_map.php3?id=393

Suggested reading

Box, G. E. P. (1966) Use and Abuse of Regression, Technometrics, 8, 625-629.
http://www.jstor.org/ or
http://www.jstor.org/stable/1266635?&Search=yes&term=abuse&term=box&list=hide&searchUri=%2Faction%2FdoAdvancedSearch%3Fq0%3Dbox%26f0%3Dau%26c0%3DAND%26q1%3Dabuse%26f1%3Dti%26c1%3DAND%26q2%3D%26f2%3Dall%26c2%3DAND%26q3%3D%26f3%3Dall%26wc%3Don%26sd%3D%26ed%3D%26la%3D%26jo%3D%26dc.Statistics%3DStatistics%26Search%3DSearch&item=1&ttl=1&returnArticleService=showArticle

Helpful articles from JSTOR http://www.jstor.org/

1. The Analysis of Repeated Measures: A Practical Review with Examples B. S. Everitt *The Statistician*, Vol. 44, No. 1. (1995), pp. 113-135.
2. The hat matrix in regression and anova. D. Hoaglin and R. Welsh, *American Statistician*, Vol 32, (1978), pp. 17-22.
3. The Use of Nonparametric Methods in the Statistical Analysis of the Two-Period Change-Over Design Gary G. Koch *Biometrics*, Vol. 28, No. 2. (Jun., 1972), pp. 577-584.

Some Web Addresses

Web page for Sheather's text
http://www.stat.tamu.edu/~sheather/

Amazon for Sheather's text (required)
http://www.amazon.com/Modern-Approach-Regression-Springer-Statistics/dp/0387096078/ref=tmm_hrd_title_0/186-7302133-0606755?ie=UTF8&qid=1315493088&sr=1-1

Alternative text used several years ago (optional alternative, not suggested)
http://www.amazon.com/Applied-Regression-Analysis-Multivariable-Methods/dp/0495384968/ref=sr_1_1?s=books&ie=UTF8&qid=1315493363&sr=1-1

Good supplement about R (optional, suggested)
http://www.amazon.com/Data-Analysis-Graphics-Using-Example-Based/dp/0521762936/ref=sr_1_1?s=books&ie=UTF8&qid=1315493138&sr=1-1

Review basic statistics, learn basic R (optional, use if you need it)
http://www.amazon.com/Introductory-Statistics-R-Computing/dp/0387790535/ref=sr_1_1?s=books&ie=UTF8&qid=1315493184&sr=1-1

Excellent text, alternative to Sheather, more difficult than Sheather
http://www.amazon.com/Applied-Regression-Analysis-Probability-Statistics/dp/0471170828/ref=sr_1_1?s=books&ie=UTF8&qid=1315493220&sr=1-1

Good text, alternative/supplement to Sheather, easier than Sheather
http://www.amazon.com/Regression-Analysis-Example-Probability-Statistics/dp/0471746967/ref=tmm_hrd_title_0?ie=UTF8&qid=1315493316&sr=1-1

Free R manuals at R home page. Start with "An Introduction to R"
http://cran.r-project.org/
--> Manuals --> An Introduction to R
--> Search --> Paradis --> R for Beginners

My web page (bulk pack, course data)
http://www-stat.wharton.upenn.edu/~rosenbap/index.html

Computing

How do I get R for free? http://cran.r-project.org/

After you have installed R, you can get the **course data** in the R-workspace on my web page: http://www-stat.wharton.upenn.edu/~rosenbap/index.html

I will probably add things to the R-workspace during the semester. So you will have to go back to my web page to **get the latest version**.

A common problem: You go to my web page and download the latest R-workspace, but it looks the same as the one you had before – the new stuff isn't there. This happens when your web browser thinks it has downloaded the file before and will save you time by not downloading it again. Bad web browser. You need to clear the cache; then it will get the new version.

Most people find an R book helpful. I recommend Maindonald and Braun, *Data Analysis and Graphics Using R*, published by Cambridge. A more basic book is Dalgaard, *Introductory Statistics with R*, published by Springer.

At http://cran.r-project.org/, click on **manuals** to get free documentation. "An Introduction to R" is there, and it is useful. When you get good at R, do a search at the site for Paradis' "R for Beginners," which is very helpful, but not for beginners.

Textbook

My sense is that students need a textbook, not just the lectures and the bulk pack.

The 'required' textbook for the course is Sheather (2009) *A Modern Approach to Regression with R*, NY: Springer. There is a little matrix algebra in the book, but there is none in the course. Sheather replaces the old text, Kleinbaum, Kupper, Muller and Nizam, *Applied Regression and other Multivariable Methods*, largely because this book has become very expensive. An old used edition of Kleinbaum is a possible alternative to Sheather – it's up to you. Kleinbaum does more with anova for experiments. A book review by Gudmund R. Iversen of Swathmore College is available at: http://www.jstor.org/stable/2289682?&Search=yes&term=kleinbaum&term=kupper&list=hide&searchUri=%2Faction%2FdoAdvancedSearch%3Fq0%3Dkleinbaum%26f0%3Dau%26c0%3DAND%26q1%3Dkupper%26f1%3Dau%26c1%3DAND%26q2%3D%26f2%3Dall%26c2%3DAND%26q3%3D%26f3%3Dall%26wc%3Don%26re%3Don%26sd%3D%26ed%3D%26la%3D%26jo%3D%26dc.Statistics%3DStatistics%26Search%3DSearch&item=6&ttl=7&returnArticleService=showArticle

Some students might prefer one of the textbooks below, and they are fine substitutes.

If you would prefer an easier, less technical textbook, you might consider *Regression by Example* by Chatterjee and Hadi. The book has a nice chapter on transformations, but it barely covers anova. An earlier book, now out of print, with the same title by Chatterjee and Price is very similar, and probably available inexpensively used.
http://www.amazon.com/Regression-Analysis-Example-Probability-Statistics/dp/0471746967/ref=sr_1_2?ie=UTF8&s=books&qid=1252524629&sr=1-2

If you know matrix algebra, you might prefer the text *Applied Regression Analysis* by Draper and Smith. It is only slightly more difficult than Kleinbaum, and you can read around the matrix algebra.
http://www.amazon.com/Applied-Regression-Analysis-Probability-Statistics/dp/0471170828/ref=sr_1_1?ie=UTF8&s=books&qid=1252524403&sr=1-1

If you use R, then as noted previously, I recommend the additional text Maindonald and Braun, *Data Analysis and Graphics Using R*, published by Cambridge. It is in its third edition, which is a tad more up to date than the first or second editions, but you might prefer an inexpensive used earlier edition if you can find one.

Graded Work

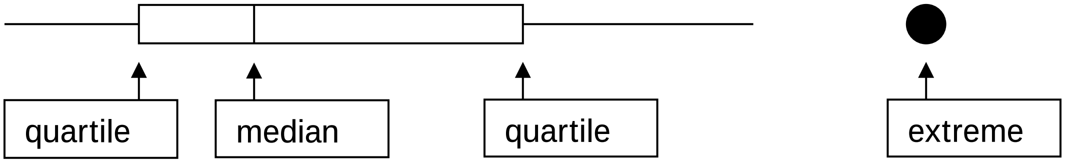
Your grade is based on three exams. Copies of old exams are at the end of this bulkpack. The first two exams are take-homes in which you do a data-analysis project. They are exams, so you do the work by yourself. The first exam covers the basics of multiple regression. The second exam covers diagnostics, model building and variable selection. The final exam is sometimes in-class, sometimes take home. The date of the final exam is determined by the registrar – see the page above for Common Questions. The decision about whether the final is in-class or take-home will be made after the first take-home is graded. That will be in the middle of the semester. If you need to make travel arrangements before the middle of the semester, you will need to plan around an in-class final.

The best way to learn the material is to practice using the old exams. There are three graded exams. If for each graded exam, you did two practice exams, then you would do nine exams in total, which means doing nine data analysis projects. With nine projects behind you, regression will start to be familiar.

Review of Basic Statistics – Some Statistics

- The review of basic statistics is a quick review of ideas from your first course in statistics.
- n measurements: $X_1, X_2, \ldots, X_n$
- **mean** (or average): $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$
- **order statistics** (or data sorted from smallest to largest): Sort $X_1, X_2, \ldots, X_n$ placing the smallest first, the largest last, and write $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$, so the smallest value is the first order statistic, $X_{(1)}$, and the largest is the nth order statistic, $X_{(n)}$. If there are n=4 observations, with values $X_1 = 5, X_2 = 4, X_3 = 9, X_4 = 5$, then the n=4 order statistics are $X_{(1)} = 4, X_{(2)} = 5, X_{(3)} = 5, X_{(4)} = 9$.
- **median** (or middle value): If n is odd, the median is the middle order statistic – e.g., $X_{(3)}$ if n=5. If n is even, there is no middle order statistic, and the median is the average of the two order statistics closest to the middle – e.g., $\frac{X_{(2)} + X_{(3)}}{2}$ if n=4. Depth of median is $\frac{n+1}{2}$ where a “half” tells you to average two order statistics – for n=5, $\frac{n+1}{2} = \frac{5+1}{2} = 3$, so the median is $X_{(3)}$, but for n=4, $\frac{n+1}{2} = \frac{4+1}{2} = 2.5$, so the median is $\frac{X_{(2)} + X_{(3)}}{2}$. The median cuts the data in half – half above, half below.
- **quartiles**: Cut the data in quarters – a quarter above the upper quartile, a quarter below the lower quartile, a quarter between the lower quartile and the median, a quarter between the median and the upper quartile. The **interquartile range** is the upper quartile minus the lower quartile.

- **boxplot**: Plots median and quartiles as a box, calls attention to extreme observations.



- **sample standard deviation**: square root of the typical squared deviation from the mean, sorta,

$$s = \sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}}$$

  however, you don't have to remember this ugly formula.
- **location**: if I add a constant to every data value, a measure of location goes up by the addition of that constant.
- **scale**: if I multiply every data value by a constant, a measure of scale is multiplied by that constant, but a measure of scale does not change when I add a constant to every data value.

Check your understanding: What happens to the mean if I drag the biggest data value to infinity? What happens to the median? To a quartile? To the interquartile range? To the standard deviation? Which of the following are measures of location, of scale or neither: median, quartile, interquartile range, mean, standard deviation? In a boxplot, what would it mean if the median is closer to the lower quartile than to the upper quartile?

Topic: Review of Basic Statistics – Probability

- **probability space**: the set of everything that can happen, Ω. Flip two coins, dime and quarter, and the sample space is Ω = {HH, HT, TH, TT} where HT means "head on dime, tail on quarter", etc.
- **probability**: each element of the sample space has a probability attached, where each probability is between 0 and 1 and the total probability over the sample space is 1. If I flip two fair coins: prob(HH) = prob(HT) = prob(TH) = prob(TT) = ¼.
- **random variable**: a rule **X** that assigns a number to each element of a sample space. Flip to coins, and the number of heads is a random variable: it assigns the number **X**=2 to HH, the number **X**=1 to both HT and TH, and the number **X**=0 to TT.
- **distribution of a random variable**: The chance the random variable **X** takes on each possible value, x, written prob(**X**=x). Example: flip two fair coins, and let **X** be the number of heads; then prob(**X**=2) = ¼, prob(**X**=1) = ½, prob(**X**=0) = ¼.
- **cumulative distribution of a random variable**: The chance the random variable **X** is less than or equal to each possible value, x, written prob(**X**≤x). Example: flip two fair coins, and let **X** be the number of heads; then prob(**X**≤ 0) = ¼, prob(**X**≤1) = ¾, prob(**X**≤2) = 1. Tables at the back of statistics books are often cumulative distributions.
- **independence of random variables**: Captures the idea that two random variables are unrelated, that neither predicts the other. The formal definition which follows is not intuitive – you get to like it by trying many intuitive examples, like unrelated coins and taped coins, and finding the definition always works. Two random variables, **X** and **Y**, are independent if the chance that simultaneously **X**=x and **Y**=y can be found by multiplying the separate probabilities

  prob(**X**=x and **Y**=y) = prob(**X**=x) prob(**Y**=y) for every choice of x,y**.**

Check your understanding: Can you tell exactly what happened in the sample space from the value of a random variable? Pick one: Always, sometimes, never. For people, do you think **X**=height and **Y**=weight are independent? For undergraduates, might **X**=age and **Y**=gender (1=female, 2=male) be independent? If I flip two fair coins, a dime and a quarter, so that prob(HH) = prob(HT) = prob(TH) = prob(TT) = ¼, then is it true or false that getting a head on the dime is independent of getting a head on the quarter?

Topic: Review of Basics – Expectation and Variance

- **Expectation**: The expectation of a random variable **X** is the sum of its possible values weighted by their probabilities,
$$E(\mathbf{X}) = \sum_x x \cdot prob(\mathbf{X} = x)$$
- **Example**: I flip two fair coins, getting **X**=0 heads with probability ¼, **X**=1 head with probability ½, and **X**=2 heads with probability ¼; then the expected number of heads is $E(\mathbf{X}) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$, so I expect 1 head when I flip two fair coins. Might actually get 0 heads, might get 2 heads, but 1 head is what is typical, or expected, on average.
- **Variance and Standard Deviation**: The standard deviation of a random variable **X** measures how far **X** typically is from its expectation *E*(**X**). Being too high is as bad as being too low – we care about errors, and don't care about their signs. So we look at the squared difference between **X** and *E*(**X**), namely $\mathbf{D} = \{\mathbf{X} - E(\mathbf{X})\}^2$, which is, itself, a random variable. The variance of **X** is the expected value of **D** and the standard deviation is the square root of the variance, $\text{var}(\mathbf{X}) = E(\mathbf{D})$ and $st.dev.(\mathbf{X}) = \sqrt{\text{var}(\mathbf{X})}$.
- **Example**: I independently flip two fair coins, getting **X**=0 heads with probability ¼, **X**=1 head with probability ½, and **X**=2 heads with probability ¼. Then *E*(**X**)=1, as noted above. So $\mathbf{D} = \{\mathbf{X} - E(\mathbf{X})\}^2$ takes the value **D** =

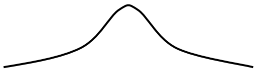
$(0-1)^2=1$ with probability ¼, the value **D** = $(1-1)^2=0$ with probability ½, and the value **D** = $(2-1)^2=1$ with probability ¼. The variance of **X** is the expected value of **D** namely: var(**X**) = $E(\mathbf{D})=1\cdot\frac{1}{4}+0\cdot\frac{1}{2}+1\cdot\frac{1}{4}=\frac{1}{2}$. So the standard deviaiton is $st.dev.(\mathbf{X})=\sqrt{\mathrm{var}(\mathbf{X})}=\sqrt{\frac{1}{2}}=0.707$. So when I flip two fair coins, I expect one head, but often I get 0 or 2 heads instead, and the typical deviation from what I expect is 0.707 heads. This 0.707 reflects the fact that I get exactly what I expect, namely 1 head, half the time, but I get 1 more than I expect a quarter of the time, and one less than I expect a quarter of the time.

Check your understanding: If a random variance has zero variance, how often does it differ from its expectation? Consider the height **X** of male adults in the US. What is a reasonable number for *E*(**X**)? Pick one: 4 feet, 5'9", 7 feet. What is a reasonable number for *st.dev.*(**X**)? Pick one: 1 inch, 4 inches, 3 feet. If I independently flip three fair coins, what is the expected number of heads? What is the standard deviation?

Topic: Review of Basics – Normal Distribution

- **Continuous random variable**: A continuous random variable can take values with any number of decimals, like 1.2361248912. Weight measured perfectly, with all the decimals and no rounding, is a continuous random variable. Because it can take so many different values, each value winds up having probability zero. If I ask you to guess someone's weight, not approximately to the nearest millionth of a gram, but rather exactly to all the decimals, there is no way you can guess correctly – each value with all the decimals has probability zero. But for an interval, say the nearest kilogram, there is a nonzero chance you can guess correctly. This idea is captured in by the density function.

- **Density Functions**: A density function defines probability for a continuous random variable. It attaches zero probability to every number, but positive probability to ranges (e.g., nearest kilogram). The probability that the random variable **X** takes values between 3.9 and 6.2 is the area under the density function between 3.9 and 6.2. The total area under the density function is 1.
- **Normal density**: The Normal density is the familiar “bell shaped curve”.



The standard Normal distribution has expectation zero, variance 1, standard deviation 1 = $\sqrt{1}$. About 2/3 of the area under the Normal density is between –1 and 1, so the probability that a standard Normal random variable takes values between –1 and 1 is about 2/3. About 95% of the area under the Normal density is between –2 and 2, so the probability that a standard Normal random variable takes values between –2 and 2 is about .95. (To be more precise, there is a 95% chance that a standard Normal random variable will be between –1.96 and 1.96.) If **X** is a standard Normal random variable, and μ and $\sigma > 0$ are two numbers, then $\mathbf{Y} = \mu + \sigma\mathbf{X}$ has the Normal distribution with expectation μ, variance σ^2 and standard deviation σ, which we write N(μ, σ^2). For example, $\mathbf{Y} = 3 + 2\mathbf{X}$ has expectation 3, variance 4, standard deviation 2, and is N(3,4).

- **Normal Plot**: To check whether or not data, $X_1, \ldots, X_n$ look like they came from a Normal distribution, we do a Normal plot. We get the order statistics – just the data sorted into order – or $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ and plot this ordered data against what ordered data from a standard Normal distribution should look like. The computer takes care of the details. A straight line in a

Normal plot means the data look Normal. A straight line with a couple of strange points off the lines suggests a Normal with a couple of strange points (called outliers). Outliers are extremely rare if the data are truly Normal, but real data often exhibit outliers. A curve suggest data that are not Normal. Real data wiggle, so nothing is ever perfectly straight. In time, you develop an eye for Normal plots, and can distinguish wiggles from data that are not Normal.

Topic: Review of Basics – Confidence Intervals

- Let $X_1, \ldots, X_n$ be n independent observations from a Normal distribution with expectation μ and variance σ^2. A compact way of writing this is to say $X_1, \ldots, X_n$ are iid from N(μ, σ^2). Here, iid means independent and identically distributed, that is, unrelated to each other and all having the same distribution.
- How do we know $X_1, \ldots, X_n$ are iid from N(μ, σ^2)? We don't! But we check as best we can. We do a boxplot to check on the shape of the distribution. We do a Normal plot to see if the distribution looks Normal. Checking independence is harder, and we don't do it as well as we would like. We do look to see if measurements from related people look more similar than measurements from unrelated people. This would indicate a violation of independence. We do look to see if measurements taken close together in time are more similar than measurements taken far apart in time. This would indicate a violation of independence. Remember that statistical methods come with a warrantee of good performance if certain assumptions are true, assumptions like $X_1, \ldots, X_n$ are iid from N(μ, σ^2). We check the assumptions to make sure we get the promised good performance of statistical methods. Using statistical methods when the assumptions are not

true is like putting your CD player in washing machine – it voids the warrantee.

- To begin again, having checked every way we can, finding no problems, assume $X_1, \ldots, X_n$ are iid from N(μ, σ^2). We want to estimate the expectation μ. We want an interval that in most studies winds up covering the true value of μ. Typically we want an interval that covers μ in 95% of studies, or a **95% confidence interval**. Notice that the promise is about what happens in most studies, not what happened in the current study. If you use the interval in thousands of unrelated studies, it covers μ in 95% of these studies and misses in 5%. You cannot tell from your data whether this current study is one of the 95% or one of the 5%. All you can say is the interval usually works, so I have confidence in it.
- If $X_1, \ldots, X_n$ are iid from N(μ, σ^2), then the confidence interval uses the sample mean, $\overline{X}$, the sample standard deviation, *s*, the sample size, *n*, and a critical value obtained from the t-distribution with *n-1* degrees of freedom, namely the value, $t_{0.025}$, such that the chance a random variable with a t-distribution is above $t_{0.025}$ is 0.025. If *n* is not very small, say n>10, then $t_{0.025}$ is near 2. The 95% confidence interval is:

$$\overline{X} \pm \text{(allowance for error)} \quad = \quad \overline{X} \pm \frac{t_{0.025} \cdot s}{\sqrt{n}}$$

Topic: Review of Basics – Hypothesis Tests

- **Null Hypothesis**: Let $X_1, \ldots, X_n$ be n independent observations from a Normal distribution with expectation μ and variance σ^2. We have a particular value of μ in mind, say μ_0, and we want to ask if the data contradict this value. It means something special to us if μ_0 is the correct value – perhaps it means the treatment has no effect, so the treatment should be discarded. We wish to test the null hypothesis, $H_0\colon \mu = \mu_0$. Is the null hypothesis plausible? Or do the data force us to abandon the null hypothesis?
- **Logic of Hypothesis Tests**: A hypothesis test has a long-winded logic, but not an unreasonable one. We say: Suppose, just for the sake of argument, not because we believe it, that the null hypothesis is true. As is always true when we suppose something for the sake of argument, what we mean is: Let's suppose it and see if what follows logically from supposing it is believable. If not, we doubt our supposition. So suppose μ_0 is the true value after all. Is the data we got, namely $X_1, \ldots, X_n$, the sort of data you would usually see if the null hypothesis were true? If it is, if $X_1, \ldots, X_n$ are a common sort of data when the null hypothesis is true, then the null hypothesis looks sorta ok, and we *accept* it. Otherwise, if there is no way in the world you'd ever see data anything remotely like our data, $X_1, \ldots, X_n$, if the null hypothesis is true, then we can't really believe the null hypothesis having seen $X_1, \ldots, X_n$, and we *reject* it. So the basic question is: Is data like the data we got commonly seen when the null hypothesis is true? If not, the null hypothesis has gotta go.
- **P-values or significance levels**: We measure whether the data are commonly seen when the null hypothesis is true using something called the P-value or significance level. Supposing the null hypothesis to be true, the P-value is the chance of data at least as inconsistent with the null hypothesis as

the observed data. If the P-value is ½, then half the time you get data as or more inconsistent with the null hypothesis as the observed data – it happens half the time by chance – so there is no reason to doubt the null hypothesis. But if the P-value is 0.000001, then data like ours, or data more extreme than ours, would happen only one time in a million by chance if the null hypothesis were true, so you gotta being having some doubts about this null hypothesis.

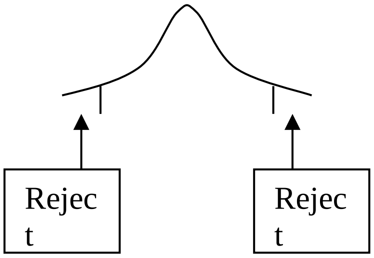
- **The magic 0.05 level:** A convention is that we "reject" the null hypothesis when the P-value is less than 0.05, and in this case we say we are testing at **level** 0.05. Scientific journals and law courts often take this convention seriously. It is, however, only a convention. In particular, sensible people realize that a P-value of 0.049 is not very different from a P-value of 0.051, and both are very different from P-values of 0.00001 and 0.3. It is best to report the P-value itself, rather than just saying the null hypothesis was rejected or accepted.
- **Example**: You are playing 5-card stud poker and the dealer sits down and gets 3 royal straight flushes in a row, winning each time. The null hypothesis is that this is a fair poker game and the dealer is not cheating. Now, there are or 2,598,960 five-card stud poker hands, and 4 of these are royal straight flushes, so the chance of a royal straight flush in a fair game is $\frac{4}{2{,}598{,}960} = 0.000001539$. In a fair game, the chance of three royal straight flushes in a row is 0.000001539x0.000001539x0.000001539 = 3.6×10^{-18}. (Why do we multiply probabilities here?) Assuming the null hypothesis, for the sake of argument, that is assuming he is not cheating, the chance he will get three royal straight flushes in a row is very, very small – that is the P-value or significance level. The data we see is highly improbable if the null hypothesis were true, so we doubt it is true. Either the dealer got very, very lucky, or he cheated. This is the logic of all hypothesis tests.

- **One sample t-test**: Let $X_1, \ldots, X_n$ be n independent observations from a Normal distribution with expectation μ and variance σ^2. We wish to test the null hypothesis, $H_0: \mu = \mu_0$. We do this using the one-sample t-test:

$$t = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{s}$$

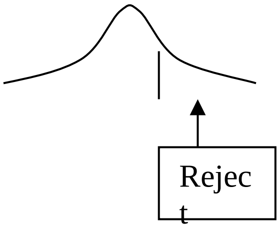
  looking this up in tables of the t-distribution with *n-1* degrees of freedom to get the P-value.

- **One-sided vs Two-sided tests**: In a two-sided test, we don't care whether $\overline{X}$ is bigger than or smaller than μ_0, so we reject at the 5% level when |t| is one of the 5% largest values of |t|. This means we reject for 2.5% of t's that are very positive and 2.5% of t's that are very negative:



In a two sided test we reject when t is big positive or big negative. If we reject when the P-value is less than 0.05, then each tail has probability 0.025.

In a one sided test, we do care, and only want to reject when $\overline{X}$ is on one particular side of μ_0, say when $\overline{X}$ is bigger than μ_0, so we reject at the 5% level when t is one of the 5% largest values of t. This means we reject for the 5% of t's that are very positive:



In a one sided test we reject on just one side, say big positive. If we reject when the P-value is less than 0.05, the tail on the right has probability 0.05.

- **Should I do a one-sided or a two-sided test**: Scientists mostly report two-sided tests.

Obtaining a Confidence Interval for a Parameter by Inverting a Test

1. We have a valid way of testing the hypothesis $H_0 : \theta = \theta_0$ that works for any real number θ_0. For example, we can test $H_0 : \theta = 0$ and $H_0 : \theta = 1$ and $H_0 : \theta = 1.263$, etc.

2. A valid level α test of $H_0 : \theta = \theta_0$ falsely rejects H_0 when it is true with probability at most α. That is the definition of "valid" in the phrase "valid test". The popular α is $\alpha = 0.05$, but that is merely a convention.

3. Suppose we test every possible value θ_0 at level $\alpha = 0.05$, keeping the values we do not reject in a set $\mathcal{C}$. So θ_0 is in $\mathcal{C}$, or equivalently $\theta_0 \in \mathcal{C}$, if we did not reject θ_0 when we tested it because its P-value was > 0.05.

4. The set $\mathcal{C}$ is a random set. We computed it from the data, and the data are random, so $\mathcal{C}$ is random.

5. There is one true value, θ^*, of θ. This true value θ^* is a number, not a random variable, a number we do not know but a number nonetheless. What is the chance that the random set $\mathcal{C}$ fails to cover the fixed number θ^*?

6. We know that we will, sooner or later, test the true value, $H_0 : \theta = \theta^*$, because we test every value. We are happy to reject false values of θ — after all, they are false — but when we test the one true value, θ^*, we don't want to reject it.

7. When we test the true value, $H_0 : \theta = \theta^*$, the chance that we falsely reject it is at most $\alpha = 0.05$. That is, again, what it means to use a valid test.

8. Therefore, the probability that the random interval $\mathcal{C}$ fails to cover θ^* is at most $\alpha = 0.05$. We call $\mathcal{C}$ a 95% confidence interval.

REGRESSION ASSUMPTIONS

Assumption	If untrue:	How to detect:
Independent errors	95% confidence intervals may cover much less than 95% of the time. Tests that reject with $p<0.05$ may reject true hypotheses more than 5% of the time. You may think you have much more information than you do.	Often hard to detect. Questions to ask yourself: (i) Are the observations clustered into groups, such as several measurements on the same person? (ii) Are observations repeated over time?
Normal errors	Thick tails and outliers may distort estimates, and they may inflate the estimated error variance, so that confidence intervals are too long, and hypothesis tests rarely reject false hypotheses.	Do a Normal quantile plot. This is the one use of the Normal quantile plot. A more or less straight line in the plot suggests the data are approximately Normal.
Errors have constant variance.	Least squares gives equal weight to all observations, but if some observations are much more stable than others, it is not sensible to give equal weight to all observations.	Plot the residuals against the predicted values. A fan shape in the plot – narrow on one end, wide on the other – suggests unequal variances. Can also plot residuals against individual x's.
Model is linear.	Linear model may not fit, or may give the wrong interpretation of the data.	Plot the residuals against the predicted values. Curves, such as a U-shape, suggest the relationship is not linear. Can also plot residuals against individual x's.

Statistics 500: Basic Statistics Review

- **Reading**: In Kleinbaum, read chapter 3.
- **Practice**: The blood pressure data we discussed in class is given below. It is from MacGregor, et. al. (1979) British Medical Journal, 2, 1106-9. It is the change in systolic blood pressure two hours after taking Captopril, in mm Hg, after-before, so a negative number means a decline in blood pressure. Use JMP or another package to do a Normal plot, a boxplot and a t-test. Think about how you would describe what you see.

Patient #	Change in bp
1	-9
2	-4
3	-21
4	-3
5	-20
6	-31
7	-17
8	-26
9	-26
10	-10
11	-23
12	-33
13	-19
14	-19
15	-23

Homework: The following data are from Kaneto, Kosaka and Nakao (1969) Endocrinology, 80, 530-536. It is an experiment on 7 dogs. Question is whether stimulation of the vagus nerve increases levels of immunoreactive insulin in the blood. Two measurements were taken on each dog, one before, one five minutes after stimulation. The measurements are blood lead levels of immunoreative insulin ($\mu U / ml$).

Dog	Before	After
1	350	480
2	200	130
3	240	250
4	290	310
5	90	280
6	370	1450
7	240	280

Do an appropriate analysis.

Topic: Simple Regression

- **Simple regression**: Fitting a line a response Y using one predictor X.
- **Data**: 48 contiguous states in 1972, *i=1,...,48.* Y = FUEL = motor fuel consumption per person in gallons per person. X = TAX = motor fuel tax rate in cents per gallon.
- **First thing you do**: Plot the data.
- **Least squares**: Fit the line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ by minimizing the sum of the squares of the residuals $Y_i - \hat{Y}_i$ around the line, $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$.
- **Plot the residuals**: After you fit a line, you plot the residuals, $Y_i - \hat{Y}_i$. They tell you where and how the line fits poorly. The minimum is: (i) a boxplot of residuals, (ii) a Normal plot of residuals, (iii) a plot of residuals vs predicted values, $Y_i - \hat{Y}_i$ vs $\hat{Y}_i$.
- **Statistical Model**: The statistical model says:

  $Y_i = \alpha + \beta X_i + \varepsilon_i$ where the ε_i are iid $N(0, \sigma^2)$,

  so the Y's were generated by a true line, $\alpha + \beta X_i$, which we do not know, plus errors ε_i that are independent of each other and Normal with mean zero and constant variance σ^2. We use the residual plots to check whether the model is a reasonable description of the data. The line fitted by least squares, $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, is our estimate of the true line $\alpha + \beta X_i$.
- **Properties of least squares estimates**: The least squares estimators are great estimators – the best there are – when the model is correct, and not so great when the model is wrong. Checking the model is checking whether we are getting good estimates. When the model is true, least squares estimates are **unbiased**, that is, correct in expectation or on average, and they have **minimum variance** among all unbiased estimates, so they are the most stable, most accurate unbiased estimates (but only if the model is correct!).

They are not robust to outliers – one weird observation can move the fitted line anywhere it wants.

- **Basic Regression Output**

Variable	Estimated Coefficient	Estimated Standard Error of Estimated Coefficient	t-ratio
Constant	$\hat{\alpha}$	se($\hat{\alpha}$)	$\frac{\hat{\alpha}}{se(\hat{\alpha})}$
X	$\hat{\beta}$	se($\hat{\beta}$)	$\frac{\hat{\beta}}{se(\hat{\beta})}$

- **Hypothesis tests**: Use the t-ratio to test the null hypothesis $H_0: \beta = 0$. Under the model, the hypothesis $H_0: \beta = 0$ implies X and Y are unrelated.
- **Confidence intervals**: Under the model, a 95% confidence interval for β is:

$$estimate \pm allowance \quad = \quad \hat{\beta} \pm t_{0.025} \cdot se(\hat{\beta})$$

where $t_{0.025}$ is the upper 2.5% point of the t-distribution with n-2 degrees of freedom. When n-2 is not small, the interval is almost (but not quite) $\hat{\beta} \pm 2 \cdot se(\hat{\beta})$.

- **Points on a line vs Predictions**: Two problems look almost the same, but really are very different. One asks: Where is the line at X=8.5? That is, what is $\alpha + \beta 8.5$? That problem gets easier as I collect more data and learn where the line really is. The other asks: Where will a new observation on Y be if X=8.5? That is, what is $\alpha + \beta 8.5 + \varepsilon_{new}$? That problem always stays pretty hard, no matter how much data I collect, because I can't predict the new error, ε_{new}, for this new observation no matter how well I know where the line is. Important thing is to make sure you know which answer you want and to use the right method for that answer. They look similar, but they're not.

- **Regression Anova Table**: Partitions the total variation (or sum of squares) in the data about the mean, namely $\sum_{i=1}^{n}(Y_i-\overline{Y})^2$ into two parts that add back to the total, namely the variation fitted by the regression, $\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$, and the variation in the residuals, $\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$. Degrees of freedom measure keep track of how many distinct numbers are really being described by a sum of squares. In simple regression, the variation fitted by the regression is just fitted by the slope, $\hat{\beta}$, which is just one number, so this sum of squares has 1 degree of freedom. A mean square is the ratio $\frac{\text{sum of squares}}{\text{degrees of freedom}}$. The F-ratio is the ratio of two mean squares, a signal to noise ratio. The F-ratio is used to test that all the slopes are zero.
- **Simple correlation**: If the data fall perfectly on a line tilted up, the correlation r is 1. If the data fall perfectly on a line tilted down, the correlation r is –1. If a line is not useful for predicting Y from X, the correlation r is 0. Correlation is always between –1 and 1. The correlation between Y and X is the regression coefficient of standardized Y on standardized X, that is, the regression of $\frac{Y_i-\overline{Y}}{st.dev(Y)}$ on $\frac{X_i-\overline{X}}{st.dev(X)}$. In simple, one-predictor regression, the square of the correlation, r^2, is the percent of variation fitted by the regression, so it summarizes the anova table. Correlation discards the units of measurement, which limits its usefulness.

Homework: Vocabulary Data

Homework: The following data are from M. E. Smith, (1926), "An investigation of the development of the sentence and the extent of vocabulary in young children." It relates the X=age of children in years to their Y=vocabulary size in words. I would like you to do a regression of Y and X, look closely at what you've done, and comment on what it all means. You should turn in (1) one paragraph of text, (2) linear regression output, (3) at most two plots you find interesting and helpful in thinking about what is special about these data. This is real data, so it is not a "trick question", but it does require some real thought about what makes sense and what is happening.

X=age	**Y=vocabulary**
0.67	0
0.83	1
1	3
1.25	19
1.5	22
1.75	118
2	272
2.5	446
3	896
3.5	1,222
4	1,540
4.5	1,870
5	2,072
5.5	2,289
6	2,562

Topic: Multiple Regression

- **Multiple regression**: Uses several predictor variables $X_1, X_2, \ldots X_k$ to fit a single response variable Y.
- **FUEL DATA**: Trying to predict Y = FUEL from X_1 = TAX and a second predictor, X_2 = LICENSES.
- **Least squares fit**: Multiple regression fits a plane $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$ making the residuals $Y_i - \hat{Y}_i$ small, in the sense that the sum of the squares of the residuals, namely, $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, is minimized.
- **Multiple Correlation**: The multiple correlation, R, is the ordinary correlation between the observed Y_i and the fitted $\hat{Y}_i$. The square of the multiple correlation, R^2, is the percent of variation fitted by the regression, that is, regression sum of squares in the ANOVA table divided by the total sum of squares of Y around its mean.
- **Fit vs Prediction**: Fit refers to how close the model is to the observed data. Predicition refers to how close the model is to new data one might collect. They are not the same. Adding variables, even junk variables, always improves the fit, but the predictions may get better or worse. R^2 is a measure of fit, not of prediction. We will develop a measure of prediction, C_P, later in the course.
- **Statistical Model**: The model underlying multiple regression says:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

where the ε_i are independent $N(0, \sigma^2)$

The true model is unknown, but the least squares fit is an estimate.

- **Hypothesis Tests and Confidence Intervals for a Coefficient**: Testing a hypothesis about a regression coefficient, say $H_0: \beta_5 = 0$, is done using the t-statistic as in simple regression. Confidence intervals are also done as in simple regression.

- **Testing that all coefficients are zero**: The F-test from the ANOVA table is used to test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$.
- **Residual analysis**: One checks the model by plotting the residuals. The minimum is a plot of residuals against predicted, a boxplot of residuals, and a Normal plot of residuals, as in simple regression.

Topic: General Linear Hypothesis

- What is it? In model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

where the ε_i are independent $N(0, \sigma^2)$,

we know how to test a hypothesis about one coefficient, say $H_0\colon \beta_5 = 0$, (t-test) and we know how to test that all of the variables are unneeded, $H_0\colon \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (F-test from regression anova table). The general linear hypothesis says that a particular subset of the coefficients is zero. For example, the hypothesis might say that the last *k–J* variables are not needed, $H_0\colon \beta_{J+1} = \beta_{J+2} = \ldots = \beta_k = 0$.

- Why do this? Generally, a hypothesis expresses an idea. Some ideas need to be expressed using more than one variable. For example, in the FUEL data, the 48 states might be divided into five regions, Northeast, Southeast, Midwest, Mountain, and Pacific, say. Later on, we will see how to code region into several variables in a regression. Testing whether "REGION" matters is testing whether all of these variables can be dropped from the model.
- Comparing Two Models: The test involves comparing two models, a reduced model which assumes the hypothesis is true, and a full model which assumes it is false. To test $H_0\colon \beta_{J+1} = \beta_{J+2} = \ldots = \beta_k = 0$, one fits the full model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

and the reduced model without variables $X_{J+1}, \ldots, X_k$,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_J X_{Ji} + \varepsilon_i,$$

and the test is based on comparing the ANOVA tables for these two models. Details in the textbook.

Topic: Woes of Regression Coefficients

- The phrase, "the woes of regression coefficients" is due to Fred Mosteller and John Tukey in a standard text, *Data Analysis and Regression.* The bulk pack contains another standard reading: George Box's paper "The use and abuse of regression". (The paper is mostly easy to read, but contains some technical material – just skip the technical material. The main points are not technical and not difficult.)
- The issue concerns the interpretation of regression coefficients. The bottom line is that it is hard to interpret regression coefficients. The reason is that whenever you add (or delete) a variable from a regression model, all of the other coefficients change to reflect the added (or deleted) variable. People want (but they can't have) a way of speaking of THE coefficient of a variable, but actually the coefficient of a variable always depends on what other variables are in the model with it. People want to say that β_j is the change in Y expected from a one unit change in X_j, but it simply isn't true. It can't be true, since β_j keeps changing as variables are added or deleted from a model, whereas changing X_j out in the world has nothing to do with which variables I put in the model.
- The bottom line is this: Whenever you hear people say that changing X_j will produce a particular change in Y, and they say they know this solely because they did a regression, you should be a little skeptical. There is more to knowing something like this than just running regressions.

Topic: Transformations

- **Key Idea**: Fit many kinds of curved (i.e., nonlinear) models by transforming the variables and fitting a linear model to the transformed variables.
- **Logs**. For b>0, the base b log has the property that:

  $$y = b^a \quad \text{is the same as} \quad \log_b(y) = a.$$

  Common choices of the base b are b=10, b=2 and b=e=2.71828… for natural logs. Outside high school, if no base is mentioned (e.g., log(y)) it usually means base e or natural logs. Two properties we use often are: log(xy)=log(x)+log(y) and $\log(y^a) = a \cdot \log(y)$.
- **Why transform?** (i) You plot the data and it is curved, so you can't fit a line. (ii) The Y's have boundaries (e.g., Y must be >0 or Y must be between 0 and 1), but linear regression knows nothing of the boundaries and overshots them, producing impossible $\hat{Y}'s$. (iii) The original data violate the linear regression assumptions (such as Normal errors, symmetry, constant variance), but perhaps the transformed variables satisfy the assumptions. (iv) If some Y's are enormously bigger than others, it may not make sense to compare them directly. If Y is the number of people who work at a restaurant business, the Y for McDonald's is very, very big, so much so that it can't be compared to the Y for Genji's (4002 Spruce & 1720 Samson). But you could compare log(Y).
- **Family of transformations**: Organizes search for a good transformation. Family is $\frac{Y^p - 1}{p}$ which tends to log(y) as p gets near 0. Often we drop the shift of 1 and the scaling of 1/p, using just $sign(p) \cdot Y^p$ for $p \neq 0$ and log(y) for p=0. Important members of this family are: (i)

p=1 for no transformation or Y, (ii) p=1/2 for $\sqrt{Y}$, (iii) p=1/3 for $\sqrt[3]{Y}$, (iv) p=0 for log(y), (v) p = –1 for 1/Y.

- **Straightening a scatterplot**: Plot Y vs X. If the plot looks curved, then do the following. Divide the data into thirds based on X, low, middle, high. In each third, find median Y and median X. Gives you three (X,Y) points. Transform Y and/or X by adjusting p until the slope between low and middle equals the slope between middle and high. Then plot the transformed data and see if it looks ok. You want it to look straight, with constant variance around a line.
- **Logit**: logit(a) =log{a/(1-a)} when a is between 0 and 1. If the data are between 0 and 1, their logits are unconstrained.
- **Picking Curves that Make Sense**: Sometimes we let the data tell us which curve to fit because we have no idea where to start. Other times, we approach the data with a clear idea what we are looking for. Sometimes we know what a sensible curve should look like. Some principles – (i) If the residuals show a fan pattern, with greater instability for larger Y's, then a log transformation may shift things to constant variance. (ii) If there is a naïve model based on a (too) simple theory (e.g., weight is proportional to volume), then consider models which include the naïve theory as a very special case. (iii) If outcomes Y must satisfy certain constraints (e.g., percents must be between 0% and 100%), consider families of models that respect those constraints.
- **Interpretable transformations**: Some transformations have simple interpretations, so they are easy to think and write about. Base 2 logs, i.e., $\log_2(y)$ can be interpreted in terms of doublings. Reciprocals, 1/Y, are often interpretable if Y is a ratio (like density) or a time. Squares and

square roots often suggest a relationship between area and length or diameter. Cubes and cube roots suggest a relationship between volume and diameter.

- **Transformations to constant variance**: A very old idea, which still turns up in things you read now and then. Idea is that certain transformations – often strange ones like the arcsin of the square root – make the variance nearly constant, and that is an assumption of regression.

Topic: Polynomials

Why fit polynomials? The transformations we talked about all keep the order of Y intact – big Y's have big transformed Y's. Often that is just what we want. Sometimes, however, we see a curve that goes down and comes back up, like a $\cup$, or goes up and comes back down, like a $\cap$, and the transformations we looked at don't help at all. Polynomials can fit curves like this, and many other wiggles. They're also good if you want to find the X that maximizes Y, the top point of the curve $\cap$.

- **Quadratic**: $y = a + bx + cx^2$ has a $\cup$ shape if c>0 and a $\cap$ shape if c<0 (why?) and is a line if c=0. Top of hill or bottom of valley is at $x = \frac{-b}{2c}$.
- **Fitting a Quadratic**: Easy – put two variables in the model, namely X and X^2.
- **Centering**: If X>0, then X is big at the same time X^2, so these two variables are highly correlated. Often a good idea to center, using X and $(X - \overline{X})^2$ instead of X and X^2. Fits the same curve, but is more stable as a computing algorithm.
- **Orthogonal polynomials**: Typically used in anova rather than in regression. Transforms X^2 so it is uncorrelated with X. Does this by regressing X^2 on X and using the residuals in place of X^2.
- **Cubics**: Can fit cubics using X, X^2 and X^3. Usually don't go beyond cubics. Usually center.
- **Polynomials in several predictors**: If I have two predictors, say x and w, the quadratic in x and w has squared terms, x^2 and w^2, but it adds something new, their crossproduct or interaction, xw:

$$y = a + b \cdot x + c \cdot w + d \cdot x^2 + f \cdot w^2 + h \cdot w \cdot x$$

- **Are quadratic terms needed?** You can judge whether you need several quadratic terms using a general linear hypothesis and its avova table.

Topic: Coded Variables (i.e., Dummy Variables)

- **Why use coded variables?** Coded or dummy variables let you incorporate nominal data (Philly vs New York vs LA) as predictors in regression.
- **Two categories**: If there are just two categories, say male and female, you include a single coded variable, say C=1 for female and C=0 for male. Fits a parallel line model. If you add interactions with a continuous variable, X, then you are fitting a two-line model, no longer a parallel line model.
- **More than Two Categories**: If there are 3 categories (Philly vs New York vs LA) then you need two coded variables to describe them (C=1, D=0 for New York; C=0, D=1 for LA; C=0, D=0 for Philly). Such a model compares each group to the group left out, the group without its own variable (here, Philly). When there are more than two categories – hence more than one coded variable – interesting hypotheses often involve several variables and are tested with the general linear hypothesis. Does it matter which group you leave out? Yes and no. Had you left out NY rather than Philly, you get the same fitted values, the same residuals, the same overall F-test, etc. However, since a particular coefficient multiplies a particular variable, changing the definition of a variable changes the value of the coefficient.

Topic: Diagnostics -- Residuals

- **Why do we need better residuals**?: We look at residuals to see if the model fits ok – a key concern for any model. But the residuals we have been looking at are not great. The problem is that least squares works very hard to fit data points with extreme X's – unusual predictors – so it makes the residuals small in those cases. A data point with unusual X's is called a high leverage point, and we will think about them in detail a little later. A single outlier (weird Y) at a high leverage point can pull the whole regression towards itself, so this point looks well fitted and the rest of the data looks poorly fitted. We need ways of finding outliers like this. We want regression to tell us what is typical for most points – we don't want one point to run the whole show.
- The model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

$$\text{where the } \varepsilon_i \text{ are independent } N(0, \sigma^2)$$

- By least squares, we estimate the model to be:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki} \text{ with residuals } E_i = Y_i - \hat{Y}_i$$

- Although the true errors, ε_i have constant variance, $\text{var}(\varepsilon_i) = \sigma^2$, the same for every unit i, the residuals have different variances, $\text{var}(E_i) = \sigma^2(1 - h_i)$ where h_i is called the leverage.

- The standardized residual we like best does two things: (i) it uses the leverages h_i to give each residual the right variance for that one residual, and (ii) it removes observation i when estimating the variance of the residual $E_i = Y_i - \hat{Y}_i$ for observation i. That is, $\text{var}(E_i) = \sigma^2(1 - h_i)$ is estimated by $\hat{\sigma}_{[-i]}{}^2(1-h_i)$, where $\hat{\sigma}_{[-i]}{}^2$ is the estimate of the residual variance we get by setting observation i aside and fitting the regression without it.

- **The residual we like has several names** – studentized, deleted, jacknife – no one of which is used by everybody. It is the residual divided by its estimated standard error:

$$r_{[i]} = \frac{E_i}{\sigma_{[-i]}\sqrt{1-h_i}}$$

- Another way to get this residual is to create a coded variable that is 1 for observation i and 0 for all other observations. Add this variable to your regression. The t-statistic for its coefficient equals $r_{[i]}$.

- We can test for outliers as follows. The null hypothesis says there are no outliers. If there are n observations, there are n deleted residuals $r_{[i]}$. Find the largest one in absolute value. To test for outliers at level 0.05, compute 0.025/n, and reject the hypothesis of no outliers if the largest absolute deleted residual is beyond the 0.025/n percentage point of the t-distribution with one less degree of freedom than the error line in the anova table for the regression. (You lose one degree of freedom for the extra coded variable mentioned in the last paragraph.)

Topic: Diagnostics -- Leverage

- **Three very distinct concepts**: An outlier is an observation that is poorly fitted by the regression – it has a response Y that is not where the other data points suggest its Y should be. A high leverage point has predictors, X's, which are unusual, so at these X's, least squares relies very heavily on this one point to decide where the regression plane should go – a high leverage point has X's that allow it to move the regression if it wants to. A high influence point is one that did move the regression – typically, such a point has fairly high leverage (weird X's) and is fairly poorly fitted (weird Y for these X's); however, it may not be the one point with the weirdest X or the one point with the weirdest Y. People often mix these ideas up without realizing it. Talk about a weird Y is outlier talk; talk about a weird X is leverage talk; talk about a weird Y for these X's is influence talk. We now will measure leverage and later influence.
- **Measuring Leverage**: Leverage is measured using the leverages h_i we encountered when we looked at the variance of the residuals. The leverages are always between 0 and 1, and higher values signify more pull on the regression.
- **When is leverage large?** If a model has k predictors and a constant term, using n observations, then the average leverage, averaging over the n observations is always $\frac{k+1}{n} = \frac{1}{n}\sum_{i=1}^{n} h_i$. A rule a thumb that works well is that leverage is large if it is at least twice the average, $h_i \geq \frac{2(k+1)}{n}$.
- **What do you do if the leverage is large?** You look closely. You think. Hard. You find the one or two or three points with $h_i \geq \frac{2(k+1)}{n}$ and you

look closely at their data. What is it about their X's that made the leverage large? How, specifically, are they unusual? Is there a mistake in the data? If not, do the X's for these points make sense? Do these points belong in the same regression with the other points? Or should they be described separately? Regression gives high leverage points a great deal of weight. Sometimes that makes sense, sometimes not. If you were looking at big objects in our solar system, and X=mass of object, you would find the sun is a high leverage point. After thinking about it, you might reasonably decide that the regression should describe the planets and the sun should be described separately as something unique. With the solar system, you knew this before you looked at the data. Sometimes, you use regression in a context where such a high leverage point is a discovery. (If you remove a part of your data from the analysis, you must tell people you did this, and you must tell them why you did it.)

- **Interpretation of leverage**: Leverage values h_i have several interpretations. You can think of them as the distance between the predictors X for observation i and the mean predictor. You can think of them as the weight that observation i gets in forming the predicted value $\hat{Y}_i$ for observation i. You can think of leverages as the fraction of the variance of Y_i that is variance of $\hat{Y}_i$. We will discuss these interpretations in class.

Topic: Diagnostics -- Influence

- **What is influence?** A measure of influence asks whether observation i *did* move the regression. Would the regression change a great deal if this one observation were removed? Not whether it *could* move the regression – that's leverage. Not whether it fits poorly – that's an outlier.
- **Measures of influence.** There are several measures of influence. They are all about the same, but no one has become the unique standard. Two common choices are DFFITS and Cook's Distance. Cook's distance is (almost) a constant times the square of DFFITS, so it makes little difference which one you use. It is easier to say what DFFITS does.
- **What is DFFITS?** Roughly speaking, DFFITS measures the change in the predicted value for observation i when observation it is removed from the regression. Let $\hat{Y}_i$ be the predicted value for observation i using all the data, and let $\hat{Y}_{i[i]}$ be the predicted value for observation i if we fit the regression without this one observation. Is $\hat{Y}_i$ close to $\hat{Y}_{i[i]}$? If yes, then this observation does not have much influence. If no, then it does have influence. DFFITS divides the difference, $\hat{Y}_i - \hat{Y}_{i[i]}$ by an estimate of the standard error of $\hat{Y}_i$, so a value of 1 means a movement of one standard erro. Recall that $\hat{\sigma}_{[-i]}{}^2$ is the estimated residual variance when observation i is removed from the regression. Then DFFITS is:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i[i]}}{\hat{\sigma}_{[-i]}\sqrt{h_i}}.$$

- **DFBETAS**: A related quantity is DFBETAS which looks at the standardized change in the regression coefficient $\hat{\beta}_j$ when observation i is removed. There is one DFBETAS for each observation and for each coefficient. DFFITS is always bigger than the largest DFBETAS, and there is only one DFFITS per observation, so many people look at DFFITS instead of all k DFBETAS.

Topic: Variable Selection

- **What is variable selection?** You have a regression model with many predictor variables. The model looks ok – you've done the diagnostic checking and things look fine. But there are too many predictor variables. You wonder if you might do just as well with fewer variables. Deciding which variables to keep and which to get rid of is variable selection.
- **Bad methods**. There are two bad methods you should not use. One bad method is to drop all the variables with small t-statistics. The problem is the t-statistic asks whether to drop a variable *providing you keep all the others*. The t-statistic tells you little about whether you can drop two variables at the same time. It might be you could drop either one but not both, and t can't tell you this. Another bad method uses the squared multiple correlation, R^2. The problem is R^2 always goes up when you add variables, and the size of the increase in R^2 is not a great guide about what to do. Fortunately, there is something better.
- **A good method**. The good method uses a quantity called C_P which is a redesigned R^2 built for variable selection. Suppose the model is, as before,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

where the ε_i are independent $N(0, \sigma^2)$, but now k is large (many predictors) and we think some β's might be zero. We fit this model and get the usual estimate $\hat{\sigma}^2$ of σ^2. A submodel has some of the k variables but not all of them, and we name the submodel by the set P of variables it contains. So the name of the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_9 X_{9i} + \varepsilon_i$ is

P={1,3,9}, and it has residual sum of squares SSE_P from the residual line in its Anova table, and p=3 variables plus one constant term or 4 parameters. (Note carefully – I let p=#variables, but a few people let p=#parameters.) We have n observations. Then the strange looking but simple formula for C_P is: $C_P = \frac{SSE_P}{\hat{\sigma}^2} - [n-2(p+1)]$. Then C_P compares the model with all variable to the model with just the variables in P and asks whether the extra variables are worth it.

- Using C_P: The quantity C_P estimates the standardized total squared error of prediction when using model P in place of the model with all the variables. We like a model P with a small C_P. If a model P contains all the variables with nonzero coefficients, then C_P tends on average to estimate p+1, the number of variables plus 1 for the constant, so a good value of C_P is not much bigger than p+1. For instance, if $C_{\{1,3,9\}}$ = 8, then that is much bigger than p+1=3+1=4, so the model seems to be missing important variables, but if $C_{\{1,3,9,11\}}$ =5.1 then that is close to p+1=4+1=5 and smaller than 8, so that model predicts better and might have all important variables.
- **Searching**: If a model has k variables, then there are 2^k submodels formed by dropping variables, or about a billion models for k=30 variables. There are various strategies for considering these models: forward selection, backward elimination, stepwise, all subsets, best subsets.
- **Cautions**: Variable selection is an exploratory method, one that looks for interesting things, but because it searches so extensively, it may find some things that don’t replicate. If we reject hypotheses when P-

value<0.05, then only 1 time in 20 do we reject a true hypothesis. But if we fit billions for regressions, calculating billions of P-values, then we reject many true hypotheses and make many mistakes. The results of variable selection need to be examined with caution avoiding overstatement. A good strategy is to split the sample, perform variable selection on one half, and confirm the results on the other. This is a simple type of cross-validation.

Topic: One Way Analysis of Variance

- **What is ANOVA**? Anova, or analysis of variance, is the decomposition of data into parts that add back up to the original data, and the summary of the parts in terms of their sizes measured by summing and squaring their numerical entries. At an abstract level, in statistical theory, anova and regression are not really different. In practice, however, they look very different. Most computer programs have separate routines for regression and anova. Center questions, issues and methods arise in anova that don't arise in regression. Anova tends to be used with structured data sets, often from carefully designed experiments, while regression is often used with data that arises naturally. However, by running enough regressions, knowing exactly what you are doing, and putting together the pieces very carefully, you can do even a complex anova using a regression program – it's easy to use an anova program. Anova has a nice geometry.
- **What is one-way anova?** One-way anova is the very simplest case. People fall into one of several groups and we want to understand the difference between the groups. Basic questions are: Do the groups differ? (F-test.) If so, how? (Multiple comparisons.) If I anticipate a specific pattern of differences between the groups, is this pattern confirmed by the data? (Contrasts.) Notation: There are I groups, i=1,…,I, and n_i people in group I, with $n = n_1 + \ldots + n_I$ people in total. Each person is in just one group and people in different groups have nothing to do with each other. Person j in group i, has response y_{ij}. The

mean response in group i is $\bar{y}_{i\bullet}$ and the mean response for everyone is $\bar{y}_{\bullet\bullet}$. The anova decomposition is:

$$y_{ij} = \bar{y}_{\bullet\bullet} + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet})$$

data = (grand mean) + (group difference) + (residual)

Model for One-Way Anova: The model for one-way anova says the observations are Normal with the same variance, are independent of each other, and have different means in the different groups. Specifically: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where the ε_{ij} are iid $N(0,\sigma^2)$ and $0 = \alpha_1 + \alpha_2 + \ldots + \alpha_I$.

- **Do the groups differ?** We test the hypothesis that the groups do not differ using the F-ratio from the one-way analysis of variance table. If F is large enough, judged by the F-table, we reject the null hypotheses and conclude there is strong evidence the groups differ. Otherwise, we conclude that we lack strong evidence that the groups differ (which is not the same thing as saying we know for certain they are the same).
- **If the groups differ, how do they differ?** It is not enough to know the groups differ – we need to understand what differences are present. There are two cases: (1) we have no idea what we are looking for, or (2) we have a clear and specific idea what we are looking for. Case 2 is the better case – if you know what you are looking for in statistics, then you can find it with a smaller sample size. We handle case 1 using multiple comparisons and case 2 using contrasts, described later. In multiple comparison, every group is compared to every other group. If there are I=10 groups, there are 45 comparisons of two groups, 1 with 2, 1 with 3, …, 9 with 10. If you did 45 t-tests to compare the groups, rejected for P-values < 0.05, then you would falsely reject a true hypothesis of no

difference in one out of 20 tests. This means that with I=10 groups and 45 comparisons, if the groups are really the same, you expect to get 45x0.05 = 2.25 significant (P-value<0.05) difference by chance alone. That's a problem – a big problem. It means you expect 2 mistakes – you expect to say a treatment worked when it didn't. Gotta do something to prevent this. There are many, many solutions to this problem – there are whole books of multiple comparison procedures. One of the first is due to John Tukey of Princeton. You can think of it as using essentially a t-statistic to compare groups in pairs, but as the number of groups, I, gets bigger, so more tests are being done, the procedure requires a bigger value of the t-statistic before declaring the difference significant. If you did this with I=10 groups and 45 comparisons of two groups, the procedure promises that if the groups are really the same, the chance of a significant difference anywhere in the 45 comparisons is less that 0.05 – if you find anything, you can believe it. For example, with just two groups and 30 degrees of freedom, we would reject at the 0.05 level if |t|>2.04, but using Tukey's method with I=10 groups, 45 comparisons, we would reject if |t|>3.41, and if t is that big, then it is unlikely to happen by chance even if you did 45 comparisons.

- **Planned Contrasts for Specific Hypotheses:** Ideally, when you do research, you have a clear idea what you are looking for and why. When this is true, you can build a test tailored to your specific hypothesis. You do this with contrasts among group means. You express your hypothesis using a set of contrast weights you pick, one weight for each group mean, summing to zero: $c_1, c_2, \ldots, c_I$ with $c_1 + c_2 + \ldots + c_I = 0$. For instance, consider a study with I=3 groups, with $n_1 = n_2 = n_3 = 100$ people in each group. The groups are two different drug treatments, A and B, and a placebo control. Then the contrast "drug vs placebo" is:

Contrast: Drug vs Placebo

Placebo	Drug A	Drug B
c_1	c_2	c_3
-1	$\frac{1}{2}$	$\frac{1}{2}$

whereas the contrast "drug A vs drug B" is:

Contrast: Drug A vs Drug B

Placebo	Drug A	Drug B
d_1	d_2	d_3
0	1	-1

- The value of contrast applies the contrast weights to the group means, $L = \sum_{i=1}^{I} c_i \cdot \overline{y}_{i\bullet}$, so for "Drug vs Placebo" it is $L = -1 \cdot \overline{y}_{1\bullet} + \frac{1}{2} \cdot \overline{y}_{2\bullet} + \frac{1}{2} \cdot \overline{y}_{3\bullet}$
- The t-test for a contrast tests the null hypothesis $H_0\colon\ 0 = \sum_{i=1}^{I} c_i \cdot \alpha_i$. Let $\hat{\sigma}^2$ be the residual mean square from the anova table, which estimates σ^2. The t-statistic is $t = \dfrac{L}{\sqrt{\hat{\sigma}^2 \cdot \sum \dfrac{c_i^2}{n_i}}}$ and the degrees of freedom are from the residual line in the anova table.
- The sum of squares for a contrast is $\dfrac{L^2}{\sum \dfrac{c_i^2}{n_i}}$. Two contrasts, $c_1, c_2, \ldots, c_I$ and $d_1, d_2, \ldots, d_I$ are orthogonal if $0 = \sum_{i=1}^{I} \dfrac{c_i \cdot d_i}{n_i}$. Example: "Drug vs Placebo" is orthogonal to "Drug A vs Drug B" because

$\sum_{i=1}^{I} \frac{c_i \cdot d_i}{n_i} = \frac{-1 \times 0}{100} + \frac{\frac{1}{2} \times 1}{100} + \frac{\frac{1}{2} \times -1}{100} = 0$. When contrasts are orthogonal, the sum of squares between groups may be partitioned into separate parts, one for each contrast. If there are I groups, then there are I-1 degrees of freedom between groups, and each degree of freedom can have its own contrast. Both of these formulas are mostly used in balanced designs where the sample sizes in the groups are the same, $n_1 = n_2 = \ldots = n_I$.

Topic: Two Way Analysis of Variance

- What is two-way ANOVA? In two-way anova, each measurement is classified into groups in two different ways, as in the rows and columns of a table. In the social sciences, the most common situation is to measure the same unit or person under several different treatments – this is the very simplest case of what is know as repeated measurements. Each person is a row, each treatment is a column, and each person gives a response under each treatment. The two-way's are person and treatment. Some people give higher responses than others. Some treatments are better than others. The anova measures both sources of variation. The units might be businesses or schools or prisons instead of people.

  Notation: There are I people, i=1,…,I, and J treatments, j=1,…,J, and person i gives response y_{ij} under treatment j. The mean for person i is $\bar{y}_{i\bullet} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}$ and the mean for treatment j is $\bar{y}_{\bullet j} = \frac{1}{I}\sum_{i=1}^{I} y_{ij}$, and the mean of everyone is $\bar{y}_{\bullet\bullet}$. The anova decomposition is:

  $y_{ij} = \bar{y}_{\bullet\bullet} + (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})$.

- **Anova table**: The anova table now has “between rows”, “between columns” and “residual”, so the variation in the data is partitioned more finely.
- **Normal model**: The model is $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ where the errors are independent Normals with mean zero and variance σ^2. Under this model, F-statistics from the anova table may be used to test the hypotheses of no difference between rows and no difference between columns. Can do multiple comparisons and contrasts using the residual line from the anova table to obtain the estimate $\hat{\sigma}^2$.

Error Rates When Performing More Than One Hypothesis Test

Table Counts Null Hypotheses

	Accepted or untested null hypotheses	Rejected null hypotheses	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	$m- m_0$
Total	m-R=m-(U+T)	R	m

Family-wise error rate: $Pr(V\geq 1)$, the probability of at least one false rejection in m tests.

The family-wise error rate is weakly controlled at $\alpha=0.05$ if $\alpha=0.05\geq Pr(V\geq 1)$ whenever $m= m_0$, that is, whenever all m null hypotheses are true.

The family-wise error rate is strongly controlled at $\alpha=0.05$ if $\alpha=0.05\geq Pr(V\geq 1)$ for all values of m_0, that is, no matter how many null hypotheses are true.

Weak control is not enough. Weak control means you are unlikely to find something when there is nothing, but you are still likely to find too much when there is something.

False discovery rate (FDR) is the expected number of false rejections, E(V/R) where E/R=0/0 is defined to be 0 (i.e., no false rejections if no rejections). This is a more lenient standard than the family-wise error rate, rejecting more true hypotheses.

If you do m tests at level $\alpha=0.05$, you expect to falsely reject 0.05 x m_0 hypotheses, and if all hypotheses are true, this might be as high as 0.05 x m. The expected ratio of false rejections to tests, E(V/m), is called the per comparison error rate.

Example
Table Counts Null Hypotheses

	Accepted or untested null hypotheses	Rejected null hypotheses	Total
True null hypotheses	U	V	100
False null hypotheses	T	S	1
Total	m-R=m-(U+T)	R	101

In this example, there are 101 hypotheses and 100 are true.
If you test each hypothesis at level α=0.05, you expect 0.05 x 100 = 20 false rejections of true null hypotheses, plus if you are lucky a rejection of the one false null hypothesis, so you expect most rejections to be false rejections.

If you strongly control the family-wise error rate at α=0.05, then the chance of at least one false rejection is at most 5%.

If you weakly control the family-wise error rate at α=0.05, then there are no promises about false rejections in this case, as one null hypothesis is false.

What are adjusted P-values? (e.g. as produced by `pairwise.t.test()`)

A test of null hypothesis H_0 either rejects H_0 or it does not.
The level, α, of the test is such that $\alpha \geq \Pr(\text{Reject } H_0)$ when H_0 is true.
The P-value is the smallest α such that we reject H_0.
This definition of a P-value continues to work with multiple hypothesis testing.

Topic: Factorial Analysis of Variance

- **Two Factor Factorial Anova**: The simplest case of factorial anova involves just two factors – similar principles apply with more than two factors, but things get large quickly. Suppose you have two drugs, A and B – then “drug” is the first factor, and it has two levels, namely A and B. Suppose each drug has two dose levels, low and high – then “dose” is the second factor, and it too has two levels, low and high. A person gets one combination, perhaps drug B at low dose. Maybe I give 50 people each drug at each level, so I have 200 people total, 100 on A, 100 on B, 100 at low dose, 100 at high dose.

  Main effects and interactions: We are familiar with main effects – we saw them in two-way anova. Perhaps drug A is better than drug B – that’s a main effect. Perhaps high dose is more effective than low dose – that’s a main effect. But suppose instead that drug A is better than drug B at high dose, but drug A is inferior to drug B at low dose – that’s an interaction. In an interaction, the effect of one factor changes with the level of the other.

- **Anova table**: The anova table has an extra row beyond that in two-way anova, namely a row for interaction. Again, it is possible to do contrasts and multiple comparisons.
- **More Complex Anova**: Anova goes on and on. The idea is to pull apart the variation in the data into meaningful parts, each part having its own row in the anova table. There may be many factors, many groupings, etc.

Some Aspects of R

Script is my commentary to you. **Bold Courier is what I type in R.** Regular Courier is what R answered.

What is R?

R is a close relative of Splus, but R is available for free. You can download R from http://cran.r-project.org/ . *R is very powerful and is a favorite (if not the favorite) of statisticians; however, it is not easiest package to use. It is command driven, not menu driven, so you have to remember things or look them up – that's the only thing that makes it hard. You can add things to R that R doesn't yet know how to do by writing a little program. R gives you fine control over graphics. Most people need a book to help them, and so Mainland & Braun's Data Analysis and Graphics Using R, Cambridge University Press. Abnother book is Dalgaard's Introductory Statistics with R, NY: Springer. Dalgaard's book is better at teaching basic statistics, and it is good if you need a review of basic statistics to go with an introduction to R. R is similar to Splus, and there are many good books about Splus. One is: Venables and Ripley Modern Applied Statistics with S-Plus (NY: Springer-Verlag).*

Who should use R?

If computers terrify you, if they cause insomnia, cold sweats, and anxiety attacks, perhaps you should stay away from R. On the other hand, if you want a very powerful package for free, one you won't outgrow, then R worth a try. If you find you need lots of help to install R or make R work, then R isn't for you. Alternatives for Statistics 500 are JMP-IN, SPSS, Systat, Stata, SAS and many others. For Statistics 501, beyond the basics, R is clearly best.

You need to download R the first time from the webpage above.

You need to get the "Rst500" workspace for the course from
http://www-stat.wharton.upenn.edu/
going to "Course downloads" and the most recent Fall semester, Statistics 500, or in one step to
http://download.wharton.upenn.edu/download/pub/stat/Fall-2006/STAT-500/
For Statistics 501,
http://stat.wharton.upenn.edu/statweb/course/Spring-2007/stat501/

Start R.
From the File Menu, select "Load Workspace".
Select "Rst500"

To see what is in a workspace, type

```
     ls()
```

or type

```
     objects()
```

```
> ls()
[1] "fuel"
```

To display an object, type its name

```
> fuel
   ID state Fuel   Tax License   Inc   Road
1   1    ME  541  9.00    52.5 3.571  1.976
2   2    NH  524  9.00    57.2 4.092  1.250
3   3    VT  561  9.00    58.0 3.865  1.586
                     .
                     .
                     .
46 46    WN  510  9.00    57.1 4.476  3.942
47 47    OR  610  7.00    62.3 4.296  4.083
48 48    CA  524  7.00    59.3 5.002  9.794
```

Fuel is a data.frame.

```
> is.data.frame(fuel)
[1] TRUE
```

You can refer to a variable in a data frame as fuel$Tax, etc. It returns one column of fuel.

```
> fuel$Tax
 [1]  9.00  9.00  9.00  7.50  8.00 10.00  8.00  8.00  8.00  7.00  8.00  7.50
[13]  7.00  7.00  7.00  7.00  7.00  7.00  7.00  8.50  7.00  8.00  9.00  9.00
[25]  8.50  9.00  8.00  7.50  8.00  9.00  7.00  7.00  8.00  7.50  8.00  6.58
[37]  5.00  7.00  8.50  7.00  7.00  7.00  7.00  7.00  6.00  9.00  7.00  7.00
```

length() and dim() tell you how big things are. There are 48 states and seven variables.

```
> length(fuel$Tax)
[1] 48
> dim(fuel)
[1] 48  7
```

To get a summary of a variable, type summary(variable)

```
> summary(fuel$Tax)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.000   7.000   7.500   7.668   8.125  10.000
```

R has very good graphics. You can make a boxplot with

```
boxplot(fuel$Fuel)
```

or dress it up with

```
boxplot(fuel$Fuel,ylab="gallons per person",main="Figure 1:
Motor Fuel Consumption")
```

To learn about a command, type help(command)

```
help(boxplot)

help(plot)

help(t.test)

help(lm)
```

Optional Trick

It can get tiresome typing fuel$Tax, fuel$Licenses, etc. If you type attach(data.frame) then you don't have to mention the data frame. Type detach(data.frame) when you are done.

```
> summary(fuel$Tax)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.000   7.000   7.500   7.668   8.125  10.000 
> summary(Tax)
Error in summary(Tax) : Object "Tax" not found
> attach(fuel)
> summary(Tax)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.000   7.000   7.500   7.668   8.125  10.000 
> summary(License)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  45.10   52.98   56.45   57.03   59.52   72.40 
> detach(fuel)
```

HELP

R contains several kinds of help. Use help(keyword) to get documentation about keyword.

```
> help(boxplot)
```

Use help("key") to find the keywords that contain "key". The quotes are needed.

```
> apropos("box")
[1] "box"              "boxplot"          "boxplot.default"
     "boxplot.stats"
```

Use help.search("keyword") to search the web for R functions that you can download related to keyword. Quotes are needed.

```
> help.search("box")

> help.search("fullmatch")
```

At http://cran.r-project.org/ there is free documentation, some of which is useful, but perhaps not for first-time users. To begin, books are better.

Some R

A variable, "change" in a data.frame bloodpressure.

```
> bloodpressure$change
 [1]  -9  -4 -21  -3 -20 -31 -17 -26 -26 -10 -23 -33 -19 -19 -23
```

It doesn't know what "change" is.

```
> change
Error: Object "change" not found
```

Try attaching the data.frame

```
> attach(bloodpressure)
```

Now it knows what "change" is.

```
> change
 [1]  -9  -4 -21  -3 -20 -31 -17 -26 -26 -10 -23 -33 -19 -19 -23

> mean(change)
[1] -18.93333
> sd(change)
[1] 9.027471
> summary(change)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -33.00  -24.50  -20.00  -18.93  -13.50   -3.00 
> stem(change)

  The decimal point is 1 digit(s) to the right of the |

  -3 | 31
  -2 | 663310
  -1 | 9970
  -0 | 943

> hist(change)
> boxplot(change)
> boxplot(change,main="Change in Blood Pressure After
Captopril",ylab="Change mmHg")
> boxplot(change,main="Change in Blood Pressure After
Captopril",ylab="Change mmHg",ylim=c(-40,40))
> abline(0,0,lty=2)
> t.test(change)

        One Sample t-test

data:  change 
t = -8.1228, df = 14, p-value = 1.146e-06
alternative hypothesis: true mean is not equal to 0 
95 percent confidence interval:
 -23.93258 -13.93409 
sample estimates:
mean of x 
-18.93333 
```

Are the Data Normal?

```
> attach(bloodpressure)
> change
 [1]  -9  -4 -21  -3 -20 -31 -17 -26 -26 -10 -23 -33 -19 -19 -23
> par(mfrow=c(1,2))
> boxplot(change)
> qqnorm(change)
```

A straight line in a Normal quantile plot is consistent with a Normal distribution.

You can also do a Shapiro-Wilk test. A small p-value suggests the data are not Normal.

```
> shapiro.test(change)

        Shapiro-Wilk normality test

data:  change
W = 0.9472, p-value = 0.4821
```

The steps below show what the qqnorm() function is plotting

```
> round(ppoints(change),3)
 [1] 0.033 0.100 0.167 0.233 0.300 0.367 0.433 0.500 0.567 0.633
[11] 0.700 0.767 0.833 0.900 0.967
```

The plotting positions in the normal plot:

```
> round(qnorm(ppoints(change)),3)
 [1] -1.834 -1.282 -0.967 -0.728 -0.524 -0.341 -0.168  0.000  0.168
[10]  0.341  0.524  0.728  0.967  1.282  1.834
```

qqnorm(change) is short for

```
> plot(qnorm(ppoints(change)),sort(change))
```

Here are Normal quantile plots of several Normal and non-Normal distributions.
Can you tell from the plot which are Normal?

```
> qqnorm(rnorm(10))
> qqnorm(rnorm(100))
> qqnorm(rnorm(1000))
> qqnorm(rcauchy(100))
> qqnorm(rlogis(100))
> qqnorm(rexp(100))
```

Regression in R

Script is my commentary to you. **Bold Courier is what I type in R.** Regular Courier is what R answered.

```
> ls()
[1] "fuel"
```

To display an object, type its name

```
> fuel
   ID state Fuel   Tax License   Inc   Road
1   1    ME  541  9.00    52.5 3.571  1.976
2   2    NH  524  9.00    57.2 4.092  1.250
3   3    VT  561  9.00    58.0 3.865  1.586
                     .
                     .
                     .
46 46    WN  510  9.00    57.1 4.476  3.942
47 47    OR  610  7.00    62.3 4.296  4.083
48 48    CA  524  7.00    59.3 5.002  9.794
```

To do regression, use lm. lm stands for linear model.

To fit Fuel = α + □ Tax + □, type

```
> lm(Fuel~Tax)

Call:
lm(formula = Fuel ~ Tax)

Coefficients:
(Intercept)          Tax
     984.01       -53.11
```

To fit Fuel = □□ + □□ Tax + □□□icense + □, type

```
> lm(Fuel~Tax+License)

Call:
lm(formula = Fuel ~ Tax + License)

Coefficients:
(Intercept)          Tax      License
     108.97       -32.08        12.51
```

To see more output, type

```
> summary(lm(Fuel~Tax))

Call:
lm(formula = Fuel ~ Tax)

Residuals:
     Min       1Q   Median       3Q      Max
-215.157  -72.269    6.744   41.284  355.736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   984.01     119.62   8.226 1.38e-10 ***
Tax           -53.11      15.48  -3.430  0.00128 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 100.9 on 46 degrees of freedom
Multiple R-Squared: 0.2037,     Adjusted R-squared: 0.1863
F-statistic: 11.76 on 1 and 46 DF,  p-value: 0.001285
```

You can save the regression in an object and then refer to it:

```
> reg1<-lm(Fuel~Tax+License)
```

Now the workspace has a new object, namely `reg1`:

```
> ls()
[1] "fuel" "reg1"
```

To see reg1, type its name:

```
> reg1

Call:
lm(formula = Fuel ~ Tax + License)

Coefficients:
(Intercept)          Tax      License
     108.97       -32.08        12.51
```

To get residuals, type

```
> reg1$residuals
```

This works only because I defined reg1 above. To boxplot residuals, type:

```
>boxplot(reg1$residuals)
```

To plot residuals against predicted values, type

```
> plot(reg1$fitted.values,reg1$residuals)
```

To do a normal plot of residuals, type

```
> qqnorm(reg1$residuals)
```

To get deleted or jackknife residuals, type

```
> rstudent(reg1)
```

To get leverages or hats, type

```
>hatvalues(reg1)
```

To get dffits

```
> dffits(reg1)
```

To get Cook's distance

```
> cooks.distance(reg1)
```

Clean up after yourself. To remove reg1, type rm(reg1)

```
> ls()
[1] "fuel" "reg1"
> rm(reg1)
> ls()
[1] "fuel"
```

Predictions

Fit a linear model and save it.

```
> mod<-lm(Fuel~Tax)
```

A confidence interval for the line at Tax = 8.5

```
>  predict(mod,data.frame(Tax=8.5),interval="confidence")
          fit      lwr      upr
[1,] 532.6041 493.4677 571.7405
```

A prediction interval for a new observation at Tax = 8.5

```
> predict(mod,data.frame(Tax=8.5),interval="prediction")
          fit      lwr      upr
[1,] 532.6041 325.7185 739.4897
```

Same point estimate, 532.6 gallons, but a very different interval, because the prediction interval has to allow for a new error for the new observation.

Multiple Regression Anova in R

The standard summary output from a linear model in R contains the key elements of the anova table, which are underlined.

```
> summary(lm(Fuel~Tax+License))
Call:
lm(formula = Fuel ~ Tax + License)

Residuals:
     Min       1Q   Median       3Q      Max 
-123.177  -60.172   -2.908   45.032  242.558 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  108.971    171.786   0.634   0.5291    
Tax          -32.075     12.197  -2.630   0.0117 *  
License       12.515      2.091   5.986 3.27e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1 

Residual standard error: 76.13 on 45 degrees of freedom
Multiple R-Squared: 0.5567,     Adjusted R-squared: 0.537 
F-statistic: 28.25 on 2 and 45 DF,  p-value: 1.125e-08
```

More explicitly, the model `lm(Fuel~1)` *fits just the constant term, and the F test compares that model (with just the constant term) to the model with all the variables (here Tax & License).*

```
> anova(lm(Fuel~1),lm(Fuel~Tax+License))
Analysis of Variance Table

Model 1: Fuel ~ 1
Model 2: Fuel ~ Tax + License
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     47 588366                                  
2     45 260834  2    327532 28.253 1.125e-08 ***
```

Most regression programs present an explicit anova table, similar to that above, rather than just the F-test.

Partial Correlation Example

Here are the first two lines of data from a simulated data set. We are interested in the relationship between y and x2, taking account of x1.

```
> partialcorEG[1:2,]
           y         x1         x2
1 -3.8185777 -0.8356356 -1.0121903
2  0.3219982  0.1491024  0.0853746
```

Plot the data. Always plot the data.

```
> pairs(partialcorEG)
```

Notice that y and x2 have a positive correlation.

```
> cor(partialcorEG)
           y        x1        x2
y  1.0000000 0.9899676 0.9535053
x1 0.9899676 1.0000000 0.9725382
x2 0.9535053 0.9725382 1.0000000
```

The partial correlation is the correlation between the residuals. Notice that y and x2 have a negative partial correlation adjusting for x1.

```
> cor(lm(y~x1)$residual,lm(x2~x1)$residual)
[1] -0.2820687
```

Notice that the multiple regression coefficient has the same sign as the partial correlation.

```
> summary(lm(y~x1+x2))
Call:
lm(formula = y ~ x1 + x2)
Residuals:
     Min       1Q   Median       3Q      Max
-1.13326 -0.27423 -0.02018  0.32216  1.07808
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.007177   0.048662   0.147  0.88305
x1           4.768486   0.243833  19.556  < 2e-16 ***
x2          -0.720948   0.248978  -2.896  0.00468 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.4866 on 97 degrees of freedom
Multiple R-Squared: 0.9816,     Adjusted R-squared: 0.9812
F-statistic:  2591 on 2 and 97 DF,  p-value: < 2.2e-16
```

Added Variable Plots

You have fit a model, mod1 say, $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$ where ε are iid $N(0,\sigma^2)$ and now want to ask about adding a new variable, x_{k+1}, to this model.

It can't hurt to plot Y against x_{k+1}. However, that plot does not tell you what x_{k+1} will do in the model above. It could happen that Y increases with x_{k+1} but $\beta_{k+1} < 0$.

The added variable plot uses the idea of regression by stages. In regression by stages, you estimate β_{k+1} by regressing the residuals from mod1 on the residuals of x_{k+1} when regressed on $x_1 , \ldots , x_k$. The added variable plot is simply the plot of these two sets of residuals, residuals of Y versus residuals of x_{k+1}. The slope in that plot estimates β_{k+1}. So the added variable plot lets you see what happens when x_{k+1} is added to mod1.

You can calculate the two set of residuals and plot them. That works fine. Or you can use **addedvarplot** in the course workspace.

```
> attach(fuel)
> head(fuel)
  ID state Fuel  Tax License   Inc  Road
1  1    ME  541  9.0    52.5 3.571 1.976
2  2    NH  524  9.0    57.2 4.092 1.250
3  3    VT  561  9.0    58.0 3.865 1.586
4  4    MA  414  7.5    52.9 4.870 2.351
5  5    RI  410  8.0    54.4 4.399 0.431
6  6    CN  457 10.0    57.1 5.342 1.333
> mod1<-lm(Fuel~Tax+License)
> addedvarplot(mod1,Inc)
The same plot is produced directly by:
> plot(lm(Inc~Tax+License)$resid,mod1$resid)
```

ADDED VARIABLE PLOTS IN THE car Package

```
> attach(fuel)
> pairs(cbind(Fuel,Tax,License))
> library(car)
> help(avPlots)
> avPlots(lm(Fuel~Tax+License))
> avPlots(lm(Fuel~Tax+License+Inc),term=~Inc)
> avPlots(lm(Fuel~Tax+License+Inc))
> summary(lm(Fuel~Tax+License+Inc))
```

car stands for “Companion to Applied Regression”. The book, *An R Companion to Applied Regression*” by John Fox and Sanford Weisberg discusses regression using this package.

Vocabulary Homework

```
> vocabulary
    Age Vocab
1  0.67     0
2  0.83     1
3  1.00     3
4  1.25    19
5  1.50    22
6  1.75   118
7  2.00   272
8  2.50   446
9  3.00   896
10 3.50  1222
11 4.00  1540
12 4.50  1870
13 5.00  2072
14 5.50  2289
15 6.00  2562

> attach(vocabulary)
```

Fit linear model (a line) and store results in "mod".

```
> mod<-lm(Vocab~Age)
```

Summary output for mod.

```
> summary(mod)
Call:
lm(formula = Vocab ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max 
-249.67 -104.98   13.14   78.47  268.25 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -621.16      74.04  -8.389 1.32e-06 ***
Age           526.73      22.12  23.808 4.17e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1 

Residual standard error: 148 on 13 degrees of freedom
Multiple R-Squared: 0.9776,     Adjusted R-squared: 0.9759 
F-statistic: 566.8 on 1 and 13 DF,  p-value: 4.170e-12
```

Plot the data. Does a line look appropriate?

```
>  plot(Age,Vocab,ylim=c(-1000,3000))
> abline(mod)
```

Plot residuals vs predicteds. Is there a pattern?

```
> plot(mod$fitted.values,mod$residuals)
```

Boxplot residuals. Unusual points? Skewness?

```
> boxplot(mod$residuals)
```

Normal plot of residuals. Do the residuals look Normal? (Is it a line?)

```
> qqnorm(mod$residuals)
```

Test of the null hypothesis that the residuals are Normal.

```
> shapiro.test(mod$residuals)

        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.9801, p-value = 0.9703
```

General Linear Hypothesis

```
> help(anova.lm)

> attach(fuel)
> fuel[1:2,]
  ID state Fuel Tax License   Inc  Road
1  1    ME  541   9    52.5 3.571 1.976
2  2    NH  524   9    57.2 4.092 1.250
```

Fit the full model.

```
> mod<-lm(Fuel~Tax+License+Inc)
> anova(mod)
```

Optional step – for your education only.

```
Analysis of Variance Table

Response: Fuel
          Df Sum Sq Mean Sq F value    Pr(>F)
Tax        1 119823  119823  27.560 4.209e-06 ***
License    1 207709  207709  47.774 1.539e-08 ***
Inc        1  69532   69532  15.992 0.0002397 ***
Residuals 44 191302    4348
---
```

Fit the reduced model.

```
> mod2<-lm(Fuel~Tax)
> anova(mod2)
```

Optional step – for your education only.

```
Analysis of Variance Table

Response: Fuel
          Df Sum Sq Mean Sq F value   Pr(>F)
Tax        1 119823  119823  11.764 0.001285 **
Residuals 46 468543   10186
```

Compare the models

```
> anova(mod2,mod)
Analysis of Variance Table

Model 1: Fuel ~ Tax
Model 2: Fuel ~ Tax + License + Inc
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     46 468543
2     44 191302  2    277241 31.883 2.763e-09 ***
```

Notice the residual sum of squares and degrees of freedom in the three anova tables!

Polynomial Regression

```
> attach(cars)
```

Quadratic in size $y = \beta_0 + \beta_1 x + \beta_2 x^2$

```
> lm(mpg~size+I(size^2))

Call:
lm(formula = mpg ~ size + I(size^2))

Coefficients:
(Intercept)         size    I(size^2)
 39.3848313   -0.1485722    0.0002286
```

Centered quadratic in size $y = \beta_0 + \beta_1 x + \beta_2 \{x\text{-mean}(x)\}^2$

```
> lm(mpg~size+I((size-mean(size))^2))

Call:
lm(formula = mpg ~ size + I((size - mean(size))^2))

Coefficients:
             (Intercept)                       size  I((size - 
mean(size))^2)
              28.8129567                 -0.0502460
0.0002286
```

Orthogonal Polynomial Quadratic in size

```
> lm(mpg~poly(size,2))

Call:
lm(formula = mpg ~ poly(size, 2))

Coefficients:
   (Intercept)  poly(size, 2)1  poly(size, 2)2
         20.74          -24.67           12.33
```

To gain understanding:

- *do all there regressions*
- *look at t-test for* β_2
- *type poly(size,2)*
- *plot poly(size,2)[,1] and poly(size,2)[,2] against size*

Centered Polynomial with Interaction

```
> fuel[1:2,]
  ID state Fuel Tax License   Inc  Road
1  1    ME  541   9    52.5 3.571 1.976
2  2    NH  524   9    57.2 4.092 1.250

> attach(fuel)
```

Construct the squared and crossproduct terms. Alternatives: use "" or ":" in model formula.*

```
> TaxC<-Tax-mean(Tax)
> LicC<-License-mean(License)
> TaxC2<-TaxC*TaxC
> LicC2<-LicC*LicC
> TaxLicC<-TaxC*LicC

> modfull<-lm(Fuel~Tax+License+TaxC2+LicC2+TaxLicC)
> summary(modfull)

Call:
lm(formula = Fuel ~ Tax + License + TaxC2 + LicC2 + TaxLicC)
Residuals:
       Min         1Q     Median         3Q        Max
-121.52425  -51.08809   -0.01205   46.27051  223.28655

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 169.7242   179.6332   0.945   0.3501
Tax         -32.4465    12.2906  -2.640   0.0116 *
License      11.2776     2.3087   4.885 1.55e-05 ***
TaxC2         1.3171     8.6638   0.152   0.8799
LicC2         0.2575     0.2868   0.898   0.3743
TaxLicC      -2.5096     2.7343  -0.918   0.3640
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 76.42 on 42 degrees of freedom
Multiple R-Squared: 0.5831,     Adjusted R-squared: 0.5335
F-statistic: 11.75 on 5 and 42 DF,  p-value: 3.865e-07
```

Test whether the three squared and interaction terms are needed:

```
> modred<-lm(Fuel~Tax+License)
> anova(modred,modfull)
Analysis of Variance Table

Model 1: Fuel ~ Tax + License
Model 2: Fuel ~ Tax + License + TaxC2 + LicC2 + TaxLicC
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 260834
2     42 245279  3     15555 0.8879 0.4552
```

Understanding Linear Models with Interactions or Polynomials

*NIDA data (DC*MADS) on birth weight of babies in DC and attributes of mom.*

```
> DCBabyCig[1:2,]
    Age Married CIGS   BW
1    17       0    0 2385
2    23       1    0 4175
```

Age x Cigarettes interaction

```
> AC<-Age*CIGS
```

Model with interaction

```
> lm(BW~Age+CIGS+AC)

Call:
lm(formula = BW ~ Age + CIGS + AC)

Coefficients:
(Intercept)          Age         CIGS           AC
    2714.81        13.99       562.66       -28.04
```

How do you understand a model with interactions?

Let's create a new data.frame with 6 moms in it. Three moms are 18, three are 35. Some smoke 0, 1 or 2 packs.

```
> new[,1]<-c(18,35,18,35,18,35)
> new[,2]<-c(0,0,1,1,2,2)
> new[,3]<-new[,1]*new[,2]
> colnames(new)<-c("Age","CIGS","AC")
> new<-data.frame(new)

> new
  Age CIGS AC
1  18    0  0
2  35    1 35
3  18    0  0
4  35    1 35
5  18    0  0
6  35    1 35
```

Now, for these six moms, let's predict birth weight of junior. It is usually easier to talk about people than about coefficients, and that is what this table does: it talks about 6 moms.

```
> round(cbind(new,predict(lm(BW~Age+CIGS+AC),new,interval="confidence")))
  Age CIGS AC  fit  lwr  upr
1  18    0  0 2967 2865 3068
2  35    0  0 3204 3073 3336
3  18    1 18 3024 2719 3330
4  35    1 35 2786 2558 3013
5  18    2 36 3082 2474 3691
6  35    2 70 2367 1919 2814
```

Interpretation of an Interaction

```
> DCBabyCig[1:6,]
   Age Married CIGS   BW
1   17       0    0 2385
2   23       1    0 4175
3   25       0    0 3655
4   18       0    0 1855
5   20       0    0 3600
6   24       0    0 2820
```

Age = mother's age
Married, 1=yes, 0=no
CIGS = packs per day, 0, 1, 2.
BW = birth weight in grams

```
> dim(DCBabyCig)
[1] 449   4
```

```
> mod<-lm(BW~Age+Married+CIGS+I(Married*CIGS))
> summary(mod)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2973.1866   152.5467  19.490  < 2e-16 ***
Age                  0.1699     6.5387   0.026  0.97928
Married            274.0662    89.2913   3.069  0.00228 **
CIGS               -88.4957    81.7163  -1.083  0.27941
I(Married * CIGS) -415.1501   160.4540  -2.587  0.00999 **

Residual standard error: 687.8 on 444 degrees of freedom
Multiple R-squared: 0.05337,    Adjusted R-squared: 0.04484
F-statistic: 6.258 on 4 and 444 DF,  p-value: 6.618e-05
```

Plot the data

```
> boxplot(BW~Married:CIGS)
```

A 25 year old mom in all combinations of Married and CIGS.

```
> DCBabyCigInter
  Age Married CIGS
1  25       0    0
2  25       0    1
3  25       0    2
4  25       1    0
5  25       1    1
6  25       1    2
```

Predicted birth weights for this mom, with confidence intervals.

```
> predict(mod,DCBabyCigInter,interval="conf")
       fit      lwr      upr
1 2977.434 2890.180 3064.688
2 2888.938 2738.900 3038.977
3 2800.443 2502.124 3098.761
4 3251.500 3114.163 3388.838
5 2747.854 2476.364 3019.345
6 2244.209 1719.423 2768.995
```

Let’s clean it up, converting to pounds (2.2 pounds per kilogram), and add the predictors:

```
> pr<-predict(mod,DCBabyCigInter,interval="conf")

> round(cbind(DCBabyCigInter,2.2*pr/1000),1)
  Age Married CIGS fit lwr upr
1  25       0    0 6.6 6.4 6.7
2  25       0    1 6.4 6.0 6.7
3  25       0    2 6.2 5.5 6.8
4  25       1    0 7.2 6.9 7.5
5  25       1    1 6.0 5.4 6.6
6  25       1    2 4.9 3.8 6.1
```

Using Restricted Cubic Splines (aka Natural Splines)

```
> library(Hmisc)
> head(cars)
       car  size  mpg group
1  ToyotaC  71.1 33.9     1
2   HondaC  75.7 30.4     1

> x<-rcspline.eval(size,nk=3) #Three knots, one additional
variable
> plot(size,x) #What does the new variable look like?
> plot(size,mpg)
> m<-lm(mpg~size+x) #Add the new variable to the model.
> points(size,m$fit,pch=16,col="red") #What does the fit
look like?
> summary(m)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.81737    1.91493  20.271  < 2e-16 ***
size        -0.11667    0.01419  -8.225 7.86e-09 ***
x            0.14618    0.02642   5.533 7.29e-06 ***
Residual standard error: 2.346 on 27 degrees of freedom
Multiple R-squared:  0.8395,    Adjusted R-squared:  0.8276
F-statistic: 70.62 on 2 and 27 DF,  p-value: 1.877e-11

> x<-rcspline.eval(size,nk=5) #Five knots, three additional
variables
> m<-lm(mpg~size+x)
> summary(m)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.28095    3.61684  12.796 1.79e-12 ***
size        -0.19689    0.03655  -5.387 1.37e-05 ***
x1           2.17972    0.97629   2.233   0.0348 *
x2          -3.52907    1.75277  -2.013   0.0550 .
x3           1.45720    0.90924   1.603   0.1216
Residual standard error: 2.199 on 25 degrees of freedom
Multiple R-squared:  0.8694,    Adjusted R-squared:  0.8485
F-statistic: 41.61 on 4 and 25 DF,  p-value: 1.054e-10

> points(size,m$fit,pch=16,col="purple")
```

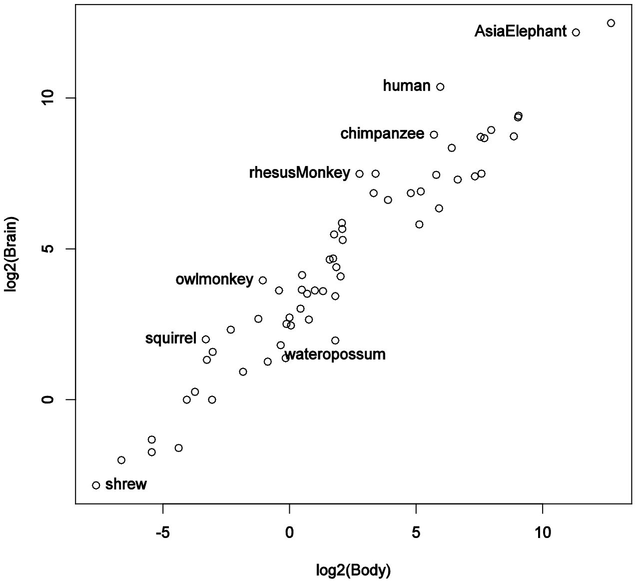
Reference: Harrell, F. (2015) *Regression Modeling Strategies*, New York: Springer, section 2.4.5.

Comment: There are many types of splines. Natural splines are linear beyond the final knots, so they wiggle less at the ends.

Dummy Variable in Brains Data

First two rows of "brains" data.

```
> brains[1:2,]
   Body  Brain    Animal Primate Human
1 3.385 44.500  articfox       0     0
2 0.480 15.499 owlmonkey       1     0

> attach(brains)

> plot(log2(Body),log2(Brain))
> identify(log2(Body),log2(Brain),labels=Animal)
```



```
> mod<-lm(log2(Brain)~log2(Body)+Primate)
> mod

Call:
lm(formula = log2(Brain) ~ log2(Body) + Primate)

Coefficients:
(Intercept)   log2(Body)      Primate
     2.8394       0.7402       1.6280
```

log2(Brain) ~ log2(Body) + Primate

is $2^{Brain} = 2^{\{\alpha + \square log2(Body) + \square Primate + \square\}} = (2^{\alpha})(Body^{\square})(2^{\square Primate})(2^{\square})$

$2^{1.628 Primate}$ = 3.1 for a primate, = 1 for a nonprimate

Computing the Diagnostics in the Rat Data

```
> ratdata[1:3,]
  BodyWgt LiverWgt Dose Percent Rat3
1     176      6.5 0.88    0.42    0
2     176      9.5 0.88    0.25    0
3     190      9.0 1.00    0.56    1

> attach(ratdata)
> mod<-lm(Percent~BodyWgt+LiverWgt+Dose)
```

Standardized residuals (first 5)

```
> rstandard(mod)[1:5]
        1         2         3         4         5 
 1.766047 -1.273040  0.807154 -1.377232 -1.123099 
```

Deleted or jackknife or "studentized" residuals (first 5)

```
> rstudent(mod)[1:5]
         1          2          3          4          5 
 1.9170719 -1.3022313  0.7972915 -1.4235804 -1.1337306 
```

dffits (first 5)

```
> dffits(mod)[1:5]
         1          2          3          4          5 
 0.8920451 -0.6087606  1.9047699 -0.4943610 -0.9094531 
```

Cook's distance (first 5)

```
> cooks.distance(mod)[1:5]
         1          2          3          4          5 
0.16882682 0.08854024 0.92961596 0.05718456 0.20291617 
```

Leverages or 'hats' (first 5)

```
> hatvalues(mod)[1:5]
        1         2         3         4         5 
0.1779827 0.1793410 0.8509146 0.1076158 0.3915382 

> dfbeta(mod)[1:3,]
   (Intercept)       BodyWgt     LiverWgt       Dose
1 -0.006874698  0.0023134055 -0.011171761 -0.3419002
2  0.027118946 -0.0007619302 -0.008108905  0.1869729
3 -0.045505614 -0.0134632770  0.005308722  2.6932347

> dfbetas(mod)[1:3,]
  (Intercept)     BodyWgt   LiverWgt       Dose
1 -0.03835128  0.31491627 -0.7043633 -0.2437488
2  0.14256373 -0.09773917 -0.4817784  0.1256122
3 -0.23100202 -1.66770314  0.3045718  1.7471972
```

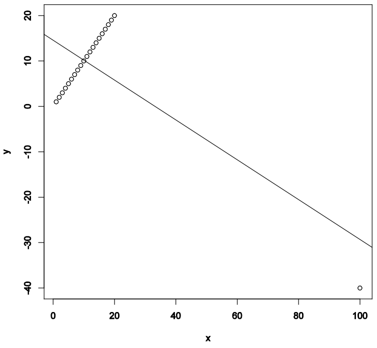
High Leverage Example

```
> t(highlev)
```

t() is transpose – make rows into columns and columns into rows – compact printing

```
  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  21
x 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 100
y 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -40

> mod<-lm(y~x)
> summary(mod)
Residuals:
     Min       1Q   Median       3Q      Max
-13.1343  -7.3790  -0.1849   7.0092  14.2034
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.57312    2.41264   6.040 8.24e-06 ***
x           -0.43882    0.09746  -4.503 0.000244 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.875 on 19 degrees of freedom
Multiple R-Squared: 0.5162,     Adjusted R-squared: 0.4908
F-statistic: 20.27 on 1 and 19 DF,  p-value: 0.0002437

> plot(x,y)
> abline(mod)
```

Puts the fitted line on the plot — What a dumb model!



The bad guy, #21, doesn't have the biggest residual!

```
> mod$residual[21]
       21
-10.69070
```

Residuals:

```
     Min       1Q   Median       3Q      Max
-13.1343  -7.3790  -0.1849   7.0092  14.2034
```

But our diagnostics find him!

```
> rstudent(mod)[21]
        21
-125137800
> hatvalues(mod)[21]
       21
0.9236378
> dffits(mod)[21]
        21
-435211362
```

Outlier Testing

Use the Bonferroni inequality with the deleted/jackknife/"studentized" residuals.

Example uses random data – should not contain true outliers

```
> x<-rnorm(1000)
> y<-rnorm(1000)
> plot(x,y)
> summary(lm(y~x))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.005469   0.031685  -0.173    0.863
x           -0.044202   0.031877  -1.387    0.166

Residual standard error: 1.002 on 998 degrees of freedom
Multiple R-Squared: 0.001923,   Adjusted R-squared: 0.0009228
F-statistic: 1.923 on 1 and 998 DF,  p-value: 0.1659
```

Look at the deleted residuals (from rstudent)

```
> summary(rstudent(lm(y~x)))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-3.189e+00 -6.596e-01  2.186e-02 -7.077e-06  6.517e-01  3.457e+00
```

The big residual in absolute value is 3.457. Is that big for the biggest of 1000 residuals?

The pt(value,df) command looks up value in the t-table with df degrees of freedom, returning Pr(t<value). You need the other tail, Pr(t>value), and you need to double it for a 2-tailed test. The degrees of freedom are ***one less than the degrees of freedom in the error for the regression****, here 997.*

```
> 2*(1-pt(3.457,997))
[1] 0.0005692793
```

This is uncorrected p-value. Multiply by the number of tests, here 1000, to correct for multiple testing. (It's an inequality, so it can give a value bigger than 1.)

```
> 1000* 2*(1-pt(3.457,997))
[1] 0.5692793
```

As this is bigger than 0.05, the null hypothesis of no outliers is not rejected – it is plausible there are no outliers present.

IS WYOMING AN OUTLIER?

```
> attach(fuel)
> dim(fuel)
[1] 48  7
> mod<-lm(Fuel~Tax+License)
> which.max(abs(rstudent(mod)))
40
40
> fuel[40,]
   ID state Fuel Tax License   Inc  Road
40 40    WY  968   7    67.2 4.345 3.905
> rstudent(mod)[40]
      40
3.816379
> wy<-rep(0,48)
> wy[40]<-1
> wy
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 0

> summary(lm(Fuel~Tax+License+wy))

Call:
lm(formula = Fuel ~ Tax + License + wy)

Residuals:
     Min       1Q   Median       3Q      Max
-122.786  -55.294    1.728   46.621  154.557

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  198.651    152.405   1.303  0.19920
Tax          -30.933     10.696  -2.892  0.00593 **
License       10.691      1.894   5.645 1.12e-06 ***
wy           267.433     70.075   3.816  0.00042 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.74 on 44 degrees of freedom
Multiple R-Squared: 0.6669,     Adjusted R-squared: 0.6442
F-statistic: 29.37 on 3 and 44 DF,  p-value: 1.391e-10

> 0.05/48
[1] 0.001041667
> 0.00042<=0.001041667
[1] TRUE
```

Testing Whether a Transformation of Y is Needed

Tukey’s One Degree of Freedom for Nonadditivity

Tukey (1949) proposed testing whether a transformation of y is needed in the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \varepsilon \quad \varepsilon \sim \text{iid } N(0,\sigma^2)$$

by adding a scaled centered version of $\hat{y}^2$ to the model, specifically $\frac{(\hat{y}-\bar{y})^2}{2\bar{y}}$; see Atkinson (1985, p. 157). The function `tukey1df(mod)` in the class workspace does this, but you could easily do it yourself.

```
> mod<-lm(BW~Age+Married+CIGS)
> summary(mod)

Call:
lm(formula = BW ~ Age + Married + CIGS)

Residuals:
     Min       1Q   Median       3Q      Max 
-2408.30  -358.49    99.69   453.34  1952.79 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2936.618    152.859  19.211  < 2e-16 ***
Age            2.557      6.515   0.392  0.69488    
Married      200.615     85.198   2.355  0.01897 *  
CIGS        -196.644     70.665  -2.783  0.00562 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 692.2 on 445 degrees of freedom
Multiple R-squared: 0.0391,     Adjusted R-squared: 0.03262 
F-statistic: 6.036 on 3 and 445 DF,  p-value: 0.0004910 

> boxplot(mod$resid)
> qqnorm(mod$resid)
> shapiro.test(mod$resid)

        Shapiro-Wilk normality test

data:  mod$resid 
W = 0.9553, p-value = 1.996e-10

> plot(mod$fit,mod$resid)
> lines(lowess(mod$fit,mod$resid))
```

To do the test, add the transformed variable to the model and look at its t-statistic.

```
> summary(lm(BW~Age+Married+CIGS+tukey1df(mod)))

Call:
lm(formula = BW ~ Age + Married + CIGS + tukey1df(mod))

Residuals:
    Min      1Q  Median      3Q     Max 
-2301.3  -334.5   107.7   420.8  1981.2 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2962.751    151.775  19.521  < 2e-16 ***
Age              0.508      6.494   0.078  0.93769    
Married        104.750     90.366   1.159  0.24701    
CIGS          -489.699    120.693  -4.057 5.86e-05 ***
tukey1df(mod)   31.507     10.567   2.982  0.00302 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 686.2 on 444 degrees of freedom
Multiple R-squared: 0.05796,    Adjusted R-squared: 0.04947 
F-statistic:  6.83 on 4 and 444 DF,  p-value: 2.428e-05
```

Box – Cox Method

An alternative approach is due to Box and Cox (1964).

```
> library(MASS)
> help(boxcox)
> boxcox(mod)
```

Andrews, D. F. (1971) A note on the selection of data transformations. Biometrika, 58, 249-254.
Atkinson, A. C. (1985) Plots, Transformations and Regression. NY: Oxford.
Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). Journal of the Royal Statistical Society B, 26, 211–252.
Tukey, J. W. (1949) One degree of freedom for nonadditivity. *Biometrics*, 5, 232-242.
Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

Tukey Tests in the `car` Package in `R`

- The function `residualPlots()` in the `car` package automates some checks for nonlinearity using the general form of Tukey's test for nonadditivity.

- If `m` is a linear model, then type `residualPlots(m)`

- For example, using the cars data:

```
> library(car)
> residualPlots(lm(mpg~size))
           Test stat Pr(>|t|)
size            4.99        0
Tukey test      4.99        0
```

- Because there is only one x in the cars data, there is only one test. It asks whether a quadratic in size would predict the residuals. The t-statistic is 4.99 with a tiny P-value, so the answer is yes.

- For example, using the fuel data:

```
> attach(fuel)
> residualPlots(lm(Fuel~Tax+License))
           Test stat Pr(>|t|)
Tax            0.202    0.841
License        1.321    0.193
Tukey test     1.633    0.103
```

- Now there are 3 tests. One is about whether $\widehat{y}^2$ predictes the residuals, the "Tukey test". Another is whether Tax^2 predicts the residuals. Another is whether License^2 predicts the residuals. In all three cases, the t-statistic is not large and the P-value is large, so it is plausible that no quadratic term is needed.

Transformations in the car package

```
> attach(cars)
> plot(size,mpg)
> library(car)
```

This is looking for a transformation of size that would improve the fit. Because y=mpg is not transformed, only x=size, the procedure can compare residual sums of squares of y (RSS) for different transformations of x. Here, lambda is the power transformation, what we called p in class.

```
> invTranPlot(mpg~size)
     lambda      RSS
1 -1.262328 123.4974
2 -1.000000 126.3604
3  0.000000 192.1809
4  1.000000 317.0362
```

It likes the -1.26 power of x as a transformation, but -1 and -1.5 are in the confidence interval.

```
> invTranEstimate(size,mpg)
$lambda
[1] -1.262328
$lowerCI
[1] -1.737292
$upperCI
[1] -0.8237128
```

This is trying to transform y=mpg, not x=size. It uses the Box-Cox likelihood method. It likes y^(-0.99) or approximately 1/y.

```
> summary(powerTransform(lm(mpg~size)))
bcPower Transformation to Normality
   Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Y1   -0.9878   0.5553          -2.0763           0.1007
Likelihood ratio tests about transformation parameters
                            LRT df         pval
LR test, lambda = (0)  3.194518  1 0.0738855550
LR test, lambda = (1) 12.478121  1 0.0004117462
```

A graphical version.

```
> boxCox(lm(mpg~size))
```

Understanding invTranPlot in the car Package

```
> x<-1:100
> y<-100+100/x+rnorm(100)

# So, truth is y = α + β/x + ε = 100 + 100/x + ε =
# with ε ~ N(0,1), α = 100, β = 100.  y is linear in 1/x.

> plot(x,y)  # Looks curved
> library(car) # Companion to Applied Regression or car package

> invTranPlot(x,y) # Tries out 4 transformations of x
     lambda         RSS
1 -1.000665    95.42204
2 -1.000000    95.42420
3  0.000000  5635.08235
4  1.000000 10589.20354

# It is just fitting linear regressions, with x transformed
# and looking at the RSS = residual sum of squares
> anova(lm(y~x))
Analysis of Variance Table  Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1  3110.5 3110.53  28.787 5.406e-07 ***
Residuals 98 10589.2  108.05

> logx<-log(x)
> anova(lm(y~logx))
Analysis of Variance Table  Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
logx       1 8064.7  8064.7  140.25 < 2.2e-16 ***
Residuals 98 5635.1    57.5

> rx<-1/x
> anova(lm(y~rx))
Analysis of Variance Table  Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
rx         1 13604.3   13604   13972 < 2.2e-16 ***
Residuals 98    95.4       1


#So the fit is vastly better, RSS is vastly smaller, with 1/x than
#with x or log(x).
```

```
# You can try your own lambdas, say lambda = -1/2

> invTranPlot(x, y, lambda=c(-1/2))
     lambda        RSS
1 -1.000665   95.42204
2 -0.500000 1633.89693


> rxhalf<-x^(-1/2)
> anova(lm(y~rxhalf))
Analysis of Variance Table   Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
rxhalf     1 12065.8 12065.8   723.7 < 2.2e-16 ***
Residuals 98  1633.9    16.7

# It will also give you a confidence interval for lambda

> invTranEstimate(x,y)
$lambda
[1] -1.000665
$lowerCI
[1] -1.020736
$upperCI
[1] -0.9809275

# So it is not sure it is 1, but the interval is narrow around 1.
```

The car package has an alternative to Tukey’s test for a transformation.

Atkinson’ method

```
> v<-boxCoxVariable(mpg)
> summary(lm(mpg~size+v))
Call:
lm(formula = mpg ~ size + v)
Residuals:
    Min      1Q  Median      3Q     Max 
-4.3645 -1.3910  0.0462  1.0028  6.4195 
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 76.651791   9.132471   8.393 5.27e-09 ***
size        -0.033455   0.004324  -7.737 2.55e-08 ***
v            2.519663   0.486271   5.182 1.87e-05 ***
```

Andrews/Tukey method

```
> summary(lm(mpg~size+tukey1df(md)))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.812957   1.008748   28.56  < 2e-16 ***
size         -0.050246   0.004508  -11.15 1.32e-11 ***
tukey1df(md)  5.577240   1.117584    4.99 3.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 2.471 on 27 degrees of freedom
Multiple R-squared:  0.8218,    Adjusted R-squared:  0.8087 
F-statistic: 62.28 on 2 and 27 DF,  p-value: 7.686e-11
```

References:

For Tukey’s method and related methods

Andrews, D. F. (1971) A note on the selection of data transformations. *Biometrika* 58, 249-254.

For the method in the car package:

Atkinson, A. C. (1973) Testing transformations to Normality. *JRSS-B* 473-479.

Checking for Non-constant Variance

The usual linear model assumes that the errors are independent and identically distributed (iid) with a Normal distribution having expectation 0 and constant variance σ^2.

Is there evidence in the data that the variance is not constant?
There is a graphical aid and a test.

In the car data, the relationship between mile-per-gallon and engine size was curved.

```
> attach(cars)
> head(cars)
       car  size  mpg group
1  ToyotaC  71.1 33.9     1
2   HondaC  75.7 30.4     1
3  Fiat128  78.7 32.4     1
4  FiatX19  79.0 27.3     1
5 LotusEur  95.1 30.4     1
6   Datsun 108.0 22.8     1
> plot(size,mpg)
```

In the car data, our first attempt to straighten the curve was to raise mpg to the -4.5 power. That fixed one problem, but created another: is was straight, but the plot of residuals versus predicted had a fan shape, suggesting non-constant variance.

```
> tmpg<-mpg^(-4.5)
> plot(size,tmpg)
```

The third plot you get from plot(model) is a substitute for looking for a fan shape. It plots the square root of the absolute value of the standardized residuals against the fitted values. A trend, up or down, in this plot suggests non-constant variance.

```
> m<-lm(tmpg~size)
> plot(m)
```

The third plot you get is the plot for non-constant variance. In this case, you see a clear trend, indicating non-constant variance. You could make this plot yourself by typing:

```
> v<-sqrt(abs(rstandard(m)))
> plot(m$fit,v)
```

Why is this better than plotting residuals versus predicted, the first plot you get from plot(model)? There are three reasons:

(i) Even when the variance of the errors is constant, the usual residuals, m$residual, do not have constant variance. However, in this case, the standardized residuals, rstandard(m), do have constant variance.

(ii) By taking absolute values, you make the standardized residuals positive, so instead of looking for a fan, you are looking for a trend, up or down. It is often easier to see the trend.

(iii) The absolute value of iid Normal data tends to be skewed right. This is distracting. Taking the square root removes much of skew.

In the car package, there is a test for non-constant variance. The null hypothesis is that the variance is constant, and a small P-value raises doubt about the null hypothesis, suggesting the variance changes with one or more of the predictors.

```
> library(car)
> m<-lm(tmpg~size)
> ncvTest(m)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 20.54211    Df = 1     p = 5.833364e-06
```

This test is similar to, but not quite the same as, regressing the squared residuals on the predicted values from the model. The residuals themselves are uncorrelated with the predicted values. Can you predict the squared residuals?

The test is due to Breusch and Pagan (1979) and Cook and Weisberg (1983).

It is easy to confuse an outlier, non-constant variance, and nonlinearity. A test of each might find all three. Graphs help you recognize which one is actually present.

Cook, R. D. and Weisberg, S. (1983) Diagnostics for heteroscedasticity in regression. *Biometrika* 70, 1-10.

Calculating C_P for the Cathedral Data

```
> attach(cathedral)
> mod<-lm(length~height+gothic+GH)
> summary(mod)
Call:
lm(formula = length ~ height + gothic + GH)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  241.833    336.471   0.719    0.480
height         3.138      4.506   0.696    0.494
gothic      -204.722    347.207  -0.590    0.562
GH             1.669      4.641   0.360    0.723
Residual standard error: 79.11 on 21 degrees of freedom
Multiple R-Squared: 0.5412,     Adjusted R-squared: 0.4757
F-statistic: 8.257 on 3 and 21 DF,  p-value: 0.0008072

> drop1(lm(length~height+gothic+GH),scale=79.11^2)
Single term deletions

Model:
length ~ height + gothic + GH

scale:  6258.392

       Df Sum of Sq    RSS     Cp
<none>              131413 3.9979
height  1      3035 134448 2.4829
gothic  1      2176 133589 2.3455
GH      1       810 132223 2.1273

> drop1(lm(length~height+gothic),scale=79.11^2)
Single term deletions
Model:
length ~ height + gothic

scale:  6258.392

       Df Sum of Sq    RSS      Cp
<none>              132223  2.1273
height  1    119103 251326 19.1582
gothic  1     37217 169440  6.0740
```

Variable Selection

Highway data. First two rows of 39 rows. More details in the Variable Selection section of this bulkpack.

```
> highway[1:2,]
  ID rate   len adt trks slim lwid shld  itg sigs acpt lane fai pa ma
1  1 4.58  4.99  69    8   55   12   10 1.20    0  4.6    8   1  0  0
2  2 2.86 16.11  73    8   60   12   10 1.43    0  4.4    4   1  0  0
```

Highway data has 39 rows, 15 columns, of which y=rate, and columsn 3 to 15 or 3:15 are predictors. Want to select predictors.

```
> dim(highway)
[1] 39 15

> attach(highway)
```

To use "leaps" for best subsets regression, need to get it from the library. To get documentation, type help!

```
> library(leaps)
> help(leaps)
```

Easiest if you put the x's in a separate variable. These are columns 3:15 of highway, including all the rows.

```
> x<-highway[,3:15]
```

First three rows of 39 rows of x. Notices that the first two columns of highway are gone.

```
> x[1:3,]
    len adt trks slim lwid shld  itg sigs acpt lane fai pa ma
1  4.99  69    8   55   12   10 1.20    0  4.6    8   1  0  0
2 16.11  73    8   60   12   10 1.43    0  4.4    4   1  0  0
3  9.75  49   10   60   12   10 1.54    0  4.7    4   1  0  0
```

There are 13 predictors, hence 2^{13} = 8,192 possible models formed by including each variable or not.

```
> dim(x)
[1] 39 13
> 2^13
[1] 8192
```

Look at the names of your predictors: len = length of segment, ..., slim = speed limit, ..., acpt = number of access points per mile, ...

```
> colnames(x)
 [1] "len"  "adt"  "trks" "slim" "lwid" "shld" "itg"  "sigs" "acpt"
"lane" "fai"  "pa"   "ma"
```

A quick and easy, but not very complete, answer is obtained from regsubsets. Here, it gives the best model with 1 variable, the best with 2 variables, etc. Look for the *'s. The best 3 variable model is len, slim, acpt.

```
> summary(regsubsets(x=x,y=rate))
1 subsets of each size up to 8  Selection Algorithm: exhaustive
         len adt trks slim lwid shld itg sigs acpt lane fai pa  ma
1  ( 1 ) " " " " " "  " "  " "  " "  " " " "  "*"  " "  " " " " " "
2  ( 1 ) "*" " " " "  " "  " "  " "  " " " "  "*"  " "  " " " " " "
3  ( 1 ) "*" " " " "  "*"  " "  " "  " " " "  "*"  " "  " " " " " "
4  ( 1 ) "*" " " " "  "*"  " "  " "  " " "*"  "*"  " "  " " " " " "
5  ( 1 ) "*" " " " "  "*"  " "  " "  " " "*"  "*"  " "  " " "*" " "
6  ( 1 ) "*" " " "*"  "*"  " "  " "  " " "*"  "*"  " "  " " "*" " "
7  ( 1 ) "*" " " "*"  "*"  " "  " "  " " "*"  "*"  " "  " " "*" "*"
8  ( 1 ) "*" " " "*"  "*"  " "  " "  "*" "*"  "*"  " "  " " "*" "*"
```

To get the two best models of each size, type:

```
> summary(regsubsets(x=x,y=rate,nbest=2))
```

Variable Selection, Continued

Leaps does "best subsets regression". Here, x contains the predictors of y=rate. If you don't tell it the variable names, it uses 1, 2, …

```
mod<-leaps(x,rate,names=colnames(x))
```

The output from leaps, here "mod", has four parts, "which", "label", "size" and "Cp".

```
> summary(mod)
      Length Class  Mode
which 1573   -none- logical
label   14   -none- character
size   121   -none- numeric
Cp     121   -none- numeric
```

You refer to "which" for "mod" as "mod$which", etc. Also mod$size, mod$Cp.

The part, mod$which says which variables are in each model. It reports back about 121 models, and all 13 variables.

```
> dim(mod$which)
[1] 121  13
```

Here are the first 3 rows or models in mod$which. The first model has only acpt, while the second has only slim.

```
> mod$which[1:3,]
    len   adt  trks  slim  lwid  shld   itg  sigs  acpt  lane   fai    pa    ma
1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
```

Here are the last 3 rows or models in mod$which. The last model has all 13 variables.

```
> mod$which[119:121,]
    len  adt trks slim lwid shld  itg  sigs acpt lane  fai    pa   ma
12 TRUE TRUE TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE FALSE TRUE
12 TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE  TRUE TRUE
13 TRUE TRUE TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE
```

Here are the sizes of the 121 models. A 1-variable model has size 2, for constant-plus-one-slope. A 2-variable model has size 3. The final model, #121, with all 13 variables has size 14.

```
> mod$size
  [1]  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  4  4  4  4
 [25]  4  4  4  4  4  4  5  5  5  5  5  5  5  5  5  5  6  6  6  6  6  6  6  6
 [49]  6  6  7  7  7  7  7  7  7  7  7  7  8  8  8  8  8  8  8  8  8  8  9  9
 [73]  9  9  9  9  9  9  9  9 10 10 10 10 10 10 10 10 10 10 11 11 11 11 11 11
 [97] 11 11 11 11 12 12 12 12 12 12 12 12 12 12 13 13 13 13 13 13 13 13 13 13
[121] 14
```

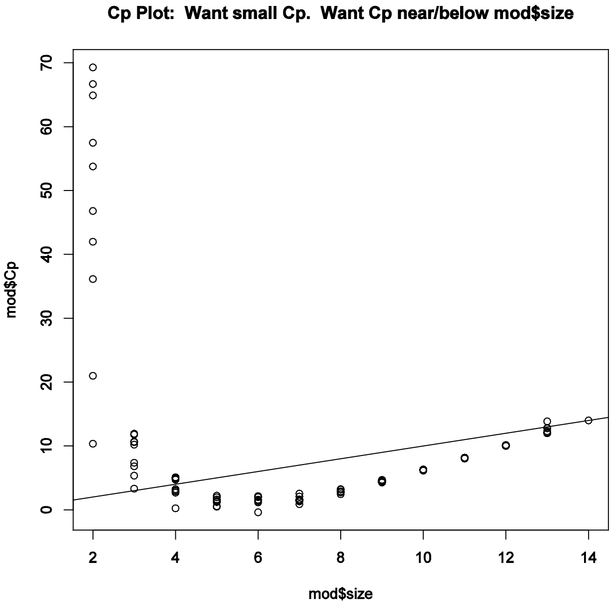
These are the C_p values for the 121 models.

```
> round(mod$Cp,2)
  [1] 10.36 20.98 36.13 41.97 46.79 53.77 57.48 64.90 66.67 69.28  3.31  5.35
 [13]  6.82  7.37 10.21 10.59 10.64 10.68 11.78 11.93  0.24  2.68  2.90  3.04
 [25]  3.24  4.72  4.88  4.99  5.05  5.10  0.49  0.56  1.21  1.26  1.33  1.39
 [37]  1.59  1.63  1.97  2.22 -0.38  1.17  1.34  1.51  1.60  2.00  2.00  2.02
 [49]  2.05  2.15  0.88  1.33  1.33  1.52  1.54  1.58  1.61  1.62  2.08  2.54
 [61]  2.45  2.73  2.74  2.82  2.82  2.87  2.88  3.18  3.19  3.23  4.27  4.33
 [73]  4.36  4.43  4.45  4.45  4.45  4.59  4.67  4.69  6.14  6.15  6.16  6.23
 [85]  6.24  6.26  6.26  6.30  6.32  6.32  8.02  8.07  8.12  8.12  8.14  8.14
 [97]  8.14  8.15  8.16  8.17 10.02 10.02 10.02 10.06 10.07 10.10 10.12 10.12
[109] 10.13 10.14 12.01 12.01 12.01 12.05 12.10 12.14 12.32 12.76 12.83 13.85
[121] 14.00
```

Variable Selection, Continued

This is the C_p plot.

```
> plot(mod$size,mod$Cp)
> abline(0,1)
```



There is one pretty good 2 variable model (size=3), with C_p near the x=y line, and one very good 3 variable model (size=4), with C_p way below the line. The best model has 5 variables (size =6) but is only trivially better than the 3 variable model. R^2 is highest for the 14 variable model, but chances are it won't predict as well as the 3 variable model.

Let's put together the pieces.

```
> join<-cbind(mod$which,mod$Cp,mod$size)
```

Let's look at the 3 variable models (size=4). The best has C_p = 0.236 and variables len, slim, and acpt.

```
> join[mod$size==4,]
  len adt trks slim lwid shld itg sigs acpt lane fai pa ma
3   1   0    0    1    0    0   0    0    1    0   0  0  0 0.2356971 4
3   0   0    1    1    0    0   0    0    1    0   0  0  0 2.6805672 4
3   1   0    0    0    0    0   0    1    1    0   0  0  0 2.8975068 4
3   1   0    0    0    0    1   0    0    1    0   0  0  0 3.0404482 4
3   1   0    1    0    0    0   0    0    1    0   0  0  0 3.2366902 4
3   1   0    0    0    1    0   0    0    1    0   0  0  0 4.7193511 4
3   0   0    0    1    0    0   0    1    1    0   0  0  0 4.8847460 4
3   1   0    0    0    0    0   0    0    1    0   0  1  0 4.9933327 4
3   1   0    0    0    0    0   0    0    1    1   0  0  0 5.0489720 4
3   1   1    0    0    0    0   0    0    1    0   0  0  0 5.1013513 4
```

The full model has C_p = 14.

```
> join[mod$size==14,]
 len  adt trks slim lwid shld  itg sigs acpt lane  fai   pa   ma
   1    1    1    1    1    1    1    1    1    1    1    1    1   14   14
```

Cp thinks that the 14 variable model will have squared errors 59 times greater than the 3 variable model with len, slim, and acpt.

```
> 14/0.2356971
[1] 59.39827
```

A key problem is that variable selection procedures overfit. Need to cross-validate!

Variable Selection, Continued
(O2Uptake Example)

Load libraries

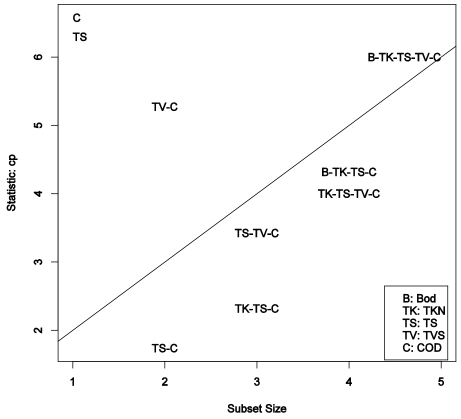
```
> library(leaps)
> library(car)

> O2Uptake[1:3,]
  Day  Bod TKN   TS  TVS  COD O2UP LogO2Up
1   0 1125 232 7160 85.9 8905 36.0  1.5563
2   7  920 268 8804 86.5 7388  7.9  0.8976
3  15  835 271 8108 85.2 5348  5.6  0.7482

> dim(O2Uptake)
[1] 20  8
```

Find best 2 models of each size.

```
> mod<-regsubsets(x=O2Uptake[,2:6],y=O2Uptake$LogO2Up,nbest=2)
> summary(mod)
2 subsets of each size up to 5
Selection Algorithm: exhaustive
         Bod TKN TS  TVS COD
1  ( 1 ) " " " " "*" " " " "
1  ( 2 ) " " " " " " " " "*"
2  ( 1 ) " " " " "*" " " "*"
2  ( 2 ) " " " " " " "*" "*"
3  ( 1 ) " " "*" "*" " " "*"
3  ( 2 ) " " " " "*" "*" "*"
4  ( 1 ) " " "*" "*" "*" "*"
4  ( 2 ) "*" "*" "*" " " "*"
5  ( 1 ) "*" "*" "*" "*" "*"
```

C_p plot

```
> subsets(mod,stat="cp")
> abline(1,1)
```



PRESS (and writing little programs in R)

We have seen many times in many ways that ordinary residuals, say E_i, tend to be too small, because Y_i was used in fitting the model, so the model is too close to Y_i. Predicting Y_i having fitted the model using Y_i is called "in-sample-prediction," and it tends to suggest that a model is better than it is, because it saying you are making progress getting close to your current Y_i's, even if you could not do well in predicting a new Y_i.

If you left i out of the regression, and tried to predict Y_i from the regression without i, the error you would make is:

$$Y_i - \hat{Y}_{i[i]} = V_i \text{ say.}$$

Here, V_i is an "out-of-sample prediction," a true effort to predict a "new" observation, because i did not get used in fitting this equation. It gives me a fair idea as to how well a model can predict an observation not used in fitting the model.

The predicted error sum of squares or PRESS is

$$\text{PRESS} = \square\ V_i^2.$$

It turns out that $V_i = E_i/(1-h_i)$ where E_i, is the residual and h_i is the leverage or hatvalue.

```
> fuel[1:2,]
  ID state Fuel Tax License   Inc  Road
1  1    ME  541   9    52.5 3.571 1.976
2  2    NH  524   9    57.2 4.092 1.250
> attach(fuel)
> modMAX<-lm(Fuel~Tax+License+Inc+Road)
```

These are the out of sample prediction errors or V_i's:

```
> V<-modMAX$residual/(1-hatvalues(modMAX))
```

Let's look at Wyoming, WY. It's residual is about 235 gallons:

```
> modMAX$residual[state=="WY"]
      40
234.9472
```

but it's out of sample prediction error is about 26 gallons larger:

```
> V[state=="WY"]
      40
260.9721
```

PRESS (and writing little programs in R), continued

PRESS is the sum of the squares of the V_i

```
> sum(V^2)
[1] 235401.1
```

How does PRESS compare to R^2? Well R^2 is an in-sample measure, while press is an out-of-sample measure. For modMAX, R^2 is:

```
> summary(modMAX)$r.squared
[1] 0.6786867
```

Let's take Road out of the model, and see what happens to R^2 and PRESS.

```
> modSmall<-lm(Fuel~Tax+License+Inc)

> summary(modSmall)$r.squared
[1] 0.6748583
```

So R^2 went down, "got worse," which it always does when you delete variables;

```
> Vsmall<-modSmall$residual/(1-hatvalues(modSmall))
> sum(Vsmall^2)
[1] 229998.9
```

however, PRESS went down too, or "got better." In other words, adding Road to the model makes the residuals smaller, as adding variables always does, but it makes the prediction errors bigger. Sometimes adding a variable makes prediction errors smaller, sometimes it makes them bigger, and PRESS tells which is true in your model.

You could compute PRESS as above each time you fit a model, but it is easier to add a little program to R. Here is how you write a program called PRESS that computes PRESS.

```
> PRESS<-function(mod){
+ V<-mod$residual/(1-hatvalues(mod))
+ sum(V^2)}
```

If you type in the name of your program, here PRESS, it prints the program for you to look at.

```
> PRESS
function(mod){
     V<-mod$residual/(1-hatvalues(mod))
     sum(V^2)}
```

Your new program will compute PRESS for you:

```
> PRESS(modMAX)
[1] 235401.1
> PRESS(modSmall)
[1] 229998.9
```

Variance Inflation Factor (VIF)

Need library DAAG. You may have to install it the first time.

```
> library(DAAG)

> fuel[1:2,]
  ID state Fuel Tax License   Inc  Road
1  1    ME  541   9    52.5 3.571 1.976
2  2    NH  524   9    57.2 4.092 1.250

> attach(fuel)
```

Run a regression, saving results.

```
> mod<-lm(Fuel~Tax+License+Inc+Road)
```

Here are the VIF's

```
> vif(mod)
    Tax License     Inc    Road
 1.6257  1.2164  1.0433  1.4969
```

You can convert the VIF's to R^2

```
> 1-1/vif(mod)
       Tax    License        Inc       Road
0.38488036 0.17790201 0.04150292 0.33195270
```

This says: If you predict Tax from License, Inc and Road, the R^2 *is 0.3849. You could do the regression and get the same answer; see below.*

```
> summary(lm(Tax~License+Inc+Road))
Call:
lm(formula = Tax ~ License + Inc + Road)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.14672    1.35167   8.247 1.79e-10 ***
License     -0.05791    0.02058  -2.814  0.00728 **
Inc          0.15455    0.19881   0.777  0.44109
Road        -0.14935    0.03232  -4.621 3.34e-05 ***

Residual standard error: 0.7707 on 44 degrees of freedom
Multiple R-Squared: 0.3849,     Adjusted R-squared: 0.3429
F-statistic: 9.177 on 3 and 44 DF,  p-value: 7.857e-05
```

Spjotvoll's Method in Variable Selection

The maximum model has variables {1,2,…,k}, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$ where the errors ε are independent and $N(0,\sigma^2)$. Suppose that T is the true model, where T is a subset of {1,2,…,k}. That is, T is the model with exactly the nonzero β_js, so $\beta_j=0$ if and only if j is not in T. We could do a general linear hypothesis F-test to test model T against the maximum model, and if we did just this one test, the chance that we would falsely reject model T at level $\alpha=0.05$ would be 5%. The problem is that we don't know T, so we end up testing many models, and might make many mistakes in all those tests.

Spjotvoll (1977) defines an inadequate model as any model that omits a variable in T, and an adequate model as any model that includes all of the variables in T, perhaps including some extra variables whose coefficients are 0. If k=5 and T={1,2}, then {1,2} and {1,2,3} are adequate models, but {1,3} and {2,3,4,5} are inadequate models. Spjotvoll wants to reject some models as inadequate. He is not worried about having too many variables, and is only worried about omitting a needed variable.

Spjotvoll (1977) declares a model Q to be inadequate at level $\alpha=0.05$ if and only if the F-test rejects both Q and every model contained in Q. With k=5, to reject Q={1,3} as inadequate, you would have to reject Q={1,3}, and also {1}, {3} and {}, where {} is the model with no variables, that is, {} is $y = \beta_0 + \varepsilon$. So to reject Q={1,3}, four F-tests have to reject $\alpha=0.05$. A different way of saying the same thing is that the maximum of the four p-values must be less than or equal to $\alpha=0.05$; that is, the maximum of the F-test p-values for {1,3}, {1}, {3}, and {} must be less than or equal to 0.05. The chance that Spjotvoll's makes at least one mistake, saying that an adequate model is inadequate, is $\alpha=0.05$, despite doing lots of tests.

It is easy to see why this works. Model Q is adequate if and only if the true model T is a subset of Q, possibly T=Q; that's the definition of "adequate". Suppose Q is adequate, so it contains (or equals) T. To declare Q inadequate at the 0.05 level, you have to reject every model formed as a subset of Q – that's Spjotvoll's method – but as T is one of those subsets, you have to reject T, and the chance that the F-test falsely rejects the true model T is α=0.05.

Instead of focusing on α=0.05, we could do this for any α in just the same way. We can define an adjusted p-value for model Q as the maximum F-test p-value for all of the models that are subsets of Q, including Q itself and the empty model. This adjusted p-value rejects Q as inadequate at level α if and only if the adjusted p-value is at most α.

Spjotvoll’s Method in Variable Selection, continued.

```
> attach(O2Uptake)
> y<-LogO2Up
> x<-O2Uptake[,2:6]
```

Test of model lm(LogO2Up~TS+COD). Compare with row 7.

```
>anova(lm(LogO2Up~TS+COD),lm(LogO2Up~Bod+TKN+TS+TVS+COD))
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     17 1.08502
2     14 0.96512  3    0.1199 0.5797 0.6379
```

The p-value from the F-test is 0.638, whereas the adjusted p-value is the maximum of the p-values for the models contained in {TS,COD}, namely 0.638 for {TS,COD}, 0.139 for {TS}, 0.129 for {COD} and 0.000 for the empty model {}, so the adjusted p-value is 0.638. So this model is not declared inadequate.

Model #16, {TKN, TVS}, is declared inadequate, but its adjusted p-value comes from model #5, {TVS}, because the largest p-value for a model contained in {TKN,TVS} is from model {TVS}.

```
> spjotvoll(x,y)
   p     Cp     Fp  pval adjusted.pval inadequate Bod TKN TS TVS COD
1  1 55.463 11.893 0.000         0.000       TRUE   0   0  0   0   0
2  2  6.297  2.074 0.139         0.139      FALSE   0   0  1   0   0
3  2  6.576  2.144 0.129         0.129      FALSE   0   0  0   0   1
4  2 13.505  3.876 0.025         0.025       TRUE   1   0  0   0   0
5  2 20.331  5.583 0.007         0.007       TRUE   0   0  0   1   0
6  2 56.861 14.715 0.000         0.000       TRUE   0   1  0   0   0
7  3  1.739  0.580 0.638         0.638      FALSE   0   0  1   0   1
8  3  5.274  1.758 0.201         0.201      FALSE   0   0  0   1   1
9  3  6.872  2.291 0.123         0.129      FALSE   0   1  0   0   1
10 3  6.885  2.295 0.122         0.139      FALSE   1   0  1   0   0
11 3  7.165  2.388 0.113         0.139      FALSE   0   0  1   1   0
12 3  7.336  2.445 0.107         0.139      FALSE   0   1  1   0   0
13 3  7.705  2.568 0.096         0.129      FALSE   1   0  0   0   1
14 3  9.097  3.032 0.065         0.065      FALSE   1   1  0   0   0
15 3 11.331  3.777 0.036         0.036       TRUE   1   0  0   1   0
16 3 21.369  7.123 0.004         0.007       TRUE   0   1  0   1   0
17 4  2.319  0.160 0.854         0.854      FALSE   0   1  1   0   1
18 4  3.424  0.712 0.508         0.638      FALSE   0   0  1   1   1
19 4  3.439  0.720 0.504         0.638      FALSE   1   0  1   0   1
20 4  5.665  1.833 0.196         0.201      FALSE   0   1  0   1   1
21 4  6.253  2.126 0.156         0.156      FALSE   1   1  1   0   0
22 4  6.515  2.258 0.141         0.141      FALSE   1   1  0   0   1
23 4  7.152  2.576 0.112         0.201      FALSE   1   0  0   1   1
24 4  8.155  3.077 0.078         0.139      FALSE   1   0  1   1   0
25 4  8.165  3.082 0.078         0.139      FALSE   0   1  1   1   0
26 4  8.681  3.341 0.065         0.065      FALSE   1   1  0   1   0
27 5  4.001  0.001 0.972         0.972      FALSE   0   1  1   1   1
28 5  4.319  0.319 0.581         0.854      FALSE   1   1  1   0   1
29 5  5.068  1.068 0.319         0.638      FALSE   1   0  1   1   1
30 5  6.776  2.776 0.118         0.201      FALSE   1   1  0   1   1
31 5  7.697  3.697 0.075         0.156      FALSE   1   1  1   1   0
32 6  6.000     NA 1.000         1.000      FALSE   1   1  1   1   1
```

Spjotvoll’s method is a case of closted testing; see Marcus et al. (1976)

Spjotvoll, E. (1977) Alternatives to plotting C_P in multiple regression. Biometrika 64, 1-8. Correction: page 241.

Marcus R, Peritz E, Gabriel KR. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63, 655—60.

ANOVA

Memory data. 36 kids randomized to form 3 groups of 12, which were given different treatments. The 'data' are columns 1 and 2, for `group` and Y=`words`. It is a "balanced design" because every group has the same sample size. The rest of memory consists of various ways of coding the 2 degrees of freedom between the three groups into two coded variables. The variables `ten` and `five` are "dummy variables" for two categories, leaving out the third category. The variables `five_ten` and `nh_ten` are used to produce effects that are deviations from a mean for all three groups. The best coding is `hier` and `info` which involve "orthogonal contrasts," discussed below. It is only with orthogonal contrasts that you partition the sum of squares between groups into single degree of freedom parts that add back to the total.

```
> memory[1:3,]
> memory
    group words five_ten nh_ten ten five hier info
1     Ten    50       -1     -1   1    0  0.5    1
2     Ten    49       -1     -1   1    0  0.5    1
3     Ten    44       -1     -1   1    0  0.5    1
4     Ten    31       -1     -1   1    0  0.5    1
5     Ten    47       -1     -1   1    0  0.5    1
6     Ten    38       -1     -1   1    0  0.5    1
7     Ten    38       -1     -1   1    0  0.5    1
8     Ten    48       -1     -1   1    0  0.5    1
9     Ten    45       -1     -1   1    0  0.5    1
10    Ten    48       -1     -1   1    0  0.5    1
11    Ten    35       -1     -1   1    0  0.5    1
12    Ten    33       -1     -1   1    0  0.5    1
13   Five    44        1      0   0    1  0.5   -1
14   Five    41        1      0   0    1  0.5   -1
15   Five    34        1      0   0    1  0.5   -1
16   Five    35        1      0   0    1  0.5   -1
17   Five    40        1      0   0    1  0.5   -1
18   Five    44        1      0   0    1  0.5   -1
19   Five    39        1      0   0    1  0.5   -1
20   Five    39        1      0   0    1  0.5   -1
21   Five    45        1      0   0    1  0.5   -1
22   Five    41        1      0   0    1  0.5   -1
23   Five    46        1      0   0    1  0.5   -1
24   Five    32        1      0   0    1  0.5   -1
25 NoHier    33        0      1   0    0 -1.0    0
26 NoHier    36        0      1   0    0 -1.0    0
27 NoHier    37        0      1   0    0 -1.0    0
28 NoHier    42        0      1   0    0 -1.0    0
29 NoHier    33        0      1   0    0 -1.0    0
30 NoHier    33        0      1   0    0 -1.0    0
31 NoHier    41        0      1   0    0 -1.0    0
32 NoHier    33        0      1   0    0 -1.0    0
33 NoHier    38        0      1   0    0 -1.0    0
34 NoHier    39        0      1   0    0 -1.0    0
35 NoHier    28        0      1   0    0 -1.0    0
36 NoHier    42        0      1   0    0 -1.0    0
```

ANOVA

```
> attach(memory)
```

The anova can be done as a linear model with a factor as the predictor.

```
> anova(lm(words~group))
Analysis of Variance Table

Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
group      2 215.06  107.53  3.7833 0.03317 *
Residuals 33 937.92   28.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Or you can use the aov command. You get the same answer.

```
> summary(aov(words~group))
            Df Sum Sq Mean Sq F value  Pr(>F)
group        2 215.06  107.53  3.7833 0.03317 *
Residuals   33 937.92   28.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Comparisons using Tukey’s Method

```
> TukeyHSD(aov(words~group))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = words ~ group)

$group
                 diff        lwr       upr
NoHier-Five -3.750000 -9.0905714  1.590571
Ten-Five     2.166667 -3.1739047  7.507238
Ten-NoHier   5.916667  0.5760953 11.257238
```

These are simultaneous 95% confidence intervals for the difference in means between two groups. The promise is that all 3 confidence intervals will cover their population differences in 95% of experiments. This is a better promise than that each one, by itself, covers in 95% of uses, because then the first interval would have a 5% chance of error, and so would the second, and so would the third, and the chance of at least one error would be greater than 5%. If the interval includes zero, as the first two intervals do, then you can't declare the two groups significantly different. If the interval excludes zero, as the third interval does, you can declare the two groups significantly different.

Tukey, Bonferroni and Holm

```
> help(pairwise.t.test)
> help(p.adjust)

> TukeyHSD(aov(words~group))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = words ~ group)
$group
                 diff        lwr       upr
NoHier-Five -3.750000 -9.0905714  1.590571
Ten-Five     2.166667 -3.1739047  7.507238
Ten-NoHier   5.916667  0.5760953 11.257238

> pairwise.t.test(words,group,p.adj = "none")
        Pairwise comparisons using t tests with pooled SD
data:  words and group
       Five  NoHier
NoHier 0.094 -
Ten    0.327 0.010
P value adjustment method: none

>  pairwise.t.test(words,group,p.adj = "bonf")
        Pairwise comparisons using t tests with pooled SD
data:  words and group
       Five  NoHier
NoHier 0.283 -
Ten    0.980 0.031
P value adjustment method: bonferroni

> pairwise.t.test(words,group,p.adj = "holm")
        Pairwise comparisons using t tests with pooled SD
data:  words and group
       Five  NoHier
NoHier 0.189 -
Ten    0.327 0.031
```

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70. http://www.jstor.org/

Wright, S. P. (1992). Adjusted P-values for simultaneous inference. *Biometrics*, 48, 1005-1013. http://www.jstor.org/

ANOVA: Many Ways to Code the Same Anova

Here are three different codings with the same Anova table. Notice that much is the same, but some things differ. >

```
summary(lm(words~ten+five))
Call:
lm(formula = words ~ ten + five)
Residuals:
    Min      1Q  Median      3Q     Max 
-11.167  -3.479   0.875   4.771   7.833 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   36.250      1.539  23.554   <2e-16 ***
ten            5.917      2.176   2.718   0.0104 *  
five           3.750      2.176   1.723   0.0943 .  

Residual standard error: 5.331 on 33 degrees of freedom
Multiple R-Squared: 0.1865,     Adjusted R-squared: 0.1372 
F-statistic: 3.783 on 2 and 33 DF,  p-value: 0.03317 

> summary(lm(words~five_ten+nh_ten))
Call:
lm(formula = words ~ five_ten + nh_ten)
Residuals:
    Min      1Q  Median      3Q     Max 
-11.167  -3.479   0.875   4.771   7.833 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  39.4722     0.8885  44.424   <2e-16 ***
five_ten      0.5278     1.2566   0.420   0.6772    
nh_ten       -3.2222     1.2566  -2.564   0.0151 *  

Residual standard error: 5.331 on 33 degrees of freedom
Multiple R-Squared: 0.1865,     Adjusted R-squared: 0.1372 
F-statistic: 3.783 on 2 and 33 DF,  p-value: 0.03317 

> summary(lm(words~hier+info))
Call:
lm(formula = words ~ hier + info)
Residuals:
    Min      1Q  Median      3Q     Max 
-11.167  -3.479   0.875   4.771   7.833 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  39.4722     0.8885  44.424   <2e-16 ***
hier          3.2222     1.2566   2.564   0.0151 *  
info          1.0833     1.0882   0.996   0.3267    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.331 on 33 degrees of freedom
Multiple R-Squared: 0.1865,     Adjusted R-squared: 0.1372 
          F-statistic: 3.783 on 2 and 33 DF,  p-value: 0.03317
```

Contrasts in ANOVA: Better Coding of Nominal Variables

The third coding is best, because the predictors (contrasts) are uncorrelated, so the sums of squares partition.

```
> cor(memory[,3:4])
         five_ten nh_ten
five_ten      1.0    0.5
nh_ten        0.5    1.0

>  cor(memory[,5:6])
      ten five
ten   1.0 -0.5
five -0.5  1.0

>  cor(memory[,7:8])
     hier info
hier    1    0
info    0    1
```

Notice that hier and info have zero correlation: they are orthogonal. Because of this, you can partition the two degrees of freedom between groups into separate sums of squares.

```
> anova(lm(words~hier+info))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
hier       1 186.89  186.89  6.5756 0.01508 *
info       1  28.17   28.17  0.9910 0.32674
Residuals 33 937.92   28.42
---
```

Reverse the order of info and hier, and you get the same answer.

```
> anova(lm(words~info+hier))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
info       1  28.17   28.17  0.9910 0.32674
hier       1 186.89  186.89  6.5756 0.01508 *
Residuals 33 937.92   28.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can't do this with correlated predictors, because they overlap, and the order of the variables changes the sum of squares for the variable, so one can't really say that what portion of the sum of squares belongs to the variable.

```
> anova(lm(words~ten+five))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
ten        1 130.68  130.68  4.5979 0.03947 *
five       1  84.38   84.38  2.9687 0.09425 .
Residuals 33 937.92   28.42

> anova(lm(words~five+ten))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
five       1   5.01    5.01  0.1764 0.67720
ten        1 210.04  210.04  7.3902 0.01037 *
Residuals 33 937.92   28.42
```

Coding the Contrasts in ANOVA in R

This is about shortcuts to get R to convert a nominal variable into several contrasts. There's no new statistics here; just R details.

The group variable, `memory$group`*, is a factor.*

```
> is.factor(memory$group)
[1] TRUE
```

This factor has 3 levels. Notice that the levels are ordered and the order matters.

```
> levels(memory$group)
[1] "Five"   "NoHier" "Ten"

> memory$group
 [1] Ten    Ten    Ten    Ten    Ten    Ten    Ten    Ten    Ten
[10] Ten    Ten    Ten    Five   Five   Five   Five   Five   Five
[19] Five   Five   Five   Five   Five   Five   NoHier NoHier NoHier
[28] NoHier NoHier NoHier NoHier NoHier NoHier NoHier NoHier NoHier
Levels: Five NoHier Ten
```

If you do nothing, R codes a factor in a linear model using 'dummy coding'.

```
> contrasts(memory$group)
       NoHier Ten
Five        0   0
NoHier      1   0
Ten         0   1
```

You can change the coding. Essentially, you can replace the little table above by whatever you want. We will build an new 3x2 table and redefine the contrasts to be this new table.

```
> hier2<-c(.5,-1,.5)
> hier2
[1]  0.5 -1.0  0.5

> info2<-c(-1,0,1)
> info2
[1] -1  0  1

> cm<-cbind(hier2,info2)
> cm
     hier2 info2
[1,]   0.5    -1
[2,]  -1.0     0
[3,]   0.5     1
```

So `cm` *is our new table, and we redefine the contrasts for* `memory$group`*.*

```
> contrasts(memory$group)<-cm
```

This replaces the 'dummy coding' by our new coding.

```
> contrasts(memory$group)
       hier2 info2
Five     0.5    -1
NoHier  -1.0     0
Ten      0.5     1
```

Coding the Contrasts in ANOVA, Continued

If you ask R to extend the contrasts into variables, it will do this with "model.matrix". Notice that this is the coding in the original data matrix, but R is happy to generate it for you using the contrasts you specified.

```
> m<-model.matrix(memory$words~memory$group)
> m
   (Intercept) memory$grouphier2 memory$groupinfo2
1            1               0.5                 1
2            1               0.5                 1
3            1               0.5                 1
4            1               0.5                 1
5            1               0.5                 1
6            1               0.5                 1
7            1               0.5                 1
8            1               0.5                 1
9            1               0.5                 1
10           1               0.5                 1
11           1               0.5                 1
12           1               0.5                 1
13           1               0.5                -1
14           1               0.5                -1
15           1               0.5                -1
16           1               0.5                -1
17           1               0.5                -1
18           1               0.5                -1
19           1               0.5                -1
20           1               0.5                -1
21           1               0.5                -1
22           1               0.5                -1
23           1               0.5                -1
24           1               0.5                -1
25           1              -1.0                 0
26           1              -1.0                 0
27           1              -1.0                 0
28           1              -1.0                 0
29           1              -1.0                 0
30           1              -1.0                 0
31           1              -1.0                 0
32           1              -1.0                 0
33           1              -1.0                 0
34           1              -1.0                 0
35           1              -1.0                 0
36           1              -1.0                 0

> hcontrast<-m[,2]
> icontrast<-m[,3]
```

We now do the anova with single degree of freedom contrasts.

```
> anova(lm(memory$words~hcontrast+icontrast))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
hcontrast  1 186.89  186.89  6.5756 0.01508 *
icontrast  1  28.17   28.17  0.9910 0.32674
Residuals 33 937.92   28.42
```

ANOVA DECOMPOSITION

```
>  mod<-aov(words~group,projections=T)

> mod$projections
   (Intercept)      group     Residuals
1     39.47222  2.6944444  7.833333e+00
2     39.47222  2.6944444  6.833333e+00
3     39.47222  2.6944444  1.833333e+00
4     39.47222  2.6944444 -1.116667e+01
5     39.47222  2.6944444  4.833333e+00
6     39.47222  2.6944444 -4.166667e+00
7     39.47222  2.6944444 -4.166667e+00
8     39.47222  2.6944444  5.833333e+00
9     39.47222  2.6944444  2.833333e+00
10    39.47222  2.6944444  5.833333e+00
11    39.47222  2.6944444 -7.166667e+00
12    39.47222  2.6944444 -9.166667e+00
13    39.47222  0.5277778  4.000000e+00
14    39.47222  0.5277778  1.000000e+00
15    39.47222  0.5277778 -6.000000e+00
16    39.47222  0.5277778 -5.000000e+00
17    39.47222  0.5277778 -9.503032e-16
18    39.47222  0.5277778  4.000000e+00
19    39.47222  0.5277778 -1.000000e+00
20    39.47222  0.5277778 -1.000000e+00
21    39.47222  0.5277778  5.000000e+00
22    39.47222  0.5277778  1.000000e+00
23    39.47222  0.5277778  6.000000e+00
24    39.47222  0.5277778 -8.000000e+00
25    39.47222 -3.2222222 -3.250000e+00
26    39.47222 -3.2222222 -2.500000e-01
27    39.47222 -3.2222222  7.500000e-01
28    39.47222 -3.2222222  5.750000e+00
29    39.47222 -3.2222222 -3.250000e+00
30    39.47222 -3.2222222 -3.250000e+00
31    39.47222 -3.2222222  4.750000e+00
32    39.47222 -3.2222222 -3.250000e+00
33    39.47222 -3.2222222  1.750000e+00
34    39.47222 -3.2222222  2.750000e+00
35    39.47222 -3.2222222 -8.250000e+00
36    39.47222 -3.2222222  5.750000e+00
attr(,"df")
(Intercept)       group   Residuals
          1           2          33
```

Orthogonal and Non-orthogonal Predictors

```
> attach(memory)
> summary(aov(words~group))
            Df Sum Sq Mean Sq F value  Pr(>F)
group        2 215.06  107.53  3.7833 0.03317 *
Residuals   33 937.92   28.42
---
```

ten and five are not orthogonal predictors – so there is not a unique sum of squares for each

```
> anova(lm(words~ten+five))
Analysis of Variance Table
Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
ten        1 130.68  130.68  4.5979 0.03947 *
five       1  84.38   84.38  2.9687 0.09425 .
Residuals 33 937.92   28.42
---
> anova(lm(words~five+ten))
Analysis of Variance Table

Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
five       1   5.01    5.01  0.1764 0.67720
ten        1 210.04  210.04  7.3902 0.01037 *
Residuals 33 937.92   28.42
```

her and info are orthogonal predictors – so there is a unique sum of squares for each

```
> anova(lm(words~hier+info))
Analysis of Variance Table

Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
hier       1 186.89  186.89  6.5756 0.01508 *
info       1  28.17   28.17  0.9910 0.32674
Residuals 33 937.92   28.42

>  anova(lm(words~info+hier))
Analysis of Variance Table

Response: words
          Df Sum Sq Mean Sq F value  Pr(>F)
info       1  28.17   28.17  0.9910 0.32674
hier       1 186.89  186.89  6.5756 0.01508 *
Residuals 33 937.92   28.42
```

Simulating in R

Ten observations from the standard Normal distribution:

```
> rnorm(10)
 [1]  0.8542301 -1.3331572  1.4522862  0.8980641  0.1456334
 [6]  0.4926661 -0.4366962  0.6204263 -0.1582319 -0.6444449
```

Fixed integer sequences

```
> 1:2
[1] 1 2

> 1:5
[1] 1 2 3 4 5

> 0:1
[1] 0 1
```

20 coin flips

```
> sample(0:1,20,r=T)
 [1] 0 1 0 0 1 1 0 0 0 1 0 1 0 0 0 1 0 1 1 1
```

10 random numbers from 1 to 5

```
> sample(1:5,10,r=T)
 [1] 5 2 3 5 2 2 1 1 3 2
```

More information:

```
     help(sample)
     help(rnorm)
```

PROBLEM SET #1 STATISTICS 500 Fall 2018: DATA PAGE 1

Due at noon in class Tuesday 23 Oct 2018.

This is an exam. Do not discuss it with anyone.

A study by Feuer et al. (2015) (Feuer AJ, Demmer RT, Thai A, Vogiatzi MG. 2015. Use of selective serotonin reuptake inhibitors and bone mass in adolescents. Bone 78:28-33) suggested that use of a certain class of drugs, selective serotonin reuptake inhibitors (SSRIs) may reduce bone mineral density in children. We will reexamine the data they used from NHANES 2005-2010. The data for this problem set are in an object, **bonessri**, in the course workspace. (You can obtain the complete NHANES data at https://www.cdc.gov/nchs/nhanes/ , but there is no reason to do this unless you want to.)

The SSRIs are PAROXETINE, FLUOXETINE, SERTRALINE, FLUVOXAMINE, CITALOPRAM, and ESCITALOPRAM. They are often used to treat psychiatric disorders, including depression. The key variable for you is ssriyears, which is total years receiving an SSRI, capped at 3 years. Bone density is obtained from dual-energy x-ray absorptiometry. **Use femurbmd**, not femurbmc. The variables in bonessri are:

SEQN = NHANES id number
nhanes = Which NHANES? 0506, 0708, 0910
age = age in years, 8-20
female = 1 for female, 0 for male
povertyr = ratio of family income to the poverty level, capped at 5xPoverty
weight = weight in kilograms
height = height in centimeters
bmi = body mass index
waist = waist in centimeters
femurbmd = Total femur **bone mineral density** (gm/cm^2)
femurbmc = Total femur bone mineral content (gm)
ssridays = days using an SSRI
ssriyears = years using an SSRI, capped at 3 years
druguse = uses a prescription drug
drugcount = number of prescription drugs

You should look at the data in various ways.

```
dim(bonessri)
[1] 6053   15
attach(bonessri)
sum(ssriyears>0) # Only 69 children received an ssri
boxplot(femurbmd)
boxplot(povertyr)
boxplot(bmi)
plot(ssriyears,femurbmd)
lines(lowess(ssriyears,femurbmd))
```

PROBLEM SET #1 STATISTICS 500 Fall 201: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due at noon in class Tuesday 23 Oct 2018

The models have different Greek symbols so that different things have different names.

Model 1: (Make sure y is femurbmd)

femurbmd = $\gamma_0 + \gamma_1$ ssriyears + ε

where ε is iid $N(0,\omega^2)$.

Model 2:

femurbmd = β_0+$\beta_1$ssriyears+$\beta_2$age+$\beta_3$female+ε

where ε is iid $N(0,\sigma^2)$

Model 3:

femurbmd = λ_0+$\lambda_1$ssriyears+$\lambda_2$female+ε

where ε is iid $N(0,\kappa^2)$

Model 4: (Note that y is age, not femurbmd)

age = η_0+$\eta_1$ssriyears+$\eta_2$female+ε

where ε is iid $N(0,\theta^2)$

Question 5 is about the added variable plot. It asks you to work with the residuals from models 3 and 4 to get your own added variable plot. It asks you to add the lowess smooth to the added variable plot. Get y = residuals from model 3 and x = residuals from model 4. Do

```
plot(x,y)
lines(lowess(x,y),col="red",lwd=3)
abline(lm(y~x))
```

Call this the **Question 5 Plot.** You might compare it to the avPlot() for age in model 2 obtained from the car package. The only difference is the red lowess curve.

Important: Write your name on both sides of the answer page, **last name first. Turn in only the answer page.** Do not turn in plots. Brief answers suffice. **Circle** the correct answer, but do not cross out an answer. A circled answer may be correct or incorrect, but every crossed out answer is incorrect. Do not give one answer adding a note explaining why a different answer is correct. If a true/false questions says (A and B and C), and A is false, then the answer is false even if B and C are true. If a true/false question says (A because of B), and A is true, but not because of B, then the answer is false.

This is an exam. Do not discuss the exam with anyone. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name (**Last**, First):_____________________ ID# _____________

PROBLEM SET #1 STATISTICS 500 Fall 2018: ANSWER PAGE 1

This is an exam. Do not discuss it. Due noon Oct 23.

Part 1. Fit **Model #1** from the data page for part 1.	Fill in or **circle** the correct answer.
1.1 In Model 1, test the null hypothesis about one coefficient, H_0: $\gamma_1=0$. What is the name of the test? What is the numerical value of the test statistic? What is the **two-sided** P-value? Is the null hypothesis plausible?	Name:_________ Value:_______ P-value:________ **Circle one** Plausible Not plausible
1.2 Give the 95% confidence interval for γ_1 in Model 1.	[,]
1.3 In this one regression, the 0.05-level test in 1.1 and the 95% confidence interval in 1.2 happen to agree about whether $\gamma_1=0$ is plausible, but they might disagree in some other regression.	**Circle one** True False

Fit **model 2** on the data page. Use model 2 for the questions in part 2.	Fill in or **circle** the correct answer.
2.1 In model 2, test the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. What is the **name** of the test, the **value** of the test statistic, the **P-value**. What are the degrees of freedom (DF) needed to determine the P-value for this test statistic? Is H_0 **plausible**?	Name:_________ Value:_______ P-value:________ DF = ________________ **Circle one** Plausible Not plausible
2.2 The Normal plot of residuals shows one outlying child with an extremely negative residual, incompatible with a Normal distribution.	**Circle one** True False

Name (**Last**, First):_____________________ ID# ______________

PROBLEM SET #1 STATISTICS 500 Fall 2018: ANSWER PAGE 2

This is an exam. Do not discuss it. Due noon Oct 23.

Use model 2 to answer questions in part 3.	Fill in or **circle** the correct answer
3.1 Give the 95% confidence interval for β_1 in Model 2.	[,]
3.2 What is the numerical value of the estimate of σ^2 in model 2? (Not σ!) If model 2 were true, this estimate would be biased downward because of overfitting.	Estimate of σ^2: __________ (Not the estimate of σ.) **Biased down? Circle one.** True False

4. In model 2, test the null hypothesis H_0: $\beta_2 = \beta_3 = 0$. Fill in the anova table.

Source of Variation	Sums of square	Degrees of freedom	Mean square	F-ratio
Full Model				
ssriyears alone				xxxxxxxxxx xxxxxxxxxx
Added by age and female				
Residual from full model				xxxxxxxxxx xxxxxxxxxx

See the data page regarding question 5.	Fill in or **circle** the correct answer
5.1 Do a simple regression of the residuals of model 3 on the residuals of model 4, and give the fitted slope. Give the estimate of β_2 in Model 2, that is, of age.	Fitted slope:_____________ Estimate of β_2 from Model 2: _____________ (Coefficient of age)
5.2 Do the Question 5 Plot. That plot clearly supports the idea of linearly relating bone density and age in model 2.	**Circle one.** True False

ANSWERS

PROBLEM SET #1 STATISTICS 500 Fall 2018: ANSWER PAGE 1

Part 1. Fit **Model #1** from the data page for part 1.	10 points per question
1.1 In Model 1, test the null hypothesis about one coefficient, H_0: $\gamma_1=0$. What is the name of the test? What is the numerical value of the test statistic? What is the **two-sided** P-value? Is the null hypothesis plausible?	Name:t-test Value: -1.528 P-value: 0.127 **Circle one** [✓ Plausible] Not plausible
1.2 Give the 95% confidence interval for γ_1 in Model 1.	[-.0465, .0058]
1.3 In this one regression, the 0.05-level test in 1.1 and the 95% confidence interval in 1.2 happen to agree about whether $\gamma_1=0$ is plausible, but they might disagree in some other regression.	**Circle one** True [✓ False] The 95% confidence interval excludes 0 if the P-value testing H_0: $\gamma_1=0$ is less than or equal to 0.05.

Fit **model 2** on the data page. Use model 2 for the questions in part 2.	Fill in or **circle** the correct answer.
2.1 In model 2, test the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. What is the **name** of the test, the **value** of the test statistic, the **P-value**. What are the degrees of freedom (DF) needed to determine the P-value for this test statistic? Is H_0 **plausible**?	Name: F-test Value: 1977 P-value: 2.2 x 10^{-16} DF = (3, 6049) **Circle one** Plausible [✓ Not plausible]
2.2 The Normal plot of residuals shows one outlying child with an extremely negative residual, incompatible with a Normal distribution.	**Circle one** True [✓ False]

ANSWERS

PROBLEM SET #1 STATISTICS 500 Fall 2018: ANSWER PAGE 2

Use model 2 to answer questions in part 3.	Fill in or **circle** the correct answer
3.1 Give the 95% confidence interval for β_1 in Model 2.	[-0.061, -0.023]
3.2 What is the numerical value of the estimate of σ^2 in model 2? (Not σ!) If model 2 were true, this estimate would be biased downward because of overfitting.	Estimate of σ^2: 0.01798 (Not the estimate of σ.) **Biased down? Circle one.** True [✓ False] We divide by DF = n-(k+1) to avoid bias.

4. In model 2, test the null hypothesis H_0: $\beta_2 = \beta_3 = 0$. Fill in the anova table.

Source of Variation	Sums of squares	Degrees of freedom	Mean square	F-ratio
Full Model	106.56	3	35.52	1976.6
ssriyears alone	0.083	1	0.083	xxxxxxxxxx xxxxxxxxxx
Added by age and female	106.47	2	53.235	2962.6
Residual from full model	108.70	6049	0.01797	xxxxxxxxxx xxxxxxxxxx

See the data page regarding question 5.	Fill in or **circle** the correct answer
5.1 Do a simple regression of the residuals of model 3 on the residuals of model 4, and give the fitted slope. Give the estimate of β_2 in Model 2, that is, of age.	Fitted slope: 0.03605 Estimate of β_2 from Model 2: 0.03605 (Coefficient of age)
5.2 Do the Question 5 Plot. That plot clearly supports the idea of linearly relating bone density and age in model 2.	**Circle one.** True [✓ False] lowess smooth is curved!

Doing the Problem Set in R

PROBLEM SET #1 STATISTICS 500 Fall 2018

```
dat<-read.csv("bone0510.csv")
ssridays<-
dat$PAROXETINEdays+dat$FLUOXETINEdays+dat$SERTRALINEdays+dat$FLUVOXAMINEdays+
  dat$CITALOPRAMdays+dat$ESCITALOPRAMdays
ssriyears<-ssridays/365
ssriyears<-pmin(ssriyears,3)
attach(dat)
d<-
data.frame(SEQN,nhanes,age,female,povertyr,weight,height,bmi,waist,femurbmd,femurbmc,
              ssridays,ssriyears,druguse,drugcount)
detach(dat)
rm(ssridays,ssriyears)
bonessri<-d[complete.cases(d),]
rm(d,dat)
attach(bonessri)
sum(ssriyears>0) # Only 69 children received an ssri
table(ssriyears)
boxplot(femurbmd)
boxplot(age)
boxplot(povertyr)
boxplot(bmi)
m1<-lm(femurbmd~ssriyears)
summary(m1) # Question 1.1
confint(m1) # Question 1.2
m2<-lm(femurbmd~ssriyears+age+female)
confint(m2) # Question 3.1
summary(m2) # Question 3.2
0.1341^2
anova(m1,m2) # Question 4
#Question 5
x<-lm(age~ssriyears+female)$residuals
y<-lm(femurbmd~ssriyears+female)$residuals
plot(x,y)
lines(lowess(x,y),col="red",lwd=3)
abline(lm(y~x))
library(car)
avPlot(m2,age)
```

PROBLEM SET #2 STATISTICS 500 Fall 2018: DATA PAGE 1

Due at noon in class Tuesday 20 November 2018.

This is an exam. Do not discuss it with anyone.

As in Problem Set #1, the data refer to a study by Feuer et al. (2015) *Bone* 78:28-33, suggesting that use of a certain class of drugs, selective serotonin reuptake inhibitors (SSRIs) may reduce bone mineral density in children. The data from NHANES for this problem set are in an object, **bonessri**, in the course workspace. **Download the workspace again** to obtain the function mtransform for question 4.

The SSRIs are PAROXETINE, FLUOXETINE, SERTRALINE, FLUVOXAMINE, CITALOPRAM, and ESCITALOPRAM. They are often used to treat psychiatric disorders, including depression. The key variable for you is ssriyears, which is total years receiving an SSRI, capped at 3 years. Bone density is obtained from dual-energy x-ray absorptiometry. **Use femurbmd**, not femurbmc. The variables you need in bonessri are:

age = age in years, 8-20
female = 1 for female, 0 for male
femurbmd = Total femur **bone mineral density** (gm/cm^2)
ssriyears = years using an SSRI, capped at 3 years

STEP 1 is to make "Fig 1" and "Fig 2" as follows. Do not turn in the figures, just answer the question about them.

```
attach(bonessri)
f<-(female==1)
s<-(ssriyears>0)
table(f,s)
boxplot(femurbmd[f]~age[f],xlab="Age",ylab="Bone Density",main="Fig 1")
points(age[s&f]-7,femurbmd[s&f],pch=16,col="red")
boxplot(femurbmd[!f]~age[!f],xlab="Age",ylab="Bone Density",main="Fig 2")
points(age[s&!f]-7,femurbmd[s&!f],pch=16,col="red")
```

Question 2.1 speaks about a 3-dimensional space with 3 coordinates (femurbmd, ssriyears, age) with fumurbmd vertical, ssriyears horizontal, and age as depth, say.

Question 2.3 asks you to do the following and interpret it.

```
library(car)
invTranPlot(age,femurbmd)
tage<-age^(-.4)    # transformed age
invTranPlot(tage,femurbmd)
residualPlots(lm(femurbmd~age))
residualPlots(lm(femurbmd~tage))
plot(tage,femurbmd)
lines(lowess(tage,femurbmd),col="blue")
abline(lm(femurbmd~tage),col="red")
```

Models have different Greek letters so that different things have different names. The letter does not matter.

Model 0: (Make sure y is femurbmd)

femurbmd = $\eta_0 + \eta_1$ age$^{-0.4}$ + ε where ε is iid $N(0,\iota^2)$.

Model 1: (Make sure y is femurbmd)

femurbmd = $\gamma_0 + \gamma_1$ ssriyears + γ_2 age + $\gamma_3$female + ε
where ε is iid $N(0,\omega^2)$.

Model 2: (and related polynomial models)

femurbmd = $\lambda_0 + \lambda_1$ ssriyears + λ_2 age + λ_3 age^2 + $\lambda_4$female + ε
where ε is iid $N(0,\kappa^2)$.

PROBLEM SET #2 STATISTICS 500 Fall 201: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due at noon in class Tuesday 20 November 2018

For Question 3, use orthogonal polynomials (R command poly()), to fit model 2 and other models involved for questions in **Part 3.** Be careful to use the correct tests. You want to know if age^3 needs to be added to a quadratic model, not if age^3 needs to be added to a quartic model. It is easy to mess up if you don't think about it. (A quartic has x, x^2, x^3, x^4.)

Question 4 asks about matched transformations from "Power Transformations for Data Analysis" by Stoto and Emerson, *Sociological Methodology*, 1983-1984, 14, 126-168 (expressions (11) and (12)). Stoto and Emerson like matched transformations because they rescale so the transformed data looks a bit like the original data, and age$^{-0.4}$ doesn't look so strange. Their paper is in jstor on the library web page, https://www.jstor.org/stable/270905, but there is no need to look at it unless you want to. Function **mtransform** in the course workspace or below is all you need:

```
mtransform<-function(y,p){
  m<-median(y)
  if (p!=0)  m+((y^p)-(m^p))/(p*(m^(p-1)))
  else m*(1+log(y)-log(m))}
```

Do the following:

mtage<-mtransform(age,-.4)

Note that question 4 asks about –(age$^{-0.4}$), not (age$^{-0.4}$).

Model 3: Define **interact**<-mtage*female

femurbmd = $\beta_0+\beta_1$ssriyears$+\beta_2$mtage$+\beta_3$female$+\beta_4$interact$+\varepsilon$

where ε is iid N$(0,\sigma^2)$

Important: Write your name on both sides of the answer page, **last name first. Turn in only the answer page.** Do not turn in plots. Brief answers suffice. **Circle** the correct answer, but do not cross out an answer. A circled answer may be correct or incorrect, but every crossed out answer is incorrect. Do not give one answer adding a note explaining why a different answer is correct. If a true/false questions says (A and B and C), and A is false, then the answer is false even if B and C are true. If a true/false question says (A because of B), and A is true, but not because of B, then the answer is false.

This is an exam. Do not discuss the exam with anyone. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name (**Last**, First):_____________________ ID# _____________

PROBLEM SET #2 STATISTICS 500 Fall 2018: ANSWER PAGE 1

This is an exam. Do not discuss it. Due 20 November 2018

Do **STEP 1** on the data page and use it for part 1 below	**Circle** the answer.
1.1: 44 females and 25 males have ssriyears > 0.	True False
1.2: !f takes the value TRUE for a male	True False
1.3: In Figure 2, does points(8,.6,col="green",pch=15) add a point to the boxplot for 8 year olds?	Yes No
1.4: The red points in Figure 2 identify the females.	True False
1.5: In Figure 1, the boxplot at age 11 describes 261 females, one of whom has current exposure to an SSRI, with a bone density that is below the lower quartile in this boxplot.	True False

Part 2. See data page	Fill in/**circle** the answer.
2.1 In 3-space, in model 1 on the data page, the fitted values fall on two parallel planes, one for males, the other for females.	**Circle one** True False
2.2 Starting with **Model 1** on the data page, test model 1 against the alternative model that has a different slope for age for males and females. Given the name of the test, its value, the two-sided P-value, and state whether the null hypothesis is plausible.	Name:_________ Value:_______ P-value:________ **Circle one** Hypothesis of equal age slopes for male and female is: Plausible Not plausible
2.3 **a.** Tukey's test rejects **Model 0.** **b.** Transformation age$^{-0.4}$ is more extreme than both log(age) and 1/age in our family of transformations.	**Circle one for a and b** **a.** True False **b.** True False

Name (**Last**, First):_____________________ ID# ______________

PROBLEM SET #2 STATISTICS 500 Fall 2018: ANSWER PAGE 2

This is an exam. Do not discuss it. Due 20 November 2018

Part 3 **See the data page.**	Fill in / **circle** the answer
3 Test the null hypothesis that **Model 1** is adequate against the alternative that the quadratic in **Model 2** is needed. Repeat by testing **Model 2** against a model adding a cubic in age. Test this cubic model against a model adding a quartic in age. What is the name of the test. Give values of the test statistic, P-values, and state whether age^k is needed for k=2, 3, 4. **See the data page.**	Name of test:_________ Circle "Needed" if the simpler model without this term is rejected at the two-sided, .05 level.

	Value	P-value	
Age^2	_______	_______	Needed
Age^3	_______	_______	Needed
Age^4	_______	_______	Needed

4. See the data page.	**Circle** the answer for a,b,c
a. The correlation between mtage and $-(\text{age}^{-0.4})$ is 1.00. **b.** The median of age equals the median of mtage. **c.** Plotting y=mtage (vertical) against x=age (horizontal), the x=y line of equality just touches the point (x,y) = (median(age),median(mtage)).	**a.** True False **b.** True False **c.** True False

5. Use Model 3 on the data page.	**Circle/Fill in** the Answer
5.1 In **Model 3**, based on the 0.05-level, two-sided t-test, exposure to SSRIs is associated with greater bone density.	True False
5.2 In **Model 3**: What is the numerical value of the t-statistic for Tukey's test of the null hypothesis that no transformation is needed.	Value: ___________
5.3 A lowess curve in the plot of residuals vs fitted values for **Model 3** shows a dramatic inverted-U shape.	True False

ANSWERS (5 points each, except as noted)

PROBLEM SET #2 STATISTICS 500 Fall 2018: ANSWER PAGE 1

Do **STEP 1** on the data page and use it for part 1 below	**Circle** the answer.
1.1: 44 females and 25 males have ssriyears > 0.	[✓ True] False
1.2: !f takes the value TRUE for a male	[✓ True] False
1.3: In Figure 2, does `points(8,.6,col="green",pch=15)` add a point to the boxplot for 8 year olds?	Yes [✓ No]
1.4: The red points in Figure 2 identify the females.	True [✓ False]
1.5: In Figure 1, the boxplot at age 11 describes 261 females, one of whom has current exposure to an SSRI, with a bone density that is below the lower quartile in this boxplot.	[✓ True] False

Part 2. See data page	Fill in/**circle** the answer.
2.1 In 3-space, in model 1 on the data page, the fitted values fall on two parallel planes, one for males, the other for females.	**Circle one** [✓ True] False
2.2 Starting with **Model 1** on the data page, test model 1 against the alternative model that has a different slope for age for males and females. Given the name of the test, its value, the two-sided P-value, and state whether the null hypothesis is plausible.	15 points Name: t-test Value: -10.772 P-value: 2×10^{-16} **Circle one** Hypothesis of equal age slopes for male and female is: Plausible [✓ Not plausible]
2.3 a. Tukey's test rejects **Model 0**. **b.** Transformation age$^{-0.4}$ is more extreme than both log(age) and 1/age in our family of transformations.	**a.** True [✓ False] **b.** True [✓ False] p=-0.4 is between p=0 (log) and p=-1 (reciprocal)

ANSWERS

PROBLEM SET #2 STATISTICS 500 Fall 2018: ANSWER PAGE 2

<table>
<tr><th>Part 3 See the data page.</th><th>Fill in / circle if needed</th></tr>
<tr><td>3 Test at 0.05 the null hypothesis that Model 1 is adequate against the alternative that the quadratic in Model 2 is needed. Repeat by testing Model 2 against a model adding a cubic in age. Test this cubic model against a model adding a quartic in age. What is the name of the test. Give values of the test statistic, P-values, and state whether age^k is needed for k=2, 3, 4. See the data page.</td>
<td>15 points
Name of test: t-test
Circle “Needed” if the simpler model without this term is rejected at the two-sided, .05 level.
<table>
<tr><th></th><th>Value</th><th>P-value</th><th></th></tr>
<tr><td>Age^2</td><td>-17.43</td><td>$2\text{x}10^{-16}$</td><td>[✓ Needed]</td></tr>
<tr><td>Age^3</td><td>-9.55</td><td>$2\text{x}10^{-16}$</td><td>[✓ Needed]</td></tr>
<tr><td>Age^4</td><td>1.66</td><td>0.096</td><td>Needed</td></tr>
</table></td></tr>
</table>

4. See the data page.	**Circle** the answer for a,b,c
a. The correlation between mtage and $-(age^{-0.4})$ is 1.00.	**a.** [✓ True] False
b. The median of age equals the median of mtage.	**b.** [✓ True] False
c. Plotting y=mtage (vertical) against x=age (horizontal), the x=y line of equality just touches the point (x,y) = (median(age),median(mtage)).	**c.** [✓ True] False The median is unchanged and the derivative at the median is 1, so this transformation does little near the median.

5. Use Model 3 on the data page.	**Circle/Fill in** the Answer
5.1 In **Model 3**, based on the 0.05-level, two-sided t-test, exposure to SSRIs is associated with greater bone density.	True [✓ False] Lower bone density!
5.2 In **Model 3**: What is the numerical value of the t-statistic for Tukey’s test of the null hypothesis that no transformation is needed.	Value: 2.489
5.3 A lowess curve in the plot of residuals vs fitted values for **Model 3** shows a dramatic inverted-U shape.	True [✓ False] Any curvature is small.

Problem Set 2, Fall 2018
Doing the Problem Set in R

```
dim(bonessri)
attach(bonessri)
# Question 1.1
f<-(female==1)
s<-(ssriyears>0)
table(f,s)
boxplot(femurbmd[f]~age[f],xlab="Age",ylab="Bone Density",main="Fig 1")
points(age[s&f]-7,femurbmd[s&f],pch=16,col="red")
boxplot(femurbmd[!f]~age[!f],xlab="Age",ylab="Bone Density",main="Fig 2")
points(age[s&!f]-7,femurbmd[s&!f],pch=16,col="red")
# Question 1.3
points(8,.6,col="green",pch=15) # Boxplot 8, not the age=8 boxplot
# Question 1.5
table(age[f])
summary(femurbmd[f&(age==11)])
femurbmd[f&(age==11)&(ssriyears>0)]

# Question 2.2
m1<-lm(femurbmd~ssriyears+age+female)
m1a<-lm(femurbmd~ssriyears+age+female+age*female)
summary(m1a)
# or equivalently you could do
anova(m1,m1a)
(-10.772)^2
# because t^2 = F when F has 1 degree of freedom in the numerator
# Question 2.3
library(car)
invTranPlot(age,femurbmd)
tage<-age^(-.4)    # transformed age
invTranPlot(tage,femurbmd)
residualPlots(lm(femurbmd~age))
residualPlots(lm(femurbmd~tage)) #2.3a
plot(tage,femurbmd)
lines(lowess(tage,femurbmd),col="blue") #2.3b
abline(lm(femurbmd~tage),col="red") #2.3b

# Question 3
summary(lm(femurbmd~ssriyears+female+poly(age,2)))
summary(lm(femurbmd~ssriyears+female+poly(age,3)))
summary(lm(femurbmd~ssriyears+female+poly(age,4)))
```

```
# Question 4
# The mtransform function is in the workspace, but
# here it is again
mtransform<-function(y,p){
       m<-median(y)
       if (p!=0)  m+((y^p)-(m^p))/(p*(m^(p-1)))
       else m*(1+log(y)-log(m))}
mtage<-mtransform(age,-.4)
tage<-age^(-.4)
cor(mtage,-tage)
median(age)
median(mtage)
plot(age,mtage)
abline(0,1) # x=y line, slope=derivative=1 at median=14

# Question 5
interact<-mtage*female
m3<-lm(femurbmd~ssriyears+mtage+female+interact)
# Question 5.1
summary(m3)
# Question 5.2
residualPlots(m3)
# Question 5.3
plot(m3) # Although Tukey rejects in this large sample,
# there is almost no curvature in this plot.
```

PROBLEM SET #3 STATISTICS 500 Fall 2018: DATA PAGE 1

Due at noon Thursday 20 December 2018.

This is an exam. Do not discuss it with anyone.

As in Problem Sets #1 & 2, the first data set refers to a study by Feuer et al. (2015) *Bone* 78:28-33, suggesting that use of a certain class of drugs, selective serotonin reuptake inhibitors (SSRIs) may reduce bone mineral density in children. The data from NHANES for this problem set are in an object, **bonessri**, in the course workspace. The function mtransform is also in the workspace.

The SSRIs are PAROXETINE, FLUOXETINE, SERTRALINE, FLUVOXAMINE, CITALOPRAM, and ESCITALOPRAM. They are often used to treat psychiatric disorders, including depression. The key variable for you is ssriyears, which is total years receiving an SSRI, capped at 3 years. Bone density is obtained from dual-energy x-ray absorptiometry. **Use femurbmd**, not femurbmc. The variables you need in bonessri are:

age = age in years, 8-20
female = 1 for female, 0 for male
femurbmd = Total femur **bone mineral density** (gm/cm^2)
ssriyears = years using an SSRI, capped at 3 years
povertyr = ratio of family income to the poverty level, capped at 5
weight = weight in kilograms
bmi = body mass index

```
mtransform<-function(y,p){
  m<-median(y)
  if (p!=0)  m+((y^p)-(m^p))/(p*(m^(p-1)))
  else m*(1+log(y)-log(m))}
```

Do the following:

mtage<-mtransform(age,-.4)

Model 1: Define **interact**<-mtage*female

$$femurbmd = \beta_0+\beta_1 ssriyears+\beta_2 mtage+\beta_3 female+\beta_4 interact$$
$$+\beta_5 povertyr+\beta_6 weight +\beta_7 bmi +\varepsilon$$

where ε is iid $N(0,\sigma^2)$

Question 1.4 asks you to count submodels of model 1. The submodels of model 1 are the models formed with a subset of the 7 predictors in model 1, including the subset that contains all 7 predictors and the subset that contains no predictors (just β_0).

The second data set for **Part 2** is in an object **abbkk** in the course workspace. You will have to download the workspace again, and you may have to clear your web browser's memory to do that. The data are from a paper, Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. American Economic Review, 92(5), 1535-1558. The paper is in jstor on the library webpage, but there is no need to look at the paper unless you want to. Aided by funding from the World Bank, the Colombian government set up a program called PACES that ran a lottery to provide subsidies for private secondary education to disadvantaged residents of Bogota, Columbia. In effect, there was a randomized experiment offering vouchers to attend private rather than public secondary education. The data you will examine simplify certain issues that are developed in more detail in the original paper. In abbkk, you have outcome y = highest grade completed on the date of the

This is an exam. Do not discuss it with anyone.
Due at noon Thursday 20 December 2018

survey, female = 1 for female, 0 for male, voucher = 1 if offered the voucher at random by the lottery, 0 otherwise, and grp = 4 level factor combining female and voucher. Type table(female,grp) and table(voucher,grp) to understand grp. There are four groups of 285 children.

Model 2:

hgrade = $\gamma_0+\gamma_j+\varepsilon$ for j=1,2,3,4 for the 4 levels of grp
where ε is iid $N(0,\omega^2)$, $0 = \gamma_1+\gamma_2+\gamma_3+\gamma_4$

Question 2.3 and 2.4 ask you to compare the 4 groups defined by grp in all 6 pairs, **listing the pairs that differ significantly** at the 0.05 level, where 0.05 refers to strong control of the family-wise error. In these questions, if Femalev and FemaleNov differed significantly, you would write (Femalev,FemaleNov) as a pair that differed significantly. If no pair differs significantly, write NONE. You will either write NONE, or you will list 1 or 2 or 3 or 4 or 5 or 6 pairs.

Question 2.5: Suppose that there are four groups, as in 2.3 and 2.5, with population means μ_1, μ_2, μ_3, μ_4. Then there are 6 hypotheses about 2 means, $\mu_1=\mu_2$, $\mu_1=\mu_3$, $\mu_1=\mu_4$, $\mu_2=\mu_3$, $\mu_2=\mu_4$, and $\mu_3=\mu_4$. Perhaps all six hypotheses are true, so the smallest P-value is multiplied by 6 in the Bonferroni, Holm and Shaffer procedures. **If $\mu_1\neq\mu_2$, then what is the maximum possible number of true hypotheses among the 6 hypotheses?** The answer is an integer in {0,1,2,3,4,5,6}. (This is how Shaffer beats Holm with four groups.)

Question 3 asks you to fill in an anova table after rewriting model 2 in terms of three **orthogonal** contrasts, gender, voucher and their interaction.

Important: Write your name on both sides of the answer page, **last name first. Turn in only the answer page.** Brief answers suffice. **Circle** the correct answer, but do not cross out an answer.

Turning in the exam: You may turn the exam in on Thursday 20 December at noon in the statistics department. You may turn it in early by placing it in my mailbox in the statistics department in an envelope addressed to me. You may scan and email your answer page, but please **DO NOT** email a cellphone photo of your exam. **Do not submit your exam twice in two different ways. Make and keep a photocopy of your answer page.** (That way, you have a backup if it gets lost in transit, and you can compare your photocopy to the answer key when it is posted on-line. I do not return finals.)

This is an exam. Do not discuss the exam with anyone. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name (**Last**, First):_____________________ ID# _____________

PROBLEM SET #3 STATISTICS 500 Fall 2018: ANSWER PAGE 1

This is an exam. Do not discuss it. Due 20 December 2018

<table>
<tr><td>Part 1. Use Model 1 from the data page in part 1.</td><td>Fill in/circle the answer.</td></tr>
<tr><td>1.1 In model 1, the one child with the largest hatvalue or leverage is a 19 year old male with a bmi above 27 and at least 3 years on ssri’s.</td><td>Circle one

True False</td></tr>
<tr><td>1.2 How many children have ssriyears > 0? How many have ssriyears = 0? In model 1, how many children have hatvalues > 6 times the mean hatvalue? Of children with hatvalues > 6 times the mean hatvalue, how many have ssriyears ≥ 1?</td><td>Give 4 counts of children

ssriyears>0:______ =0:_____
hatvalues > 6×mean(hatvalue):_______

children with hatvalues> 6×mean(hatvalue) and ssriyears ≥ 1: _____________</td></tr>
<tr><td>1.3 Test for outliers in model 1. What is the largest t-statistic in absolute value for a single-child binary indicator variable added to model 1? What are the 2 two-sided P-values for this t-statistic, one uncorrected for multiple tested, the other corrected by the Bonferroni method?</td><td>t-statistic: ____________<table><tr><td>Uncorrected P</td><td>Corrected P</td></tr><tr><td></td><td></td></tr></table>Exactly two children are significant outliers at the 0.05 level by Bonferroni?

Circle one TRUE FALSE</td></tr>
<tr><td>1.4 Model 1 has how many submodels (see the data page)? What is the value of Mallows’ C_P for model 1? Which submodel of model 1 has the smallest value of C_P? What is the value of this smallest C_P?</td><td># Submodels: ____________<table><tr><td>C_P For model 1</td><td>Smallest C_P?</td></tr><tr><td></td><td></td></tr></table>List the predictors in the model with the smallest smallest C_P?</td></tr>
<tr><td>1.5 The model with the smallest C_P contains all of the predictors with nonzero coefficients, $\beta_j \neq 0$.</td><td>Circle one

TRUE FALSE</td></tr>
<tr><td>1.6 Use the diagnostic plot and the score test (from car) to check the assumption that the variance of the errors in model 1 is constant.</td><td>“It is plausible that the variance of the errors in model 1 is constant.” Circle one:

TRUE FALSE</td></tr>
</table>

Name (**Last**, First):_____________________ ID# ______________

PROBLEM SET #3 STATISTICS 500 Fall 2018: ANSWER PAGE 2

This is an exam. Do not discuss it. Due 20 December 2018

Part 2 **Use the abbkk data for part 2.**	Fill in / **circle** the answer
2.1 In model 2, test the null hypothesis H_0: $0=\gamma_1=\gamma_2=\gamma_3=\gamma_4$. What is the name of the test? What is the value of the test statistic? What are the degrees of freedom? What is the P-value? Is the null hypothesis plausible?	Name:________ Value:________ Degrees of freedom:_(____,____) P-value:_______ Circle one Plausible Not Plausible
2.2 What is the maximum hatvalue in model 2? What is the minimum hatvalue?	Max=_________ Min=_________
2.3 Under model 2, compare the 4 groups in all 6 pairs. Controlling the familywise error rate at 0.05 using **Bonferroni**'s method, which groups differ significantly?	List 1 to 6 pairs or write NONE. (See data page)
2.4 Under model 2, compare the 4 groups in all 6 pairs. Controlling the familywise error rate at 0.05 using **Holm**'s method, which groups differ significantly?	List 1 to 6 pairs or write NONE. (See data page)
2.5 (See the data page) If $\mu_1 \neq \mu_2$, then what is the maximum possible number of true hypotheses among the 6 hypotheses?	Enter an integer, 0-6 Max = ____________________

3. Fill in the following anova table. See the data page.

	Sum of squares	Degrees of Freedom	Mean Square	F-ratio	P-value
Voucher					
Gender					
Interaction					
Residual				XXXXXXX XXXXXXX	XXXXXXX XXXXXXX

PROBLEM SET #3 STATISTICS 500 Fall 2018: ANSWERS

This is an exam. Do not discuss it. Due 20 December 2018

Part 1. Use Model 1 from the data page in part 1.	Fill in/**circle** the answer.
1.1 In model 1, the one child with the largest hatvalue or leverage is a 19 year old male with a bmi above 27 and at least 3 years on ssri's.	**Circle one** [✓ True] False
1.2 How many children have ssriyears > 0? How many have ssriyears = 0? In model 1, how many children have hatvalues > 6 times the mean hatvalue? Of children with hatvalues > 6 times the mean hatvalue, how many have ssriyears ≥ 1?	Give 4 counts of children ssriyears>0: 69 =0: 5984 hatvalues > 6×mean(hatvalue): 31 children with hatvalues> 6×mean(hatvalue) and ssriyears ≥ 1: 31
1.3 Test for outliers in model 1. What is the largest t-statistic in absolute value for a single-child indicator added to model 1? What are the 2 two-sided P-values for this t-statistic, one **uncorrected** for multiple tested, the other **corrected** by the Bonferroni method?	t-statistic: 5.937 Uncorrected P: 3.065×10^{-9} Corrected P: 1.855×10^{-5} Exactly two children are significant outliers at the 0.05 level by Bonferroni? **Circle one TRUE [✓ FALSE]**
1.4 Model 1 has how many submodels (see the data page)? What is the value of Mallows' C_P for model 1? Which submodel of model 1 has the smallest value of C_P? What is the value of this smallest C_P?	# Submodels: $2^7 = 128$ C_P For model 1: 8 Smallest C_P?: 4.936 **List the predictors** in the model with the smallest smallest C_P? ssriyears, mtage, female, weight, bmi
1.5 The model with the smallest C_P contains all of the predictors with nonzero coefficients, $\beta_j \neq 0$.	**Circle one** **TRUE [✓ FALSE]**
1.6 Use the diagnostic plot and the score test (from car) to check the assumption that the variance of the errors in model 1 is constant.	"It is plausible that the variance of the errors in model 1 is constant." Circle one: **TRUE [✓ FALSE]**

ANSWERS

PROBLEM SET #3 STATISTICS 500 Fall 2018: ANSWER PAGE 2

This is an exam. Do not discuss it. Due 20 December 2018

Part 2 **Use the abbkk data for part 2.**	Fill in / **circle** the answer
2.1 In model 2, test the null hypothesis H_0: $0=\gamma_1=\gamma_2=\gamma_3=\gamma_4$. What is the name of the test? What is the value of the test statistic? What are the degrees of freedom? What is the P-value? Is the null hypothesis plausible?	Name: F-test Value: 6.165 Degrees of freedom:_(3,1136) P-value:0.000372 Circle one Plausible [✓ Not Plausible]
2.2 What is the maximum hatvalue in model 2? What is the minimum hatvalue?	Max=0.003509 Min=0.003509
2.3 Under model 2, compare the 4 groups in all 6 pairs. Controlling the familywise error rate at 0.05 using **Bonferroni's** method, which groups differ significantly?	List 1 to 6 pairs of group **names** or write NONE. (See data page) (MaleNoV,FemaleV) (MaleV,MaleNoV)
2.4 Under model 2, compare the 4 groups in all 6 pairs. Controlling the familywise error rate at 0.05 using **Holm's** method, which groups differ significantly?	List 1 to 6 pairs of group **names** or write NONE. (See data page) (MaleNoV,FemaleV) (MaleV,MaleNoV) (MaleNoV,FemaleNoV)
2.5 (See the data page) If $\mu_1\neq\mu_2$, then what is the maximum possible number of true hypotheses among the 6 hypotheses?	Max = 3. For instance, we could have $\mu_2=\mu_3$ and $\mu_2=\mu_4$ and $\mu_3=\mu_4$. Holm pays for 5 possibly true hypotheses, but there are ≤ 3.

3. Fill in the following anova table. See the data page.

	Sum of squares	Degrees of Freedom	Mean Square	F-ratio	P-value
Voucher	10.61	1	10.61	11.88	0.000586
Gender	2.95	1	2.95	3.30	0.069
Interaction	2.95	1	2.95	3.30	0.069
Residual	1014.43	1136	0.893	XXXXXXX XXXXXXX	XXXXXXX XXXXXXX

Doing the Problem Set in R

Problem 3, Fall 2018

```
bonessri<-read.csv("bonessri.csv")
dim(bonessri)
attach(bonessri)
mtransform<-function(y,p){
  m<-median(y)
  if (p!=0)  m+((y^p)-(m^p))/(p*(m^(p-1)))
  else m*(1+log(y)-log(m))}
mtage<-mtransform(age,-.4)
interact<-mtage*female
# Part 1
m1<-
lm(femurbmd~ssriyears+mtage+female+interact+povertyr+weight
+bmi)
hv<-hatvalues(m1)
bonessri[which.max(hv),] # Question 1.1
# Question 1.2
table(ssriyears>0)
ssriyears[which(hv>(6*mean(hv)))]
length(ssriyears[which(hv>(6*mean(hv)))])
min(ssriyears[which(hv>(6*mean(hv)))])
# Question 1.3
library(car)
outlierTest(m1)
summary(rstudent(m1))
dim(bonessri)
qt(.025/6053,6044)
# Question 1.4
library(leaps)
x<-
data.frame(ssriyears,mtage,female,interact,povertyr,weight,
bmi)
subreg<-leaps(x=x,y=femurbmd,names=colnames(x))
which.min(subreg$Cp)
min(subreg$Cp)
subreg$size[38]
subreg$which[38,]
# Question 1.6
ncvTest(m1)
plot(m1) # Third plot
```

```
abbkk<-read.csv("abbkk.csv")
attach(abbkk)
table(female,grp)
table(voucher,grp)
# Question 3.1
m2<-aov(hgrade~grp)
summary(m2)
summary(hatvalues(m2))
# Questions 3.3 and 3.4
pairwise.t.test(hgrade,grp,p.adjust.method="bonf")
pairwise.t.test(hgrade,grp,p.adjust.method="holm")
contrasts(grp)
gender<-c(1,1,-1,-1)
voucher<-c(-1,1,-1,1)
inter<-gender*voucher
contrasts(grp)<-cbind(gender,voucher,inter)
contrasts(grp)
mm<-model.matrix(hgrade~grp)
gender1<-mm[,2]
voucher1<-mm[,3]
inter1<-mm[,4]
m3<-lm(hgrade~gender1+voucher1+inter1)
cor(cbind(gender1,voucher1,inter1))
anova(m3)
```

PROBLEM SET #1 STATISTICS 500 Fall 2017: DATA PAGE 1

Due at noon on Tuesday, 24 Oct 2017 in class.

This is an exam. Do not discuss it with anyone.

The data are from NHANES 2009-2010. You can obtain the original data at https://www.cdc.gov/nchs/nhanes/ , but there is no reason to do this unless you want to. The data relate lung function to smoking. The data are in an object **smokelung** in the course workspace. You will have to download the workspace again, and may need to clear your web browser's cache to get the new version. The data are also available as a csv file on my web page using the button data.cv. There are 2360 people, of whom 1842 never smoked and 518 are daily smokers, meaning that they smoked at least 5 cigarettes per day every day of the last 30 days. For an explanation of the lung function measures, fvc, fev1, and ratio = fev1/fvc, see http://oac.med.jhmi.edu/res_phys/Encyclopedia/ForcedExpiration/ForcedExpiration.HTML For an explanation of bmi, see https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm The variables are

SEQN = NHANES id number
fvc = forced vital capacity in ml
fev1 = forced expiratory volume in 1 second, in ml
ratio = fev1/fvc
cigsperday = cigarettes smoked per day, 0 for never smokers
smoke = daily or never
female = 1 for female, 0 for male
age = age in years, >=20
bmi = body mass index
educ, educf = education, 1=<9th grade, 2=9-11th grade, 3=high school or equivalent, 4="some college", say a 2 year associates degree, 5="college", >= 4 year BA degree.
income, incomef = ratio of family income to the poverty level, capped at 5xPoverty.
cotinine = cotinine in blood, ng/ml, a marker for recent tobacco exposure
lead = lead in blood, ug/dL
cadmium = cadmium in blood ug/L

```
> dim(smokelung)
[1] 2360   16
> attach(smokelung)
```

You should plot the data in various ways, such as

```
plot(cigsperday,ratio)
lines(lowess(cigsperday,ratio))
boxplot(ratio~(bmi>30))
boxplot(ratio~educf)
boxplot(ratio~incomef)
```

PROBLEM SET #1 STATISTICS 500 Fall 2017: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due at noon on Tuesday, 24 Oct 2017 in class.

Model 1:

ratio= $\beta_0+\beta_1$cigsperday$+\beta_2$bmi$+\beta_3$female$+\beta_4$age$+\varepsilon$

where ε is iid $N(0,\sigma^2)$.

Important: Write your name on both sides of the answer page, **last name first**. **Turn in only the answer page**. Do not turn in plots. Brief answers suffice. **Circle** the correct answer, but do not cross out an answer. A circled answer may be correct or incorrect, but every crossed out answer is incorrect. Do not give one answer adding a note explaining why a different answer is correct.

This is an exam. Do not discuss the exam with anyone. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name (**Last**, First):____________________ ID# ____________

PROBLEM SET #1 STATISTICS 500 Fall 2017: ANSWER PAGE 1

This is an exam. Do not discuss it. Due noon Oct 24.

Part 1.	Fill in or **circle** the correct answer.
1.1 What is the value of the smallest ratio? How old is this person? Does this person smoke?	ratio = ______ age = _______ **Circle one** Daily-smoker Never-smoker
1.2 What is the value of the largest bmi? How old is this person? Does this person smoke?	bmi = ______ age = _______ **Circle one** Daily-smoker Never-smoker
1.3 The distribution of ratio is skewed right, with a longer tail to the right (high ratios) than to the left (low ratios).	**Circle one** True False

Fit **model 1** on the data page. Use model 1 for the questions in part 2.	Fill in or **circle** the correct answer.
2.1 Test the null hypothesis $H_0: \beta_1 = 0$, that the coefficient of cigsperday is zero. What is the **name** of the test, the **value** of the test statistic, the 2-sided **P-value**. Is H_0 **plausible**?	Name:_________ Value:_______ P-value:________ **Circle one** Plausible Not plausible
2.2 Give the 2-sided 95% confidence interval for β_1, the coefficient of cigsperday.	[_____ , _____]
2.3 Test the null hypothesis $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. What is the **name** of the test, the **value** of the test statistic, the 2-sided **P-value**. Is H_0 **plausible**?	Name:_________ Value:_______ P-value:________ **Circle one** Plausible Not plausible
2.4 What is the value of the correlation between ratio and its fitted values from model 1? What is the square of this correlation?	Correlation = _________ Correlation^2= _________

Name (**Last**, First):_____________________ ID# _____________

PROBLEM SET #1 STATISTICS 500 Fall 2017: ANSWER PAGE 1

This is an exam. Do not discuss it. Due noon Oct 24.

3.1 In model 1, test the null hypothesis $H_0: \beta_2 = \beta_3 = \beta_4 = 0$. Fill in the anova table.

Source of Variation	Sums of square	Degrees of freedom	Mean square	F-ratio
Full Model				
cigsperday Alone				xxxxxxxxxx xxxxxxxxxx
Added by bmi, age, and female				
Residual from full model				xxxxxxxxxx xxxxxxxxxx

3.2 In model 1, test the null hypothesis $H_0: \beta_2 = \beta_3 = \beta_4 = 0$. Fill in the anova table. What is the **name** of the test, the **value** of the test statistic, the 2-sided **P-value**. Is H_0 **plausible**?	Name:_________ Value:_______ P-value:________ **Circle one** Plausible Not plausible

Use model 1 to answer the following questions.	Fill in or **circle** the correct answer
4.1 The plot of residuals against fitted values shows a clear U shape, with the largest positive residuals at the largest and smallest fitted values.	**Circle one** True False
4.2 The Normal quantile plot of residuals indicates that they look Normal.	**Circle one** True False
4.3 There are 8 residuals <= -0.3, and no residuals >=0.3	**Circle one** True False
4.4 The Shapiro-Wilk test applied to the residuals from model 1 accepts, at the 0.05 level, the null hypothesis that the residuals are Normal.	**Circle one** True False

ANSWERS
PROBLEM SET #1 STATISTICS 500 Fall 2017: ANSWER PAGE 1
This is an exam. Do not discuss it. Due noon Oct 24.

Part 1. 7 points each, except as noted.	Fill in or **circle** the correct answer.
1.1 What is the value of the smallest ratio? How old is this person? Does this person smoke?	ratio = 0.3130 age = 69 **Circle one** [✓ Daily-smoker] Never-smoker
1.2 What is the value of the largest bmi? How old is this person? Does this person smoke?	bmi = 82.1 age = 64 **Circle one** Daily-smoker [✓ Never-smoker]
1.3 The distribution of ratio is skewed right, with a longer tail to the right (high ratios) than to the left (low ratios).	**Circle one** True [✓ False]

Fit **model 1** on the data page. Use model 1 for the questions in part 2.	Fill in or **circle** the correct answer.
2.1 Test the null hypothesis $H_0: \beta_1 = 0$, that the coefficient of cigsperday is zero. What is the **name** of the test, the **value** of the test statistic, the 2-sided **P-value**. Is H_0 **plausible**?	Name:t-test Value:-13.85 P-value: 2x10^{-16} **Circle one** Plausible [✓ Not plausible]
2.2 Give the 2-sided 95% confidence interval for β_1, the coefficient of cigsperday.	[-.00306,-0.00230]
2.3 Test the null hypothesis $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. What is the **name** of the test, the **value** of the test statistic, the 2-sided **P-value**. Is H_0 **plausible**?	Name: F-test Value: 198.6 P-value: 2x10^{-16} **Circle one** Plausible [✓ Not plausible]
2.4 What is the value of the correlation between ratio and its fitted values from model 1? What is the square of this correlation?	Correlation = 0.502 Correlation^2= 0.252

ANSWERS

PROBLEM SET #1 STATISTICS 500 Fall 2017: ANSWER PAGE 1

This is an exam. Do not discuss it. Due noon Oct 24.

3.1 In model 1, test the null hypothesis H_0: $\beta_2 = \beta_3 = \beta_4 = 0$. Fill in the anova table. (16 points)

Source of Variation	Sums of square	Degrees of freedom	Mean square	F-ratio
Full Model	3.8825	4	0.970625	198.6
cigsperday Alone	1.1541	1	1.1541	xxxxxxxxxx xxxxxxxxxx
Added by bmi, age, and female	2.7284	3	0.9094667	186.12
Residual from full model	11.508	2355	0.00489	xxxxxxxxxx xxxxxxxxxx

3.2 In model 1, test the null hypothesis H_0: $\beta_2 = \beta_3 = \beta_4 = 0$. Fill in the anova table. What is the **name** of the test, the **value** of the test statistic, the 2-sided **P-value**. Is H_0 **plausible**?	Name:F-test Value:186.12 P-value: 2×10^{-16} **Circle one** Plausible [✓ Not plausible]

Use model 1 to answer the following questions.	Fill in or **circle** the correct answer
4.1 The plot of residuals against fitted values shows a clear U shape, with the largest positive residuals at the largest and smallest fitted values.	**Circle one** True [✓ False]
4.2 The Normal quantile plot of residuals indicates that they look Normal.	**Circle one** True [✓ False]
4.3 There are 8 residuals <= -0.3, and no residuals >=0.3	**Circle one** [✓ True] False
4.4 The Shapiro-Wilk test applied to the residuals from model 1 accepts, at the 0.05 level, the null hypothesis that the residuals are Normal.	**Circle one** True [✓ False]

PROBLEM SET #1 STATISTICS 500 Fall 2017

DOING THE PROBLEM SET IN R

```
attach(smokelung)
plot(cigsperday,ratio)
lines(lowess(cigsperday,ratio))
boxplot(ratio~(bmi>30))
boxplot(ratio~educf)
boxplot(ratio~incomef)
boxplot(bmi)

#Part 1
which.min(ratio)
smokelung[1744,]
which.max(bmi)
smokelung[1449,]

#Model 1
m<-lm(ratio~cigsperday+bmi+female+age)

#Part 2
summary(m)
confint(m)
cor(m$fitted.values,ratio)
cor(m$fitted.values,ratio)^2

#Part 3
mr<-lm(ratio~cigsperday)
anova(mr,m)

#Part 4
plot(m$fitted.values,m$residuals)
lines(lowess(m$fitted.values,m$residuals),col="red")
qqnorm(m$residuals)
qqline(m$residuals)
sum(m$residuals<=-.3)
sum(m$residuals>=.3)
shapiro.test(m$residuals)
detach(smokelung)
```

PROBLEM SET #2 STATISTICS 500 Fall 2017: DATA PAGE 1

Due at noon on Tuesday, 24 Oct 2017 in class.

This is an exam. Do not discuss it with anyone.

The data are the same as Problem Set 1, from NHANES 2009-2010. The data relate lung function to smoking. The data are in an object **smokelung** in the course workspace. The data are also available briefly as a csv file on my web page using the button data.cv. There are 2360 people, of whom 1842 never smoked and 518 are daily smokers, meaning that they smoked at least 5 cigarettes per day every day of the last 30 days. For an explanation of the lung function measures, fvc, fev1, and ratio = fev1/fvc, see http://oac.med.jhmi.edu/res_phys/Encyclopedia/ForcedExpiration/ForcedExpiration.HTML For an explanation of bmi, see https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm The variables are

SEQN = NHANES id number
fvc = forced vital capacity in ml
fev1 = forced expiratory volume in 1 second, in ml
ratio = fev1/fvc
cigsperday = cigarettes smoked per day, 0 for never smokers
smoke = daily or never
female = 1 for female, 0 for male
age = age in years, >=20
bmi = body mass index
educ, educf = education, 1=<9th grade, 2=9-11th grade, 3=high school or equivalent, 4="some college", say a 2 year associates degree, 5="college", >= 4 year BA degree.
income, incomef = ratio of family income to the poverty level, capped at 5xPoverty.
cotinine = cotinine in blood, ng/ml, a marker for recent tobacco exposure
lead = lead in blood, ug/dL
cadmium = cadmium in blood ug/L

```
> dim(smokelung)
[1] 2360   16
> attach(smokelun
```

As always, you should plot the data in various ways. **IMPORTANT**: If you look at the interaction of a binary (1 or 0) variable and a continuous variable, then **do not** center either variable before multiplying. You may mess up several questions if you do this incorrectly. Also, make sure you construct **mini** as described below.

PROBLEM SET #2 STATISTICS 500 Fall 2017: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due at noon on Tuesday, 24 Oct 2017 in class.

You will need to construct several new variables.

```
smoker<-1*(smokelung$smoke=="Daily")
smokerAge<-smoker*smokelung$age
d<-cbind(smokelung,smoker,smokerAge)
rm(smoker,smokerAge)
attach(d)
mini<-d[c(1781,1726,554,1394),]
mini<-mini[,c(8,9,18,19)]
mini
     female age smoker smokerAge
1781      1  25      0         0
1726      1  25      1        25
554       1  60      0         0
1394      1  60      1        60
```

So, mini is a data.frame with just 4 people. Look at mini carefully, so you understand it. It has 4 females, two aged 25, two aged 60, two daily smokers, two never smokers.

Model 1: (based on d)

ratio= $\beta_0+\beta_1$age$+\beta_2$female$+\beta_3$smoker$+\varepsilon$

where ε is iid $N(0,\sigma^2)$.

Model 2: (based on d)

ratio= $\gamma_0+\gamma_1$age$+\gamma_2$female$+\gamma_3$smoker$+\gamma_4$smokerAge$+\varepsilon$

where ε is iid $N(0,\omega^2)$.

For question 2.2, use the **restricted cubic spline** function, rcspline.eval(), in the Hmisc package, using the default to let the function decide the spacing of the 5 knots. The question asks you test for curvature in **age** in model 2. For understanding, you might plot the fitted values against age for the two models in 2.2.

For question 3.1, the **studentized residual** is obtained using **rstudent**() in R. It uses the deleted estimate of ω^2.

Important: Write your name on both sides of the answer page, **last name first**. **Turn in only the answer page**. Brief answers suffice. **Circle** the correct answer, but do not cross out an answer. A circled answer may be correct or incorrect, but every crossed out answer is incorrect.

This is an exam. Do not discuss the exam with anyone. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name (**Last**, First):____________________ ID# ____________

PROBLEM SET #2 STATISTICS 500 Fall 2017: ANSWER PAGE 1

This is an exam. Do not discuss it. Due noon Oct 24.

<table>
<tr><th></th><th>Fill in/CIRCLE the answer</th></tr>
<tr><td>1.1 Fit model 1 and use Tukey’s test to see if a transformation of y would be needed if this model were used. Give the value of the t-statistic and the P-value. Does the test indicate that a transformation is needed? Repeat the calculation for model 2, giving just the P-value.</td><td>Model 1:
t-statistic: ___________

P-value: _______________
Transformation is:
(CIRCLE ONE)
NEEDED NOT NEEDED

Model 2:
P-value: _______________</td></tr>
<tr><td>1.2 In models 1 and 2, is it plausible that the relationship between the relationship between ratio and age is parallel for smokers and nonsmokers? Give the name of the test statistic, its numerical value, its degrees of freedom, its P-value and state whether parallelism is plausible.</td><td>Name of test:_____________
Value of the
test statistic: __________

Degrees of freedom: ______

P-value: _________________
Parallelism is:
(CIRCLE ONE)
PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>1.3 The estimated coefficient of smoker in model 2 is positive and not significantly different from zero. Recall that y=ratio is a measure of lung function.</td><td>From the fact on the left, we can reasonably conclude that model 2 provides no indication that smoking is associated with lower values of ratio. CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>1.4 Use the predict function applied to model 2 and to mini to estimate the expected ratio for female smokers and nonsmokers aged 25 and 60. Put the four estimates in the table.</td><td><table><tr><th></th><th>age=25</th><th>age=60</th></tr><tr><td>smoke=0</td><td></td><td></td></tr><tr><td>smoke=1</td><td></td><td></td></tr></table>(Round to 2 digits after the decimal.)</td></tr>
<tr><td>1.5 Use the predict function applied to model 2 and to mini to obtain two-sided 95% confidence intervals for the expected ratio in part 1.4.</td><td><table><tr><th></th><th>age=25</th><th>age=60</th></tr><tr><td>smoke=0</td><td></td><td></td></tr><tr><td>smoke=1</td><td></td><td></td></tr></table>(Round to 2 digits).</td></tr>
</table>

Name (**Last**, First):____________________ ID# ____________

PROBLEM SET #2 STATISTICS 500 Fall 2017: ANSWER PAGE 2

This is an exam. Do not discuss it. Due noon Oct 24.

	Fill in/**CIRCLE** the answer
2.1 Who said: “To find out what happens to a system when you interfere with it you have to interfere with it”? Circle **ONE**.	Gauss Cauchy Sheather Box Cox Fox Cook Fisher Markov Tukey Harrell Everitt Weisberg Ripley Titterington Trump
2.2 Add to model #2 **restricted cubic splines** in **age** with 5 knots. **See the data page.** Test that model 2 is adequate against the alternative that the splines in age are needed. Give the **name** of the test, the **value** of the test statistic, the **degrees of freedom (df)**, the **P-value**, and state whether the test **indicates curvature** in age.	Name: _______ Value:_______ Degrees of Freedom: ________ P-value: __________ Indicates curvature in age? CIRCLE ONE YES NO

Use **model 2** for part 3.	Fill in/**CIRCLE** the answer
3.1 Which observation in model 2 has the largest **absolute** studentized residual? Give **signed** studentized residual, the row number and SEQN, and circle TRUE or FALSE about “This person is an outlier at the 0.05 level, taking account of the number of tests done.” Give **DF** and adjusted **P-value** for test.	Row:________ SEQN:________ Studentized Residual:______ This person has the **lowest ratio**. CIRCLE: TRUE FALSE **Outlier**: CIRCLE: TRUE FALSE DF=degrees of freedom: DF=__________ Adusted P-value:____________
3.2 Which observation has the largest **absolute** dffits? Give **signed** dffits, the row number and SEQN, and circle TRUE or FALSE. Does this observation move its yhat up 4.48 standard errors.	Row:________ SEQN:________ dffits:____________________ CIRCLE: YES NO
3.3 Which observation in model 2 has the largest **leverage**? Give **hatvalue**, row # and SEQN.	hatvalue:_______________ Row:________ SEQN:________

PROBLEM SET #2 STATISTICS 500 Fall 2017: ANSWER PAGE 1

This is an exam. Do not discuss it.

<table>
<tr><th></th><th>Fill in/**CIRCLE** the answer</th></tr>
<tr><td>1.1 Fit model 1 and use Tukey’s test to see if a transformation of y would be needed if this model were used. Give the value of the t-statistic and the P-value. Does the test indicate that a transformation is needed? **Repeat** the calculation for model 2, giving just the P-value.</td><td>**Model 1**:
t-statistic: -4.195

P-value: <0.001
Transformation is:
(CIRCLE ONE)
[✓ NEEDED] NOT NEEDED

Model 2:
P-value: 0.596</td></tr>
<tr><td>1.2 In models 1 and 2, is it plausible that the relationship between the relationship between ratio and age is parallel for smokers and nonsmokers? Give the name of the test statistic, its numerical value, its degrees of freedom, its P-value and state whether parallelism is plausible.</td><td>Name of test: t-test
Value of the
test statistic: -5.530

Degrees of freedom: 2355

P-value: 3.56 x 10^{-8}
Parallelism is:
(CIRCLE ONE)
PLAUSIBLE [✓ NOT PLAUSIBLE]</td></tr>
<tr><td>1.3 The estimated coefficient of smoker in model 2 is positive and not significantly different from zero. Recall that y=ratio is a measure of lung function.</td><td>From the fact on the left, we can reasonably conclude that model 2 provides no indication that smoking is associated with lower values of ratio. CIRCLE ONE
TRUE [✓ FALSE]</td></tr>
<tr><td>1.4 Use the predict function applied to model 2 and to mini to estimate the expected ratio for female smokers and nonsmokers aged 25 and 60. Put the four estimates in the table.</td><td><table>
<tr><th></th><th>age=25</th><th>age=60</th></tr>
<tr><td>smoke=0</td><td>0.85</td><td>0.78</td></tr>
<tr><td>smoke=1</td><td>0.82</td><td>0.71</td></tr>
</table>(Round to 2 digits)</td></tr>
<tr><td>1.5 Use the predict function applied to model 2 and to mini to obtain two-sided 95% **confidence intervals for the expected ratio** in part 1.4.</td><td><table>
<tr><th></th><th>age=25</th><th>age=60</th></tr>
<tr><td>smoke=0</td><td>[.84,.85]</td><td>[.78,.79]</td></tr>
<tr><td>smoke=1</td><td>[.81,.83]</td><td>[.70,.72]</td></tr>
</table>(Round to 2 digits).</td></tr>
</table>

PROBLEM SET #2 STATISTICS 500 Fall 2017: ANSWER PAGE 2

This is an exam. Do not discuss it. Due noon Oct 24.

	Fill in/**CIRCLE** the answer
2.1 Who said: “To find out what happens to a system when you interfere with it you have to interfere with it”? Circle **ONE**.	Gauss Cauchy Sheather [✓ Box] Cox Fox Cook Fisher Markov Tukey Harrell Everitt Weisberg Ripley Titterington Trump
2.2 Add to model #2 **restricted cubic splines** in **age** with 5 knots. **See the data page.** Test that model 2 is adequate against the alternative that the splines in age are needed. Give the **name** of the test, the **value** of the test statistic, the **degrees of freedom (df)**, the **P-value**, and state whether the test **indicates curvature** in age.	Name: F-test Value:4.04 Degrees of Freedom: 3 & 2352 *F-tests have degrees of freedom for numerator & denominator* P-value: 0.007079 Indicates curvature in age? CIRCLE ONE [✓ YES] NO

Use **model 2** for part 3.	Fill in/**CIRCLE** the answer
3.1 Which observation in model 2 has the largest **absolute** studentized residual? Give **signed** studentized residual, the row number and SEQN, and circle TRUE or FALSE about “This person is an outlier at the 0.05 level, taking account of the number of tests done.” Give **DF** and adjusted **P-value** for test.	Row: 1212 SEQN: 67220 Studentized Residual: -6.49 This person has the **lowest ratio**. CIRCLE: TRUE [✓ FALSE] **Outlier**: CIRCLE: [✓ TRUE] FALSE DF=degrees of freedom: DF= 2354 Adusted P-value: 2.4591x10^{-7}
3.2 Which observation has the largest **absolute** dffits? Give **signed** dffits, the row number and SEQN, and circle TRUE or FALSE. Does this observation move its yhat up 4.48 standard errors.	Row: 1744 SEQN: 69468 dffits: -0.4484 CIRCLE: YES [✓ NO]
3.3 Which observation in model 2 has the largest **leverage**? Give **hatvalue**, row # and SEQN.	hatvalue: 0.013067 Row: 674 SEQN: 65071

Doing the Problem Set in R

```
smoker<-1*(smokelung$smoke=="Daily")
smokerAge<-smoker*smokelung$age
d<-cbind(smokelung,smoker,smokerAge)
rm(smoker,smokerAge)
attach(d)
mini<-d[c(1781,1726,554,1394),]
mini<-mini[,c(8,9,18,19)]

m1<-lm(ratio~age+female+smoker)
m2<-lm(ratio~age+female+smoker+smokerAge)
library(car)
#1.1
residualPlots(m1)
residualPlots(m2)
#1.2
summary(m2)
#1.4 and 1.5
predict(m2,mini,interval="confidence")
#2.1  Last sentence of Box's Use and Abuse of Regression.
(Assigned reading.)
#2.2
library(Hmisc)
sp<-rcspline.eval(age)
m3<-lm(ratio~age+female+smoker+smokerAge+sp)
anova(m2,m3)
#For understanding
plot(age,m2$fitted.values)
plot(age,m3$fitted.values)
#3.1
outlierTest(m2)
which.max(abs(rstudent(m2)))
rstudent(m2)[1212]
d[1212,]
m2$df.residual-1
#3.2
which.max(abs(dffits(m2)))
dffits(m2)[1744]
d[1744,]
#3.3
which.max(hatvalues(m2))
hatvalues(m2)[674]
d[674,]
```

PROBLEM SET #3 STATISTICS 500 Fall 2017: DATA PAGE 1

Due at noon on Friday 15 December 2017, Statistics Dept.

This is an exam. Do not discuss it with anyone.

The first data set expands the object **smokelung** in the course workspace. **You must download the workspace again.** The data are also available briefly as a csv file on my web page using the button data.csv. There are 2360 people, of whom 1842 never smoked and 518 are daily smokers, meaning that they smoked at least 5 cigarettes per day every day of the last 30 days. For an explanation of the lung function measures, fvc, fev1, and ratio = fev1/fvc, see http://oac.med.jhmi.edu/res_phys/Encyclopedia/ForcedExpiration/ForcedExpiration.HTML For an explanation of bmi, see https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm The variables are

SEQN = NHANES id number
fvc = forced vital capacity in ml
fev1 = forced expiratory volume in 1 second, in ml
ratio = fev1/fvc
cigsperday = cigarettes smoked per day, 0 for never smokers
smoke = daily or never
female = 1 for female, 0 for male
age = age in years, >=20
bmi = body mass index
educ = education, 1=<9th grade, 2=9-11th grade, 3=high school or equivalent, 4="some college", say a 2 year associates degree, 5="college", >= 4 year BA degree.
income = ratio of family income to the poverty level, capped at 5xPoverty.

You will need to construct several new variables.

```
smoker<-1*(smokelung$smoke=="Daily")
smokerAge<-smoker*smokelung$age
d<-cbind(smokelung,smoker,smokerAge)
rm(smoker,smokerAge)
```

Use randomhalf in d to split the data into two parts.
Use the random division in d; DO NOT make a new one.
One half is used for exploration, dex. The other half is used for validation, dva.

```
dex<-d[d$randomhalf=="Explore",]
dva<-d[d$randomhalf=="Validate",]
dim(d)
[1] 2360   20
dim(dex)
[1] 1180   20
dim(dva)
[1] 1180   20
```

PROBLEM SET #3 STATISTICS 500 Fall 2017: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due at noon on Friday 15 December 2017, Statistics Dept.

Model 1:

$\text{ratio}= \beta_0+\beta_1\text{age}+\beta_2\text{female}+\beta_3\text{smoker}+\beta_4\text{smokerAge}$
$\quad +\beta_5\text{bmi}+\beta_6\text{educ}+\beta_7\text{income}+\varepsilon$

where ε is iid $N(0,\sigma^2)$.

You will work with Model 1 and its 2^7=128 submodels (including Model 1 and the model with no predictors). In **question 1**, you use the exploratory random half, dex. In **question 2**, you check what you found using the validation half, dva. Spjotvoll's method corrects for multiple testing, so you don't have to split the data to use it. In **question 3**, you use all of the data in d, 2360 observations, and all 2^7=128 submodels in Spjotvoll's method. Use Spjotvoll's method at simultaneous level 0.05, so the chance of rejecting a true model as inadequate is at most 5% despite looking at 2^7=128 submodels.

The **second data set** is **nls500**. It is from the National Longitudinal Study (NLS). Decades ago, before charter schools, people used the NLS comparing Catholic high schools and public high schools. The data you have are simplified in several ways, so don't use these data to decide about the education of your children. There are three variables. **income50** indicates whether family income was < or >= $50,000. **school** is Catholic or public. **math** is the change in a math test score from before high school to the end of high school. Define **insc<-income50:school**.

```
> attach(nls500)          > table(income50,school)
> insc<-income50:school   > boxplot(math~insc)
```

Important: In question 4, refer to "<50000:Catholic" as "<C", …, ">=50000:Public" as ">P".

Model 2: $\text{math}_{ij}=\mu+\tau_j+\varepsilon_{ij}$, i=1,…,163, j=1,…,4, ε iid $N(0,\sigma^2)$

Important: Write your name on both sides of the answer page, **last name first**. **Turn in only the answer page**. Brief answers suffice. **Circle** the correct answer. A circled answer may be correct or incorrect, but every crossed out answer is incorrect. **Turn in the exam** at noon on Friday 15 December 2017 in the Statistics Dept, 4th floor Huntsman. Give to Noelle at the front desk or place in an envelope addressed to me in my mailbox in Statistics. **Make and keep a photocopy of your answer page**. The answer key will be posted on-line in the revised bulk-pack.

This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam. **Have a great holiday!**

Name (**Last**, First):____________________ ID# ____________

PROBLEM SET #3 STATISTICS 500 Fall 2017: ANSWER PAGE 1

This is an exam. Do not discuss it. Due noon December 15

1. For questions in part 1, **use only the exploratory half of the data in dex**, n=1180. Do not use dva.	FILL IN OR CIRCLE THE CORRECT ANSWER
1.1 Fit model 1 to the 1180 observations in dex. Which predictor (x variable) has the largest variance-inflation-factor (VIF)? What is its numerical value? What is the R^2 for this predictor as derived from VIF. CICLE True or False.	Predictor:____________ VIF value:____________ R^2 Value:______________ This VIF value means that this predictor is a poor predictor of y=ratio. CIRCLE ONE TRUE FALSE
1.2 When fitted to the 1180 observations in dex, which model P of the 2^7 submodels of model 1 has the smallest C_P? List the variables in P, give the value of C_P, the "size" of the model.	Predictors in P: _____________________________ C_P=____________ size=_______
1.3 C_P estimates of J_P. If the true J_P equaled its estimate, C_P, then the total expected squared error for the model in 1.2 would be more than twice as large as model 1 with 7 predictors.	CIRCLE ONE TRUE FALSE
1.4 The value of C_P in 1.2 suggests that the model in 1.2 omits at least one important predictor.	CIRCLE ONE TRUE FALSE

In 2, use dva, not dex.	Fill in or CIRCLE the answer
2.1 Fit the model in 1.2 to the 1180 validation observations in dva. List the **predictors** and their **t-statistics.**	Variable names t-statistics
2.2 Using dva, what is the correlation between the fitted values in 2.1 and the fitted values for model 1?	Correlation: __________

Name (**Last**, First):____________________ ID# ____________

PROBLEM SET #3 STATISTICS 500 Fall 2017: ANSWER PAGE 2

This is an exam. Do not discuss it. Due noon December 15

In 3 use d, not dex nor dva.	Fill in or CIRCLE the answer
3.1 Use Spjotvoll's method and all 2360 observations in d to examine the $2^7=128$ submodels of model 1. At the 0.05 level, how many of the 128 models are **not** judged "inadequate"? Which of these "not inadequate" models has the fewest predictors? (List them.)	**How many**? ______________ List the predictors in the one "not inadequate" model with the fewest predictors: **Predictor names**:

<table>
<tr><th>In 4, use the nls500 data.</th><th>Fill in or CIRCLE the answer</th></tr>
<tr><td>4.1 Under model 2, do a 4-group one-way anova of math score changes by insc group. Test the null hypothesis $H_0:\tau_1=\tau_2=\tau_3=\tau_4=0$. Give the name of the test, the value of the test statistic, the P-value, degrees of freedom (DF) and circle an answer.</td><td>Name:_________ Value:_______
P-value:______ DF:_________
CIRCLE ONE
H_0 IS
PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>4.2 Use Holm's method to test 6 hypotheses $H_0:\tau_j=\tau_k$ controlling the familywise error rate at 0.05. Use the notation from the data page (e.g., "<C") to indicate pairs of that differ significantly (eg "(<C, >C)"</td><td>List all pairs of groups that differ significantly. If none, write "none".</td></tr>
<tr><td>4.3 Give three orthogonal contrasts with integer weights for income (<50,000, >50,000), school (Catholic, Public) and their interaction.</td><td>
<table>
<tr><th></th><th><C</th><th><P</th><th>>C</th><th>>P</th></tr>
<tr><td>Income</td><td></td><td></td><td></td><td></td></tr>
<tr><td>School</td><td></td><td></td><td></td><td></td></tr>
<tr><td>Interaction</td><td></td><td></td><td></td><td></td></tr>
</table>
</td></tr>
<tr><td>4.4 Test the hypothesis that the interaction contrast in the τ_j does not differ significantly from zero.</td><td>P-value:______
CIRCLE ONE: H_0 IS
PLAUSIBLE NOT PLAUSIBLE</td></tr>
</table>

PROBLEM SET #3 STATISTICS 500 Fall 2017: ANSWER PAGE 1

This is an exam. Do not discuss it.

1. For questions in part 1, **use only the exploratory half of the data in dex**, n=1180. Do not use dva.	FILL IN OR CIRCLE THE CORRECT ANSWER
1.1 Fit model 1 to the 1180 observations in dex. Which predictor (x variable) has the largest variance-inflation-factor (VIF)? What is its numerical value? What is the R^2 for this predictor as derived from VIF. CICLE True or False.	Predictor: smokerAge VIF value: 10.1410 R^2 Value: 0.90139 This VIF value means that this predictor is a poor predictor of y=ratio. CIRCLE ONE TRUE [✓ FALSE]
1.2 When fitted to the 1180 observations in dex, which model P of the 2^7 submodels of model 1 has the smallest C_P? List the variables in P, give the value of C_P, the “size” of the model.	Predictors in P: age, female, smokerAge C_P=1.92 size=3
1.3 C_P estimates of J_P. If the true J_P equaled its estimate, C_P, then the total expected squared error for the model in 1.2 would be more than twice as large as model 1 with 7 predictors.	CIRCLE ONE TRUE [✓ FALSE]
1.4 The value of C_P in 1.2 suggests that the model in 1.2 omits at least one important predictor.	CIRCLE ONE TRUE [✓ FALSE]

In 2, use dva, not dex.	Fill in or CIRCLE the answer
2.1 Fit the model in 1.2 to the 1180 validation observations in dva. List the **predictors** and their **t-statistics.**	Variable names t-statistics age -16.47 female 4.36 smokerAge -11.62
2.2 Using dva, what is the correlation between the fitted values in 2.1 and the fitted values for model 1?	Correlation: 0.9897

PROBLEM SET #3 STATISTICS 500 Fall 2017: ANSWER PAGE 2

This is an exam. Do not discuss it.

In 3 use d, not dex nor dva.	Fill in or CIRCLE the answer
3.1 Use Spjotvoll's method and all 2360 observations in d to examine the 2^7=128 submodels of model 1. At the 0.05 level, how many of the 128 models are **not** judged "inadequate"? Which of these "not inadequate" models has the fewest predictors? (List them.)	**How many**? 16 List the predictors in the one "not inadequate" model with the fewest predictors: **Predictor names**: age, female, smokerAge (Every adequate model includes these three!)

<table>
<tr><th>In 4, use the nls500 data.</th><th>Fill in or CIRCLE the answer</th></tr>
<tr><td>4.1 Under model 2, do a 4-group one-way anova of math score changes by insc group. Test the null hypothesis $H_0:\tau_1=\tau_2=\tau_3=\tau_4=0$. Give the name of the test, the value of the test statistic, the P-value, degrees of freedom (DF) and circle an answer.</td><td>Name: F-test Value: 5.085
P-value:0.00174 DF: 3, 648
CIRCLE ONE
H_0 IS
PLAUSIBLE [✓ NOT PLAUSIBLE]</td></tr>
<tr><td>4.2 Use Holm's method to test 6 hypotheses $H_0:\tau_j=\tau_k$ controlling the familywise error rate at 0.05. Use the notation from the data page (e.g., "<C") to indicate pairs of that differ significantly (eg "(<C, >C)"</td><td>List all pairs of groups that differ significantly. If none, write "none".
(<C, <P)
(>C, <P)</td></tr>
<tr><td>4.3 Give three orthogonal contrasts with integer weights for income (<50,000, >50,000), school (Catholic, Public) and their interaction.</td><td>
<table>
<tr><th></th><th><C</th><th><P</th><th>>C</th><th>>P</th></tr>
<tr><td>Income</td><td>1</td><td>1</td><td>-1</td><td>-1</td></tr>
<tr><td>School</td><td>1</td><td>-1</td><td>1</td><td>-1</td></tr>
<tr><td>Interaction</td><td>1</td><td>-1</td><td>-1</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>4.4 Test the hypothesis that the interaction contrast in the τ_j does not differ significantly from zero.</td><td>P-value: 0.31830
CIRCLE ONE: H_0 IS
[✓ PLAUSIBLE] NOT PLAUSIBLE</td></tr>
</table>

Doing the Problem Set in R: Problem 3, Fall 2017

```
smoker<-1*(smokelung$smoke=="Daily")
smokerAge<-smoker*smokelung$age
d<-cbind(smokelung,smoker,smokerAge)
rm(smoker,smokerAge)
dex<-d[d$randomhalf=="Explore",]
dva<-d[d$randomhalf=="Validate",]

attach(dex)
maxm<-lm(ratio~age+female+smoker+smokerAge+bmi+educ+income)

#1.1
library(car)
vif(maxm)
1-1/vif(maxm)

#1.2
x<-data.frame(age,female,smoker,smokerAge,bmi,educ,income)
library(leaps)
mex<-leaps(x=x,y=ratio,names=colnames(x))
cbind(mex$which,mex$Cp,mex$size)

#2.1
detach(dex)
attach(dva)
summary(lm(ratio~age+female+smokerAge))
#2.2
cor(lm(ratio~age+female+smokerAge)$fitted,lm(ratio~age+fema
le+smoker+smokerAge+bmi+educ+income)$fitted)

#3
sp<-spjotvoll(x,ratio)
dim(sp[!sp$inadequate,])
sp[!sp$inadequate,]
```

```
#4.1
summary(aov(math~insc))

#4.2
pairwise.t.test(math,insc)

#4.3
cpschool<-c(1,-1,1,-1)
inc<-c(1,1,-1,-1)
interact<-cpschool*inc
contrasts(insc)<-cbind(cpschool,inc,interact)
contrasts(insc)

#4.4
x<-model.matrix(lm(math~insc))
head(x)
x<-as.data.frame(x)
m<-lm(math~x$insccpschool+x$inscinc+x$inscinteract)
summary(m)
anova(m)
```

PROBLEM SET #1 STATISTICS 500 FALL 2016: DATA PAGE 1

Due in class at noon in class on Thursday, 27 Oct 2016.

This is an exam. Do not discuss it with anyone.

The data are from NHANES 2013-2014 (http://www.cdc.gov/nchs/nhanes/). They describe adults, age>=20, who are not known to be pregnant.

The data are in an object in your workspace, nh14food. The data are also at data.csv on my web page (for a limited time).

The diet information describes the foods eaten on one interview day.

```
> dim(nh14food)
[1] 3996   13
```

The data for the first 3 people is as follows:

```
t(nh14food[1:3,])
person         1                2                3
seqn        73557            73558            73559
female      0                0                0
age         69               54               72
educ        High School/GED High School/GED Some College
bmi         26.7             28.6             28.9
sad         20.5             24.2             25.8
carbpct     59.59320         33.90749         51.29706
alcoholpct   0.0000          16.6625           0.0000
fatpct      29.55471         22.37556         33.93266
proteinpct 10.85209         27.05446         14.77028
calories    1574             5062             1743
fibercarb   0.04507701       0.03940724       0.04411961
sugarcarb   0.7365499        0.1061636        0.4585766
```

The variables are:

seqn is the NHANES id number.

female=1 for female, =0 for male

age in years

educ education in 5 categories

bmi = body mass index, a measure of obesity
https://www.cdc.gov/healthyweight/assessing/bmi/

sad = Sagittal abdominal diameter, another measure of obesity (cm)
https://en.wikipedia.org/wiki/Sagittal_abdominal_diameter

carbpct = percent of calories from carbs

alcoholpct = percent of calories from alcohol

fatpct = percent of calories from fat

proteinpct = percent of calories from protein

calories = calories consumed

fibercarb = (grams of fiber)/(grams of carbs)

sugarcarb = (grams of sugar)/(grams of carbs)

PROBLEM SET #1 STATISTICS 500 FALL 2016: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due in class at noon in class on Thursday, 27 Oct 2016.

There has been much discussion of the role of carbohydrates (carbs) in obesity. Think of various attempts to tax soda, for example. What can you see about this from NHANES? The food data describes one day, so it is not ideal, but it is worth a look. You should plot the data in various ways, e.g., plot(fibercarb,sad).

Model 1: sad = $\beta_0 + \beta_1$ carbpct + β_2 fibercarb + β_3 sugarcarb + ε where ε is iid $N(0, \sigma^2)$

Model 2: sad = $\gamma_0 + \gamma_1$ carbpct + γ_2 fibercarb + γ_3 sugarcarb + γ_4 age + γ_5 female + ζ where ζ is

iid $N(0, \omega^2)$

Model 1 has betas and model 2 has gammas so that different things have different symbols. It does not matter which Greek letter we use.

Follow instructions. **Write your name**, last name first, on both sides of the answer page. Also, write your id number on the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Last name:_____________________ First Name: _________________ ID#:___________________

Statistics 500, Problem 1, Fall 2016, p1. **This problem set is an exam. Do not discuss it with anyone**.

<table>
<tr><td>Part 1: Fit model 1 on the data page and assume it is true for the purpose of answering the questions in part 1.</td><td>Fill in or CIRCLE the correct answer.</td></tr>
<tr><td>1.1 The person with the largest sad consumed no calories from alcohol on the interview day.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>1.2 The person with the smallest sad consumed no calories from alcohol on the interview day.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>1.3 Give the (ordinary Pearson) correlation between sad and the three predictors, carbpct, fibercarb, and sugarcarb.</td><td><table><tr><td></td><td>carbpct</td><td>fibercarb</td><td>sugarcarb</td></tr><tr><td>sad</td><td></td><td></td><td></td></tr></table></td></tr>
<tr><td>1.4 In model 1, test the null hypothesis H_0: β_3=0 that the coefficient of sugarcarb is zero. What is the name of the test statistic? What is the numerical value of the test statistic? What are its degrees of freedom (df)? What is the two-sided P-value? Is the null hypothesis plausible?</td><td>Name:_____________ Value: _______________
Df: _______________ P-value:______________
CIRCLE ONE
PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>1.5 If the P-value testing H_0: β_3=0 were 0.001, then there is 1 chance in 1000 that β_3=0 is true.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>1.6 Give the two-sided 95% confidence interval for β_1 in model 1, the coefficient of carbpct.</td><td>[___________, ___________]</td></tr>
<tr><td>1.7 If the two-sided 95% confidence interval for β_1 excludes 0, then the two-sided P-value testing H_0: β_1=0 is less than 0.05.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>1.8 Test the null hypothesis that H_0: β_1=β_2=β_3=0 in model 1. What is the name of the test statistic? What is the numerical value of the test statistic? What are its degrees of freedom (df)? What is the two-sided P-value? Is the null hypothesis plausible?</td><td>Name:_____________ Value: _______________
Df: _______________ P-value:______________
CIRCLE ONE
PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>1.9 What fraction of the variation in sad about its mean is fitted by model 1? (That is, of the total sum of squares around the mean, what fraction is fitted by model 1?) Give a **fraction**, between 0 and 1, not a percent.</td><td>Fraction: __________________</td></tr>
<tr><td>1.10 The square root of the fraction in 1.9 is the correlation between sad and fitted-sad in model 1.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>1.11 Do a Normal quantile plot and a Shapiro-Wilk test for the residuals from model 1. What is the P-value from the Shapiro-Wilk test? Do the residuals appear to be Normal?</td><td>P-value: ________________
CIRCLE ONE
NORMAL NOT NORMAL</td></tr>
<tr><td>1.12 In question 1.11, there are many outliers in the lower tail, outliers with negative residuals.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
</table>

Last name:______________________ First Name: _________________ ID#:___________________

Statistics 500, Problem 1, Fall 2016, p2. **This problem set is an exam. Do not discuss it with anyone**.

<table>
<tr><td>Fit model 2 from the data page, and assume it is true for the purpose of answering questions in part 2.</td><td>Fill in or CIRCLE the correct answer.</td></tr>
<tr><td>2.1 In model 1, the coefficient β_2 of fibercarb differed significantly from zero at the 0.05-level, but in model 2, the coefficient γ_2 of fibercarb does not differ significantly from zero, because women eat more fiber and gender is included in model 2.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>2.2 Test the null hypothesis that $H_0:\gamma_1=\gamma_2=\gamma_3=0$ in model 2, i.e., the hypothesis that the three dietary variables have zero coefficients once we include age and gender in the model. What is the name of the test statistic? What is the numerical value of the test statistic? What are its degrees of freedom (df)? What is the two-sided P-value? Is the null hypothesis plausible?</td><td>Name:______________ Value: ________________
Df: ________________ P-value:_______________
CIRCLE ONE
PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>2.3 Adding predictors to a linear regression model – e.g., adding female and age to model 1 to get model 2 – never reduces R^2.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>2.4 When the residuals from model 2 are plotted against the fitted values, we see a clear U-shape, with too many positive residuals at small and large fitted values and too few positive residuals and mid-sized fitted values.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>2.5 Give the sample standard deviation of sad, the estimate of σ from model 1 and the estimate of ω from model 2. Note that σ and ω, not σ^2 and ω^2, are requested.</td><td><table><tr><th>Stand.dev(sad)</th><th>Estimate of σ</th><th>Estimate of ω</th></tr><tr><td></td><td></td><td></td></tr></table></td></tr>
<tr><td>2.6 Look at the absolute values of the residuals from model 2. Which individual had the largest absolute residual? Give the NHANES seqn. What is the sad for this person? What is the fitted sad from model 2? What is the (signed) value of the residual? This is a skinny man who ate lots of fiber (true or false).</td><td>Seqn: ___________ sad:_________________
Fitted sad:__________ residual:___________
CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>2.7 Use avPlots from the car package to plot model 2. The plot for fibercarb has a slope that tilts upward (positive slope), while the plot for age has a slope that tilts downward (negative slope).</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
<tr><td>2.8 Look at the form of models 1 and 2. From the mathematical form of models 1 and 2, because the first variable is carbpct in both models, without fitting the models or looking at the data, you knew that the sign of the estimate of β_1 would be the same as the sign of the estimate of γ_1.</td><td>CIRCLE ONE
TRUE FALSE</td></tr>
</table>

Statistics 500, Problem 1, Fall 2016, p1. **Answers**

<table>
<tr><td>Part 1: Fit model 1 on the data page and assume it is true for the purpose of answering the questions in part 1.</td><td>Fill in or CIRCLE the correct answer.</td></tr>
<tr><td>1.1 The person with the largest sad consumed no calories from alcohol on the interview day.</td><td>CIRCLE ONE
[✓ TRUE] FALSE</td></tr>
<tr><td>1.2 The person with the smallest sad consumed no calories from alcohol on the interview day.</td><td>CIRCLE ONE
[✓ TRUE] FALSE</td></tr>
<tr><td>1.3 Give the (ordinary Pearson) correlation between sad and the three predictors, carbpct, fibercarb, and sugarcarb.</td><td><table><tr><td></td><td>carbpct</td><td>fibercarb</td><td>sugarcarb</td></tr><tr><td>sad</td><td>-0.0337</td><td>-0.0512</td><td>0.0224</td></tr></table></td></tr>
<tr><td>1.4 In model 1, test the null hypothesis H_0: $\beta_3=0$ that the coefficient of sugarcarb is zero. What is the name of the test statistic? What is the numerical value of the test statistic? What are its degrees of freedom (df)? What is the two-sided P-value? Is the null hypothesis plausible?</td><td>Name: t-test Value: 0.899
Df: 3992 P-value: 0.369
CIRCLE ONE
[✓ PLAUSIBLE] NOT PLAUSIBLE</td></tr>
<tr><td>1.5 If the P-value testing H_0: $\beta_3=0$ were 0.001, then there is 1 chance in 1000 that $\beta_3=0$ is true.</td><td>Hypotheses are true or false, no probability involved.
TRUE [✓ FALSE]</td></tr>
<tr><td>1.6 Give the two-sided 95% confidence interval for β_1 in model 1, the coefficient of carbpct.</td><td>[-0.03217, -0.00587]</td></tr>
<tr><td>1.7 If the two-sided 95% confidence interval for β_1 excludes 0, then the two-sided P-value testing H_0: $\beta_1=0$ is less than 0.05.</td><td>TRUE FALSE
This is the connection between tests and confidence intervals</td></tr>
<tr><td>1.8 Test the null hypothesis that H_0: $\beta_1=\beta_2=\beta_3=0$ in model 1. What is the name of the test statistic? What is the numerical value of the test statistic? What are its degrees of freedom (df)? What is the two-sided P-value? Is the null hypothesis plausible?</td><td>Name: F-test Value: 6.198
Df: 3 and 3992 P-value: 0.0003381
CIRCLE ONE
PLAUSIBLE [✓ NOT PLAUSIBLE]</td></tr>
<tr><td>1.9 What fraction of the variation in sad about its mean is fitted by model 1? (That is, of the total sum of squares around the mean, what fraction is fitted by model 1?) Give a fraction, between 0 and 1, not a percent.</td><td>Fraction: 0.004636
This is just R^2</td></tr>
<tr><td>1.10 The square root of the fraction in 1.9 is the correlation between sad and fitted-sad in model 1.</td><td>CIRCLE ONE
[✓ TRUE] FALSE</td></tr>
<tr><td>1.11 Do a Normal quantile plot and a Shapiro-Wilk test for the residuals from model 1. What is the P-value from the Shapiro-Wilk test? Do the residuals appear to be Normal?</td><td>P-value: 2.2×10^{-16}
CIRCLE ONE
NORMAL [✓ NOT NORMAL]</td></tr>
<tr><td>1.12 In question 1.11, there are many outliers in the lower tail, outliers with negative residuals.</td><td>CIRCLE ONE
TRUE [✓ FALSE]
The distribution is skewed right, long upper tail, short lower tail. Compare the Normal plot and boxplot.</td></tr>
</table>

Statistics 500, Problem 1, Fall 2016, p2. **Answers**

<table>
<tr><th>Fit model 2 from the data page, and assume it is true for the purpose of answering questions in part 2.</th><th>Fill in or CIRCLE the correct answer.</th></tr>
<tr><td>2.1 In model 1, the coefficient β_2 of fibercarb differed significantly from zero at the 0.05-level, but in model 2, the coefficient γ_2 of fibercarb does not differ significantly from zero, because women eat more fiber and gender is included in model 2.</td><td>CIRCLE ONE
TRUE [✓ FALSE]</td></tr>
<tr><td>2.2 Test the null hypothesis that $H_0:\gamma_1=\gamma_2=\gamma_3=0$ in model 2, i.e., the hypothesis that the three dietary variables have zero coefficients once we include age and gender in the model. What is the name of the test statistic? What is the numerical value of the test statistic? What are its degrees of freedom (df)? What is the two-sided P-value? Is the null hypothesis plausible?</td><td>General linear hypothesis comparing two models.
Name: F-test Value: 13.288
Df: 3 and 3990 P-value: 1.251 x 10^{-8}
CIRCLE ONE
PLAUSIBLE [✓ NOT PLAUSIBLE]</td></tr>
<tr><td>2.3 Adding predictors to a linear regression model – e.g., adding female and age to model 1 to get model 2 – never reduces R^2.</td><td>R^2 goes up when you add variables. This limits its usefulness.
[✓ TRUE] FALSE</td></tr>
<tr><td>2.4 When the residuals from model 2 are plotted against the fitted values, we see a clear U-shape, with too many positive residuals at small and large fitted values and too few positive residuals and mid-sized fitted values.</td><td>CIRCLE ONE
TRUE [✓ FALSE]</td></tr>
<tr><td>2.5 Give the sample standard deviation of sad, the estimate of σ from model 1 and the estimate of ω from model 2. Note that σ and ω, not σ^2 and ω^2, are requested.</td><td><table><tr><th>Stand.dev(sad)</th><th>Estimate of σ</th><th>Estimate of ω</th></tr><tr><td>4.449</td><td>4.44</td><td>4.291</td></tr></table></td></tr>
<tr><td>2.6 Look at the absolute values of the residuals from model 2. Which individual had the largest absolute residual? Give the NHANES seqn. What is the sad for this person? What is the fitted sad from model 2? What is the (signed) value of the residual? This is a skinny man who ate lots of fiber (true or false).</td><td>Seqn: 79612 sad: 37.5
Fitted sad: 20.68 residual: 16.8
CIRCLE ONE
TRUE [✓ FALSE]</td></tr>
<tr><td>2.7 Use avPlots from the car package to plot model 2. The plot for fibercarb has a slope that tilts upward (positive slope), while the plot for age has a slope that tilts downward (negative slope).</td><td>CIRCLE ONE
TRUE [✓ FALSE]</td></tr>
<tr><td>2.8 Look at the form of models 1 and 2. From the mathematical form of models 1 and 2, because the first variable is carbpct in both models, without fitting the models or looking at the data, you knew that the sign of the estimate of β_1 would be the same as the sign of the estimate of γ_1.</td><td>CIRCLE ONE
TRUE [✓ FALSE]
The coefficient of a variable typically changes when you add or delete variables from a model, and it can change sign.</td></tr>
</table>

Doing the Problem Set in R

```
attach(nh14food)
which.max(sad) #Question 1.1
nh14food[2685,] #Question 1.1
which.min(sad) #Question 1.2
nh14food[894,] #Question 1.2
cor(cbind(sad,carbpct,fibercarb,sugarcarb)) #Question 1.3
m1<-lm(sad~carbpct+fibercarb+sugarcarb) #Question 1.4
summary(m1) #Questions 1.4, 1.8, 1.9
confint(m1) #Question 1.6
m0<-lm(sad~1) #Question 1.8
anova(m0,m1) #Question 1.8
cor(sad,m1$fit) #Question 1.8
sqrt(0.004636) #Question 1.8
cor(sad,m1$fit)^2 #Question 1.8
qqnorm(m1$residuals) #Question 1.9
par(mfrow=c(1,2)) #Question 1.9
qqnorm(m1$residuals) #Question 1.9
qqline(m1$residuals) #Question 1.9
boxplot(m1$residuals) #Question 1.9
shapiro.test(m1$residuals) #Question 1.9
m2<-lm(sad~carbpct+fibercarb+sugarcarb+age+female) #Question 2.
m3<-lm(sad~age+female)
anova(m3,m2) #Question 2.2
plot(m2) #Question 2.4
summary(m2) #Question 2.5
which.max(abs(m2$residuals)) #Question 2.6
nh14food[2371,] #Question 2.6
m2$fitted.values[2371] #Question 2.6
m2$residuals[2371] #Question 2.6
detach(nh14food) #Question 2.6
```

PROBLEM SET #2 STATISTICS 500 FALL 2016: DATA PAGE 1

Due in class at noon in class on Tuesday, November 22, 2016.

This is an exam. Do not discuss it with anyone.

The data are the same as in Problem Set 1.
The data are from NHANES 2013-2014 (http://www.cdc.gov/nchs/nhanes/).
They describe adults, age>=20, who are not known to be pregnant.
The data are in an object in your workspace, nh14food. The data are also at data.csv on my web page (for a limited time).
The diet information describes the foods eaten on one interview day.

```
> dim(nh14food)
[1] 3996   13
```

The data for the first 3 people is as follows:

```
t(nh14food[1:3,])
person         1               2               3
seqn       73557           73558           73559
female     0               0               0
age        69              54              72
educ       High School/GED High School/GED Some College
bmi        26.7            28.6            28.9
sad        20.5            24.2            25.8
carbpct    59.59320        33.90749        51.29706
alcoholpct  0.0000         16.6625          0.0000
fatpct     29.55471        22.37556        33.93266
proteinpct 10.85209        27.05446        14.77028
calories   1574            5062            1743
fibercarb  0.04507701      0.03940724      0.04411961
sugarcarb  0.7365499       0.1061636       0.4585766
```

The variables are:
seqn is the NHANES id number.
female=1 for female, =0 for male
age in years
educ education in 5 categories
bmi = body mass index, a measure of obesity
https://www.cdc.gov/healthyweight/assessing/bmi/
sad = Sagittal abdominal diameter, another measure of obesity (cm)
https://en.wikipedia.org/wiki/Sagittal_abdominal_diameter
carbpct = percent of calories from carbs
alcoholpct = percent of calories from alcohol
fatpct = percent of calories from fat
proteinpct = percent of calories from protein
calories = calories consumed
fibercarb = (grams of fiber)/(grams of carbs)
sugarcarb = (grams of sugar)/(grams of carbs)

PROBLEM SET #2 STATISTICS 500 FALL 2016: DATA PAGE 2

This is an exam. Do not discuss it with anyone.

Due in class at noon in class on Tuesday, November 22, 2016..

In the current problem set, you will try to understand the relationship between two measures of obesity, namely sad and bmi.

Model 1: sad = $\beta_0 + \beta_1$ female + β_2 age + β_3 bmi + ε where ε is iid $N(0, \sigma^2)$

Model 2: $\text{sad}^{1/2} = \gamma_0 + \gamma_1$ female + γ_2 age + γ_3 $\log_2$(bmi) + ζ where ζ is

iid $N(0, \omega^2)$

$\log_2(x)$ is the base 2 log of x, obtained in R as log2(x). $y^{1/2}$ is the square root of y, obtained in R as sqrt(y).

Model 1 has betas and model 2 has gammas so that different things have different symbols. It does not matter which Greek letter we use.

Try plotting the data in various ways. Do not turn in the plots.

```
attach(nh14food)
boxplot(bmi~female)
boxplot(sad~female)
plot(bmi,sad,col=female+1)
plot(log2(bmi),sqrt(sad),col=female+1)
```

`The seqn is the NHANES ID number. Use it in questions that ask you to identify someone.`

Follow instructions. **Write your name**, last name first, on both sides of the answer page. Also, write your id number on the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Last name:_____________________ First Name: _________________ ID#:___________________

Statistics 500, Problem 2, Fall 2016, p1. **This problem set is an exam. Do not discuss it with anyone**.

<table>
<tr><th>Fit model 1 on the data page and use it to answer the following questions.</th><th>Fill in or CIRCLE the correct answer</th></tr>
<tr><td>1.1 In model 1, which person has the largest leverage or hatvalue? Give the SEQN for this person. What is the hatvalue for this person? For model 1, what numerical value is used to judge whether a hatvalue is large? Does this individual have “high leverage” by this standard? Use SEQN as the identifier, not row number.</td><td>SEQN: __________ hatvalue: ______________
Numerical value judged large: _______________
High leverage by this standard (Circle one)

YES NO</td></tr>
<tr><td>1.2 The individual identified in question 1.1 had the highest hatvalue because she has the smallest bmi among all 3996 people in the data set.</td><td>Circle one
TRUE FALSE</td></tr>
<tr><td>1.3 The individual you identified in question 1.1 had the highest hatvalue because her sad was incompatible with her bmi.</td><td>Circle one
TRUE FALSE</td></tr>
<tr><td>1.4 Consider the person with the largest hatvalue and the person with the smallest hatvalue. Consider the variance of their predicted values or yhats. The variance of the predicted value for the person with the largest hatvalue is more than 18 times greater than the variance of the predicted value for the person with the smallest hatvalue.</td><td>Circle one

TRUE FALSE</td></tr>
<tr><td>1.5 In model 1, test for outliers in such a way that the probability of falsely finding an outlier when there is none is at most 0.05. List all outliers by their SEQN, give their t-statistics, and their “corrected” P-values, that is, the P-value corrected for multiple testing. If there are no outliers, write NONE.</td><td>
<table>
<tr><th>SEQN</th><th>t-statistic</th><th>Corrected P</th></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
</table>
</td></tr>
<tr><td>1.6 In model 1, give the SEQN for the person with the largest absolute value of dffits and the value of dffits with its sign (+/-). This person is above the regression plane and pulls his or her predicted value up by .76 times the standard error of the predicted value for this person (True or False).</td><td>SEQN: __________ dffits: _____________
Circle one

TRUE FALSE</td></tr>
<tr><td>1.7 For the intercept and each of the three predictors in model 1, there is a dfbetas for each person. For each of these four terms in the model, give the most negative (Min) and most positive (Max) dfbetas in the table. So there are 4x3996 dfbetas. Find the one that is furthest from zero, the one dfbetas with the largest absolute value of all 4x3996 dfbetas, and identify the person who produced that value, giving that person’s SEQN.</td><td>
<table>
<tr><th>Term</th><th>Min dfbetas</th><th>Max dfbetas</th></tr>
<tr><td>intercept</td><td></td><td></td></tr>
<tr><td>female</td><td></td><td></td></tr>
<tr><td>age</td><td></td><td></td></tr>
<tr><td>bmi</td><td></td><td></td></tr>
</table>

SEQN for biggest: ___________________
</td></tr>
</table>

Last name:_____________________ First Name: _________________ ID#:___________________

Statistics 500, Problem 2, Fall 2016, p2. **This problem set is an exam. Do not discuss it with anyone**.

2. The questions in part 2 ask about the adequacy of model 1. You may have to fit additional models that add to model 1 to answer Question 2.	Fill in or CIRCLE the correct answer
2.1 Model 1 uses the same slope for men and women when relating bmi to sad. That is, the model is parallel for men and women. Test parallelism for bmi for men and women. That is, test the null hypothesis that men and women have the same bmi slope against the alternative that they have different bmi slopes. Give the name of the test statistic, the numerical value of the test statistic, the two-sided P-value, and state whether or not parallelism in bmi is plausible for men and women.	Name of test statistic:__________________ Value of test statistic: __________________ Two-sided P-value: ____________________ Circle one PLAUSIBLE NOT PLAUSIBLE
2.2 Model 1 assumes that the relationship between bmi and sad is linear. In model 1, test the null hypothesis that the relationship between bmi and sad is linear against the alternative that it is quadratic, not linear, in bmi. Give the name of the test statistic, the numerical value of the test statistic, the two-sided P-value, and state whether or not linearity is plausible.	Name of test statistic:__________________ Value of test statistic: __________________ Two-sided P-value: ____________________ Circle one PLAUSIBLE NOT PLAUSIBLE
2.3 In model 1, test the null hypothesis that the error variance is constant against the alternative that it varies with the predicted sad. Give the chi-square value and the P-value and indicate whether the null hypothesis is plausible.	Chi-square:__________________ P-value: ____________________ Circle one PLAUSIBLE NOT PLAUSIBLE

<table>
<tr><td>3. Compute the log₂(bmi) and sqrt(sad) as new variables. Fit model 2 using these new variables. Use them to answer question 3.</td><td>Fill in or CIRCLE the correct answer</td></tr>
<tr><td>3.1 Suppose one person has a bmi of 40 and another person has a bmi of 20. Then log₂(bmi) for the first person minus log₂(bmi) for the second person is 1.</td><td>Circle one
TRUE FALSE</td></tr>
<tr><td>3.2 In model 2, test for outliers in such a way that the probability of falsely finding an outlier when there is none is at most 0.05. List all outliers by their SEQN, give their t-statistics, and their “corrected” P-values, that is, the P-value corrected for multiple testing. If there are no outliers, write NONE.</td><td><table><tr><td>SEQN</td><td>t-statistic</td><td>Corrected P</td></tr><tr><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td></tr></table></td></tr>
</table>

Statistics 500, Problem 2, Fall 2016, p1. **Answers.**

<table>
<tr><th>Fit model 1 on the data page and use it to answer the following questions.</th><th>Fill in or CIRCLE the correct answer</th></tr>
<tr><td>1.1 In model 1, which person has the largest leverage or hatvalue? Give the SEQN for this person. What is the hatvalue for this person? For model 1, what numerical value is used to judge whether a hatvalue is large? Does this individual have “high leverage” by this standard? Use SEQN as the identifier, not row number.</td><td>SEQN: 82389 hatvalue: 0.009068
Numerical value judged large: 0.002002
High leverage by this standard (Circle one)
[✓ YES] NO</td></tr>
<tr><td>1.2 The individual identified in question 1.1 had the highest hatvalue because she has the smallest bmi among all 3996 people in the data set.</td><td>Circle one
TRUE [✓ FALSE]
She has the largest bmi.</td></tr>
<tr><td>1.3 The individual you identified in question 1.1 had the highest hatvalue because her sad was incompatible with her bmi.</td><td>Circle one
TRUE [✓ FALSE]
hatvalues don’t depend on y.</td></tr>
<tr><td>1.4 Consider the person with the largest hatvalue and the person with the smallest hatvalue. Consider the variance of their predicted values or yhats. The variance of the predicted value for the person with the largest hatvalue is more than 18 times greater than the variance of the predicted value for the person with the smallest hatvalue.</td><td>Circle one
[✓ TRUE] FALSE</td></tr>
<tr><td>1.5 In model 1, test for outliers in such a way that the probability of falsely finding an outlier when there is none is at most 0.05. List all outliers by their SEQN, give their t-statistics, and their “corrected” P-values, that is, the P-value corrected for multiple testing. If there are no outliers, write NONE.</td><td><table>
<tr><th>SEQN</th><th>t-statistic</th><th>Corrected P</th></tr>
<tr><td>82389</td><td>-5.871</td><td>0.00001871</td></tr>
<tr><td>82731</td><td>-5.420</td><td>0.00025223</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td>1.6 In model 1, give the SEQN for the person with the largest absolute value of dffits and the value of dffits with its sign (+/-). This person is above the regression plane and pulls his or her predicted value up by .76 times the standard error of the predicted value for this person (True or False).</td><td>SEQN: 82389 dffits: -0.5616
Circle one
TRUE [✓ FALSE]</td></tr>
<tr><td>1.7 For the intercept and each of the three predictors in model 1, there is a dfbetas for each person. For each of these four terms in the model, give the most negative (Min) and most positive (Max) dfbetas in the table. So there are 4x3996 dfbetas. Find the one that is furthest from zero, the one dfbetas with the largest absolute value of all 4x3996 dfbetas, and identify the person who produced that value, giving that person’s SEQN.</td><td><table>
<tr><th>Term</th><th>Min dfbetas</th><th>Max dfbetas</th></tr>
<tr><td>intercept</td><td>-0.0791516</td><td>0.3878053</td></tr>
<tr><td>female</td><td>-7.149e-02</td><td>6.073e-02</td></tr>
<tr><td>age</td><td>-7.866e-02</td><td>1.083e-01</td></tr>
<tr><td>bmi</td><td>-0.5403823</td><td>0.1143552</td></tr>
</table>
SEQN for biggest: 82389</td></tr>
</table>

Statistics 500, Problem 2, Fall 2016, p1. **Answers.**

2. The questions in part 2 ask about the adequacy of model 1. You may have to fit additional models that add to model 1 to answer Question 2.	Fill in or CIRCLE the correct answer
2.1 Model 1 uses the same slope for men and women when relating bmi to sad. That is, the model is parallel for men and women. Test parallelism for bmi for men and women. That is, test the null hypothesis that men and women have the same bmi slope against the alternative that they have different bmi slopes. Give the name of the test statistic, the numerical value of the test statistic, the two-sided P-value, and state whether or not parallelism in bmi is plausible for men and women. (10 points)	Name of test statistic: t-statistic Value of test statistic: -10.538 Two-sided P-value: $<10^{-16}$ Circle one PLAUSIBLE [✓ NOT PLAUSIBLE]
2.2 Model 1 assumes that the relationship between bmi and sad is linear. In model 1, test the null hypothesis that the relationship between bmi and sad is linear against the alternative that it is quadratic, not linear, in bmi. Give the name of the test statistic, the numerical value of the test statistic, the two-sided P-value, and state whether or not linearity is plausible. (10 points)	Name of test statistic: t-statistic Value of test statistic: -13.27 Two-sided P-value: $<10^{-16}$ Circle one PLAUSIBLE [✓ NOT PLAUSIBLE]
2.3 In model 1, test the null hypothesis that the error variance is constant against the alternative that it varies with the predicted sad. Give the chi-square value and the P-value and indicate whether the null hypothesis is plausible.	Chi-square: 286.2 P-value: 3.333675×10^{-64} Circle one PLAUSIBLE [✓ NOT PLAUSIBLE]

<table>
<tr><td>3. Compute the $\log_2$(bmi) and sqrt(sad) as new variables. Fit model 2 using these new variables. Use them to answer question 3.</td><td>Fill in or CIRCLE the correct answer</td></tr>
<tr><td>3.1 Suppose one person has a bmi of 40 and another person has a bmi of 20. Then $\log_2$(bmi) for the first person minus $\log_2$(bmi) for the second person is 1.</td><td>Circle one
[✓ TRUE] FALSE</td></tr>
<tr><td>3.2 In model 2, test for outliers in such a way that the probability of falsely finding an outlier when there is none is at most 0.05. List all outliers by their SEQN, give their t-statistics, and their “corrected” P-values, that is, the P-value corrected for multiple testing. If there are no outliers, write NONE.</td><td><table>
<tr><td>SEQN</td><td>t-statistic</td><td>Corrected P</td></tr>
<tr><td>NONE</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
</table></td></tr>
</table>

Doing the Problem Set in R

```
attach(nh14food)
l2bmi<-log2(bmi)
rsad<-sqrt(sad)
m1<-lm(sad~female+age+bmi) #Model 1
h<-hatvalues(m1)
dim(nh14food)
4/3996
mean(h)
2*mean(h) #Question 1.1
which.max(h) #Question 1.1
h[3482] #Question 1.1
nh14food[3482,] #Question 1.1 AND 1.2
max(h)/min(h) #Question 1.4
library(car)
outlierTest(m1) #Question 1.5
nh14food[c(3482,3615),] #Question 1.5
which.max(abs(dffits(m1))) #Question 1.6
dffits(m1)[3482] #Question 1.6
summary(dfbetas(m1)) #Question 1.7
which.max(abs(dfbetas(m1)[,4])) #Question 1.7
fbmi<-female*bmi #Question 2.1
summary(lm(sad ~ female + age + bmi + fbmi)) #Question 2.1
summary(lm(sad ~ female + age + poly(bmi,2))) #Question 2.2
ncvTest(m1) #Question 2.3
log2(40)-log2(20) #Question 3.1
outlierTest(lm(rsad~age+female+l2bmi)) #Question 3.2
```

PROBLEM SET #3 STATISTICS 500 FALL 2016: DATA PAGE 1

Due at noon on Tuesday, 20 December 2016.

This is an exam. Do not discuss it with anyone.

The data are from the paper U.Werfel1, V.Langen, I.Eickhoff, J.Schoonbrood, C.Vahrenholz, A.Brauksiepe, W.Popp and K.Norpoth (1998) Elevated DNA single-strand breakage frequencies in lymphocytes of welders exposed to chromium and nickel. Carcinogenesis 19, 413–418. It is available from the library web-page if you'd like to look at it, but there is no need to look at it. Only a subset of the data is included so that your analysis can focus on a balanced design, and hence be simpler.

The data are in the course workspace as werfelwelder or, for a short time, on my webpage as data.csv. You may need to download the workspace again, perhaps after clearing your browser's memory.

```
> head(werfelwelder)
  ID welder smoke erpcp age
1 1      C    NS 0.751  23
2 2      C    NS 0.875  23
3 3      C    NS 0.161  29
4 4      C    NS 0.630  30
5 5      C    NS 1.462  32
6 6      C    NS 0.702  35
> dim(werfelwelder)
[1] 48  5
```

Electric arc welders are exposed to chromium and nickel, which may damage DNA. Werfell et al. looked for evidence of this by comparing blood samples from welders and controls. The outcome is erpcp. It is the relative DNA elution rate through polycarbonate filters with proteinase K, and it measures DNA strand breaks, with higher values signifying greater damage.

There are 48 people, 24 electric welders and 24 controls not engaged in welding. All are men. Welders are indicated by W, controls by C. Half of each group were smokers, S, and half were nonsmokers, NS. The age of each person is given.

Create a new factor from two existing factors as follows:

```
attach(werfelwelder)
grp<-welder:smoke
is.factor(grp)
boxplot(erpcp~grp)
```

PROBLEM SET #3 STATISTICS 500 FALL 2016: DATA PAGE 2

Due at noon on Tuesday, 20 December 2016.

This is an exam. Do not discuss it with anyone.

Let y_{ij} be the erpcp value for the ith person in group j, where i=1,…,12 and j = C:NS, C:S, W:NS, W:S.

IMPORTANT: Refer to the four groups as C:NS, C:S, W:NS, W:S.

Model 1 asserts: $y_{ij} = \mu + \tau_j + \varepsilon_{ij}$ where ε_{ij} are iid $N(0,\sigma^2)$

Model 2 asserts:

$y_{ij} = \beta_0 + \beta_1 grpWvsC + \beta_2 grpSvsNS + \beta_3 grpInter + \beta_4 age + \zeta_{ij}$ where ζ_{ij} are iid $N(0,\omega^2)$

where you build grpWvsC, grpSvsNS, and grpInter in question 3 using the orthogonal contrasts in question 2.

Model 2 has $2^4 = 16$ submodels, including model 2 itself and the empty model with just the constant term.

Question 3 asks you to create 3 variables from the 3 contrasts. Each variable will have a value for each person, 48 values in total. Each variable can be used as a predictor in a regression. If in doubt about how to do this, look at the bulkpack.

Follow instructions. **Write your name and id#**, last name first, on both sides of the answer page. If a question has several parts, **answer every part**. Some questions are not in boxes; however, answer them anyway. Turn in **only the answer page**. **Do not turn in additional pages**. Do not turn in graphs. **Brief answers suffice**. If a question asks you to circle an answer, then you are correct if you **circle the correct answer**. If you cross out an answer, no matter which answer you cross out, the answer is wrong. Do not circle TRUE adding a note explaining why it is also FALSE. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam. The exam is due **on the exam date at noon, at my office, 473 JMHH**. You may turn in the exam early by placing it in an envelope addressed to me and leaving it in my mail box in statistics, 4th floor, JMHH. If you prefer, give it to Noel at the front desk in statistics. Do not slip the exam under the door of the Statistics Department when the Department is closed – the exam will get lost or thrown out by a cleaner. Make and keep a photocopy of your answer page. The answer key will be posted in the revised bulk pack on-line. You can compare the answer key to your photocopy.

Name (**Last**, First): _________________________ ID# _________________________

PROBLEM SET #3 STATISTICS 500 FALL 2016: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon, 20 December 2016.

Use model 1 in answering the questions in part 1.	Fill in or CIRCLE the correct answer
1.1 If model 1 were true, then the 48 y_{ij} are iid.	Circle one. TRUE FALSE
1.2 For model 1, do a one-way analysis of variance for its four groups. Give the sum of squares (SS), degrees of freedom (DF), mean squares (MS) and F-ratio.	<table><tr><td>Source</td><td>SS</td><td>DF</td><td>MS</td><td>F</td></tr><tr><td>Between Groups</td><td></td><td></td><td></td><td></td></tr><tr><td>Within Groups</td><td></td><td></td><td></td><td>XX XX</td></tr></table>
1.3 Test the null hypothesis that H_0: $\tau_{C:NS} = \tau_{C:S} = \tau_{W:NS} = \tau_{W:S} = 0$. Give the name, value degrees of freedom (DF)of the test statistic, the P-value. Is H_0 plausible?	Name: Value: DF: P-value: Circle one. PLAUSIBLE NOT PLAUSIBLE
1.4 The P-value in question 1.3 is one-sided, testing for higher erpcp among welders.	Circle one. TRUE FALSE
1.5 In model 1, give the **numerical value** of largest and smallest hatvalues or leverages, h_i.	Largest value: Smallest value:
1.6 In model 1, test at the 0.05 level the null hypothesis that there are no outliers. **Do you find outliers?** (YES or NO) If yes, give the t-values and id numbers of the outlier or outliers.	Outliers? Circle one. YES NO If yes, list the outlier(s) <table><tr><td>ID #</td><td>t-value</td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr></table>

2. **Create 3 orthogonal contrasts among the four groups.** Use **integer** weights.

Group	$\tau_{C:NS}$	$\tau_{C:S}$	$\tau_{W:NS}$	$\tau_{W:S}$
Welder versus Control				
Smoker versus nosmoker				
Interaction				

Name (**Last**, First): _________________________ ID# __________________________

PROBLEM SET #3 STATISTICS 500 FALL 2016: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon, 20 December 2016.

3. Create 3 variables, called grpWvsC, grpSvsNS, and grpInter, representing the 3 contrasts in question 2. Use them to fill in the following ANOVA table. Here, SS=sum of squares, DF=degrees of freedom, MS=mean square, F=F-statistic, P-value.

3.1 Source	SS	DF	MS	F	P-value
Between 4 groups					
Welder vs Control					
Smoker vs Nonsmoker					
Interaction					
Within groups				XXXXXXXX XXXXXXXX	XXXXXXXX XXXXXXXX

3.2 Test the null hypothesis that welder versus control difference is the same for smokers and nonsmokers,

H_0: $\tau_{W:S}-\tau_{C:S} = \tau_{W:NS}-\tau_{C:NS}$. What is the P-value? Is H_0 plausible?

P-value: **CIRCLE ONE**: PLAUSIBLE NOT PLAUSIBLE

4. Compare the 4 groups in all pairs under model 1, testing 6 hypotheses of equality between pairs of groups. Give the 2-sided P-value with no adjustment, adjustment by the Holm method and the Bonferroni method and.

	No adjustment	Holm	Bonferroni
H_1: $\tau_{W:NS}-\tau_{C:S}=0$			
H_2: $\tau_{W:S}-\tau_{C:NS}=0$			

5. Use model 2 as the maximum model for these question.	Fill in or CIRCLE the correct answer.
5.1 Which submodel of model 2 has the smallest C_P? **List its variables by name**. What is the **value of C_P**? Is this C_P less than or equal to the size of this model?	Submodel: Value of C_P: **Circle one**: ≤Size >Size
5.2 Give the value of PRESS for model 2 and for the model you selected in question 5.1.	PRESS for model 2: PRESS for model 5.1:

ANSWERS: PROBLEM SET #3 STATISTICS 500 FALL 2016: ANSWER PAGE 1

<table>
<tr><td>**Use model 1** in answering the questions in part 1.</td><td>Fill in or CIRCLE the correct answer</td></tr>
<tr><td>1.1 If model 1 were true, then the 48 y_{ij} are iid.</td><td>Circle one.
TRUE [✓ FALSE]</td></tr>
<tr><td>1.2 For model 1, do a one-way analysis of variance for its four groups. Give the sum of squares (SS), degrees of freedom (DF), mean squares (MS) and F-ratio.</td><td>
<table>
<tr><td>Source</td><td>SS</td><td>DF</td><td>MS</td><td>F</td></tr>
<tr><td>Between Groups</td><td>4.509</td><td>3</td><td>1.50</td><td>6.6</td></tr>
<tr><td>Within Groups</td><td>10.025</td><td>44</td><td>0.23</td><td>XX
XX</td></tr>
</table>
</td></tr>
<tr><td>1.3 Test the null hypothesis that H_0: $\tau_{C:NS} = \tau_{C:S} = \tau_{W:NS} = \tau_{W:S} = 0$. Give the name, value degrees of freedom (DF)of the test statistic, the P-value. Is H_0 plausible?</td><td>Name: F-statistic Value: 6.6
DF: 3 and 44 P-value: .000887
Circle one.
PLAUSIBLE [✓ NOT PLAUSIBLE]</td></tr>
<tr><td>1.4 The P-value in question 1.3 is one-sided, testing for higher erpcp among welders.</td><td>Circle one.
TRUE [✓ FALSE]</td></tr>
<tr><td>1.5 In model 1, give the **numerical value** of largest and smallest hatvalues or leverages, h_i.</td><td>Largest value: 0.0833
Smallest value: 0.0833</td></tr>
<tr><td>1.6 In model 1, test at the 0.05 level the null hypothesis that there are no outliers. **Do you find outliers?** (YES or NO) If yes, give the t-values and id numbers of the outlier or outliers.</td><td>Outliers? Circle one.
[✓ YES] NO
If yes, list the outlier(s)
<table>
<tr><td>ID #</td><td>t-value</td></tr>
<tr><td>28</td><td>3.97</td></tr>
<tr><td>NA</td><td>NA</td></tr>
</table>
</td></tr>
</table>

2. **Create 3 orthogonal contrasts among the four groups.** Use **integer** weights.

Group	$\tau_{C:NS}$	$\tau_{C:S}$	$\tau_{W:NS}$	$\tau_{W:S}$
Welder versus Control	-1	-1	1	1
Smoker versus nosmoker	-1	1	-1	1
Interaction	1	-1	-1	1

3. Create 3 variables, called grpWvsC, grpSvsNS, and grpInter, representing the 3 contrasts in question 2. Use them to fill in the following ANOVA table. Here, SS=sum of squares, DF=degrees of freedom, MS=mean square, F=F-statistic, P-value.

3.1 Source	SS	DF	MS	F	P-value
Between 4 groups	4.509	3	1.5029	6.597	0.0008
Welder vs Control	3.7397	1	3.7397	16.41	0.0002
Smoker vs Nonsmoker	0.1615	1	0.1615	0.7	0.4044
Interaction	0.6075	1	0.6075	2.67	0.1096
Within groups	10.0245	44	0.2278	XXXXXXXX	XXXXXXX

3.2 Test the null hypothesis that welder versus control difference is the same for smokers and nonsmokers,

H_0: $\tau_{W:S}-\tau_{C:S} = \tau_{W:NS}-\tau_{C:NS}$. What is the P-value? Is H_0 plausible?

P-value: 0.1096 **CIRCLE ONE:** [✓ **PLAUSIBLE**] NOT PLAUSIBLE

4. Compare the 4 groups in all pairs under model 1, testing 6 hypotheses of equality between pairs of groups. Give the 2-sided P-value with no adjustment, adjustment by the Holm method and theBonferroni method.

	No adjustment	Holm	Bonferroni
H_1: $\tau_{W:NS}-\tau_{C:S}=0$	0.00121	0.0061	0.0073
H_2: $\tau_{W:S}-\tau_{C:NS}=0$	0.02819	0.1128	0.1691

5. Use model 2 as the maximum model for these question.	Fill in or CIRCLE the correct answer.
5.1 Which submodel of model 2 has the smallest C_P? **List its variables by name**. What is the **value of C_P**? Is this C_P less than or equal to the size of this model?	Submodel: grpWvsNS and grpInter Value of C_P: 2.616 **Circle one:** [✓ **≤Size**] >Size
5.2 Give the value of PRESS for model 2 and for the model you selected in question 5.1.	PRESS for model 2: 12.19 PRESS for model 5.1: 11.59

DOING THE PROBLEM SET IN R
(PROBLEM 3, FALL 2016)

```
attach(werfelwelder)
grp<-welder:smoke
is.factor(grp)
boxplot(erpcp~grp)
summary(aov(erpcp~grp)) # Question 1.2 and 1.3
hatvalues(aov(erpcp~grp)) # Question 1.5
which.max(abs(rstudent(aov(erpcp~grp)))) # Question 1.6
rstudent(aov(erpcp~grp))[28] # Question 1.6
qt(0.025/48,43) # Question 1.6
sum(abs(rstudent(aov(erpcp~grp)))>=3.517654) # Question 1.6
contrasts(grp) # Question 2
WvsC<-c(-1,-1,1,1) # Question 2
SvsNS<-c(-1,1,-1,1) # Question 2
Inter<-WvsC*SvsNS # Question 2
contrasts(grp)<-cbind(WvsC,SvsNS,Inter) # Question 2
contrasts(grp) # Question 2
anova(lm(erpcp~grpWvsC +grpSvsNS +grpInter)) #Question 3
pairwise.t.test(erpcp,grp,p.adjust.method="none") #Question 4
pairwise.t.test(erpcp,grp,p.adjust.method="holm") #Question 4
pairwise.t.test(erpcp,grp,p.adjust.method="bonf") #Question 4
X<-cbind(grpWvsC,grpSvsNS,grpInter,age) #Question 5.1
library(leaps) #Question 5.1
m<-leaps(x=X,y=erpcp,names=colnames(X)) #Question 5.1
cbind(m$which,m$size,m$Cp) #Question 5.1
library(DAAG) #Question 5.2
press(lm(erpcp~grpWvsC +grpSvsNS +grpInter+age)) #Question 5.2
press(lm(erpcp~grpWvsC +grpInter)) #Question 5.2
```

PROBLEM SET #1 STATISTICS 500 FALL 2015: DATA PAGE 1

Due in class at noon in class on Tuesday 13 October 2015

This is an exam. Do not discuss it with anyone.

The data are selected for illustration from a study by Tager et al (1979), *American Journal of Epidemiology*, 110, 15-26, but have been substantially simplified. The data concern fev = forced expiratory volume of children measured in liters, which is an index of pulmonary function and is the volume of air expelled after 1 second of constant effort. Other variables are age of the child in years, height of the child in inches, female = 1 for female, 0 for male, and smoker = 1 if current smoker, 0 if not a current smoker. Each row of data is a different child.

The data are in an object called `rfev2` in the course workspace at http://www-stat.wharton.upenn.edu/~rosenbap/ If you are not using R, then the link data.csv on the same page will give you the data as a csv-file that many programs can read, including excel.

```
> head(rfev2)
    id   fev age height female smoker
1  301 1.708   9   57.0      1      0
2  451 1.724   8   67.5      1      0
3  501 1.720   7   54.5      1      0
4  642 1.558   9   53.0      0      0
5  901 1.895   9   57.0      0      0
6 1701 2.336   8   61.0      1      0

> dim(rfev2)
[1] 654   6
```

Model 1: fev = $\beta_0 + \beta_1$ age + β_2 height + ε where ε is iid $N(0, \sigma^2)$

Model 2: fev = $\gamma_0 + \gamma_1$ age + γ_2 height + γ_3 female + ζ where ζ is iid $N(0, \omega^2)$

Model 1 has betas and model 2 has gammas so that different things have different symbols. It does not matter which Greek letter we use.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Last name:______________________ First Name: _________________ ID#:____________________

Statistics 500, Problem 1, Fall 2015, p1. **This problem set is an exam. Do not discuss it with anyone**.

Fit model 1 from the data page and use it to answer the questions in part 1.	Fill in or circle the correct answer
1.1 In model 1, what is the numerical value of the estimate of σ?	Estimate of σ:
1.2 Under model 1, give the 95% two-sided confidence interval for the coefficient of age, β_1.	95% CI: [,]
1.3 In model 1, test the null hypothesis that the coefficient of height, β_2, is zero, H_0: $\beta_2=0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.	Name:_________ Value:__________ DF:___________ P-value:_____________ Circle one: H_0 is PLAUSIBLE NOT PLAUSIBLE
1.4 In model 1, test the null hypothesis that both slopes are zero, H_0: $\beta_1=\beta_2=0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.	Name:_________ Value:__________ DF:___________ P-value:_____________ Circle one: H_0 is PLAUSIBLE NOT PLAUSIBLE
1.5 In model 1, the correlation between fev and fitted fev (i.e., yhat) is 0.7664.	Circle one: TRUE FALSE

Use the residuals from model 1 on the data page to answer questions in part 2.	Fill in or circle the correct answer
2.1 Do a normal plot, a boxplot, and a Shapiro-Wilk test using the residuals from model 1. Circle the **letters** of true statements, perhaps more than one, perhaps none. E.g., you might circle a. if you thought a was true and the rest false.	Circle the LETTER for each true statement a. The residuals look Normal. b. The residuals have a short right tail and a long left tail compared to the Normal. c. The residuals have a long right tail and a long left tail compared to the Normal. d. The Shapiro-Wilk tests accepts Normality as a plausible distribution for the residuals.
2.2 Plot residuals as y against fitted values as x. Add the lowess smooth. Plot residuals as y against height as x. Add the lowess smooth. Use round() to round the fitted values to integers, forming 4 groups, 1, 2, 3, and 4; then, boxplot the residuals in these four groups. Circle the **letters** of true statements, perhaps more than one, perhaps none.	Circle the LETTER for each true statement a. The lowess smooth of the plot against **height** clearly exhibits an *inverted* (i.e., upside down) U-shape, indicating nonlinearity. b. The plots and boxplots clearly indicate that the assumption of constant variance is correct.
2.3 The child with the largest absolute residual is 15 years old, 69 inches tall, with the largest fev in the data set.	Circle one: TRUE FALSE

Last name:_____________________ First Name: _________________ ID#:___________________

Statistics 500, Problem 1, Fall 2015, p2. **This problem set is an exam. Do not discuss it with anyone**.

<table>
<tr><td>Fit model 2 and use it for questions in part 3.</td><td>Fill in or circle the correct answer</td></tr>
<tr><td>3.1 In model 2, female children are estimated to have a higher fev than male children of the same height and age.</td><td>Circle one
TRUE FALSE</td></tr>
<tr><td>3.2 In model 2, test the null hypothesis that the coefficient of age and the coefficient of female are both zero, $H_0: \gamma_1 = \gamma_3 = 0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.</td><td>Name:_________ Value:__________
DF:___________ P-value:_____________
Circle one:
H_0 is PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>3.3 For the hypothesis tested in question 3.2, fill in the following ANOVA = analysis of variance table. In the table, SS means sum-of-squares, DF means degrees-of-freedom, and MS means mean-square.</td><td><table>
<tr><td></td><td>SS</td><td>DF</td><td>MS</td></tr>
<tr><td>Full Model</td><td></td><td></td><td></td></tr>
<tr><td>Reduced Model</td><td></td><td></td><td></td></tr>
<tr><td>Added variables</td><td></td><td></td><td></td></tr>
<tr><td>Residual</td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td>3.4 Using model 2, give a 2-sided 95% confidence interval for location of the regression surface (ie, the fitted fev value) for a female, 5 foot 3 inches tall, age 15.</td><td>95% Interval [,]</td></tr>
<tr><td>3.5 Using model 2, give a 2-sided 95% interval for the fev for a new child who is a female, 5 foot 3 inches tall, age 15. This new child would be in addition to the 654 children in the data set, but is imagined to follow the same model 2.</td><td>95% Interval [,]</td></tr>
<tr><td>3.6 If the number of children in the data set, currently 654, where to increase and increase but always follow the same model 2, then the interval in question 3.4 would shrink to a point, but the interval in question 3.5 would not.</td><td>Circle one
TRUE FALSE</td></tr>
</table>

Part 4 asks hypothetical questions.	Circle the correct answer
4.1 In some data set, it could happen that a test of $H_0: \gamma_1 = \gamma_3 = 0$, as in part 3.2, could reject at the 0.05 level, but two separate test of $H_0: \gamma_1 = 0$ and $H_0: \gamma_3 = 0$ in the same model could both accept at the two-sided 0.05 level.	Circle one TRUE FALSE
4.2 The actual study from which these data were selected measured each child at several ages to study growth in fev. Imagine having 654 children, each measured at 3 ages, and fitting model 2 to these 654x3 = 1962 measurements. A problem with model 2 is that it is not very plausible that 3 fev's from one child at 3 ages will be independent.	Circle one TRUE FALSE

Statistics 500, Fall 2105, Problem 1, Answers (6 points each except as noted)

Fit model 1 from the data page and use it to answer the questions in part 1.	Fill in or circle the correct answer
1.1 In model 1, what is the numerical value of the estimate of σ?	Estimate of σ: 0.4197
1.2 Under model 1, give the 95% two-sided confidence interval for the coefficient of age, β_1.	95% CI: [0.0364, 0.0722]
1.3 In model 1, test the null hypothesis that the coefficient of height, β_2, is zero, H_0: $\beta_2=0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.	t-DF are DF for σ^2, here 651. Name: t-statistic Value: 23.263 DF: 651 P-value: $< 2 \times 10^{-16}$ Circle one: H_0 is PLAUSIBLE [✓ NOT PLAUSIBLE]
1.4 In model 1, test the null hypothesis that both slopes are zero, H_0: $\beta_1=\beta_2=0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible.	Name: F-statistic Value: 1068 DF: 2 and 651 P-value: $< 2.2 \times 10^{-16}$ Circle one: H_0 is PLAUSIBLE [✓ NOT PLAUSIBLE]
1.5 In model 1, the correlation between fev and fitted fev (i.e., yhat) is 0.7664.	Circle one: TRUE [✓ FALSE]

Use the residuals from model 1 on the data page to answer questions in part 2.	Fill in or circle the correct answer
2.1 Do a normal plot, a boxplot, and a Shapiro-Wilk test using the residuals from model 1. Circle the **letters** of true statements, perhaps more than one, perhaps none. E.g., you might circle a. if you thought a was true and the rest false.	Circle the LETTER for each true statement e. The residuals look Normal. f. The residuals have a short right tail and a long left tail compared to the Normal. [✓ g.] The residuals have a long right tail and a long left tail compared to the Normal. h. The Shapiro-Wilk tests accepts Normality as a plausible distribution for the residuals.
2.2 Plot residuals as y against fitted values as x. Add the lowess smooth. Plot residuals as y against height as x. Add the lowess smooth. Use round() to round the fitted values to integers, forming 4 groups, 1, 2, 3, and 4; then, boxplot the residuals in these four groups. Circle the **letters** of true statements, perhaps more than one, perhaps none.	Circle the LETTER for each true statement c. The lowess smooth of the plot against **height** clearly exhibits an inverted (i.e., upside down) U-shape, indicating nonlinearity. d. The plots and boxplots clearly indicate that the assumption of constant variance is correct.
2.3 The child with the largest absolute residual is 15 years old, 69 inches tall, with the largest fev in the data set.	Circle one: [✓ TRUE] FALSE

<table>
<tr><th>Fit model 2 and use it in part 3.</th><th>Fill in or circle the correct answer</th></tr>
<tr><td>3.1 In model 2, female children are estimated to have a higher fev than male children of the same height and age.</td><td>Circle one
TRUE [✓ FALSE]</td></tr>
<tr><td>3.2 In model 2, test the null hypothesis that the coefficient of age and the coefficient of female are both zero, $H_0: \gamma_1 = \gamma_3 = 0$. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), its two-sided P-value, and state whether the null hypothesis is plausible. (8 points)</td><td>An F has (numerator,denominator) DF-2 numbers.
Name: F-statistic Value: 30.213
DF: 2 and 650 P-value: 2.84 x 10^{-13}
Circle one:
H_0 is PLAUSIBLE NOT PLAUSIBLE</td></tr>
<tr><td>3.3 For the hypothesis tested in question 3.2, fill in the following ANOVA = analysis of variance table. In the table, SS means sum-of-squares, DF means degrees-of-freedom, and MS means mean-square. (8 points)</td><td>
<table>
<tr><th></th><th>SS</th><th>DF</th><th>MS</th></tr>
<tr><td>Full Model</td><td>380.27</td><td>3</td><td>126.76</td></tr>
<tr><td>Reduced Model</td><td>369.99</td><td>1</td><td>369.99</td></tr>
<tr><td>Added variables</td><td>10.28</td><td>2</td><td>5.14</td></tr>
<tr><td>Residual</td><td>110.65</td><td>650</td><td>0.170</td></tr>
</table>
</td></tr>
<tr><td>3.4 Using model 2, give a 2-sided 95% confidence interval for location of the regression surface (ie, the fitted fev value) for a female, 5 foot 3 inches tall, age 15.</td><td>95% Interval [2.97, 3.14]</td></tr>
<tr><td>3.5 Using model 2, give a 2-sided 95% interval for the fev for a new child who is a female, 5 foot 3 inches tall, age 15. This new child would be in addition to the 654 children in the data set, but is imagined to follow the same model 2.</td><td>95% Interval [2.24, 3.87]</td></tr>
<tr><td>3.6 If the number of children in the data set, currently 654, where to increase and increase but always follow the same model 2, then the interval in question 3.4 would shrink to a point, but the interval in question 3.5 would not.</td><td>With more and more data, we learn where line is, but not where a new child is.
Circle one
[✓ TRUE] FALSE</td></tr>
</table>

Part 4 asks hypothetical questions.	Circle the correct answer
4.1 In some data set, it could happen that a test of $H_0: \gamma_1 = \gamma_3 = 0$, as in part 3.2, could reject at the 0.05 level, but two separate test of $H_0: \gamma_1 = 0$ and $H_0: \gamma_3 = 0$ in the same model could both accept at the two-sided 0.05 level.	This might happen if x_1 and x_3 are highly correlated. [✓ TRUE] FALSE
4.2 The actual study from which these data were selected measured each child at several ages to study growth in fev. Imagine having 654 children, each measured at 3 ages, and fitting model 2 to these 654x3 = 1962 measurements. A problem with model 2 is that it is not very plausible that 3 fev's from one child at 3 ages will be independent.	Circle one [✓ TRUE] FALSE Multiple measures on each child are called "repeated measures" or "panel data" or "longitudinal data," and typically cannot be viewed as independent observations, violating an assumption of regression.

   Statistics 500, Problem 1, Fall 2015: Doing the Problem Set in R.

1.

```
> md1<-lm(fev~age+height)
> summary(md1)
lm(formula = fev ~ age + height)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.610466   0.224271 -20.558  < 2e-16 ***
age          0.054281   0.009106   5.961 4.11e-09 ***
height       0.109712   0.004716  23.263  < 2e-16 ***
Residual standard error: 0.4197 on 651 degrees of freedom
Multiple R-squared: 0.7664,     Adjusted R-squared: 0.7657
F-statistic:  1068 on 2 and 651 DF,  p-value: < 2.2e-16
> confint(md1)
                  2.5 %      97.5 %
(Intercept) -5.05084726 -4.17008507
age          0.03639976  0.07216159
height       0.10045104  0.11897263
> cor(md1$fitted,fev)
[1] 0.8754474
> cor(md1$fitted,fev)^2
[1] 0.7664081   This is R^2, not R.
2.
> boxplot(md1$resid)
> qqnorm(md1$resid)
> qqline(md1$resid)
> shapiro.test(md1$resid)
        Shapiro-Wilk normality test
data:  md1$resid
W = 0.9865, p-value = 9.794e-06
> plot(md1$fit,md1$resid)
> lines(lowess(md1$fit,md1$resid),col="red")
> plot(height,md1$resid)
> lines(lowess(height,md1$resid),col="red")
> boxplot(md1$resid~round(md1$fit))
> which.max(abs(md1$resid))
624
> rfev2[624,]
       id   fev age height female smoker
624 25941 5.793  15     69      0      0
> summary(fev)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.791   1.981   2.548   2.637   3.118   5.793
```

3.1

```
> md2<-lm(fev~age+height+female)
> summary(md2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.287448   0.230253 -18.621  < 2e-16 ***
age          0.061364   0.009069   6.766 2.96e-11 ***
height       0.104560   0.004756  21.986  < 2e-16 ***
female      -0.161112   0.033125  -4.864 1.45e-06 ***

Residual standard error: 0.4126 on 650 degrees of freedom
Multiple R-squared: 0.7746,     Adjusted R-squared: 0.7736
F-statistic: 744.6 on 3 and 650 DF,  p-value: < 2.2e-16
```

3.2-3.3

```
> mdh<-lm(fev~height)
> anova(mdh,md2)
Analysis of Variance Table
Model 1: fev ~ height
Model 2: fev ~ age + height + female
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    652 120.93
2    650 110.65  2    10.286 30.213 2.84e-13 ***
```

3.4

```
> predict(md2,data.frame(age=15,height=63,female=1),interval="confidence")
       fit      lwr      upr
1 3.059155 2.973354 3.144956
```

3.5

```
> predict(md2,data.frame(age=15,height=63,female=1),interval="prediction")
       fit      lwr      upr
1 3.059155 2.244461 3.873849
```

PROBLEM SET #2 STATISTICS 500 FALL 2015: DATA PAGE 1

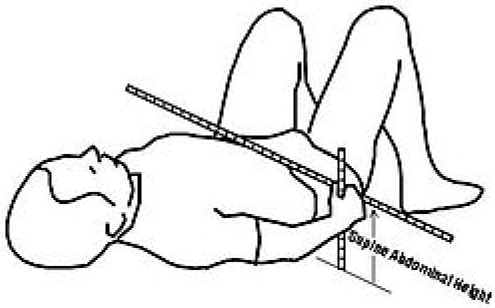
Due in class at noon in class on Tuesday 24 November 2015

This is an exam. Do not discuss it with anyone.

The data are from the 2013-2014 NHANES. The data are in an object nhanes1314st500 in the course workspace. It is also available from my web page as data.csv

```
> dim(nhanes14st500)
[1] 4576   13
> head(nhanes14st500)
    seqn lbxgh female age systolic diastolic pulse weight height  bmi waist  sad          y
1  73557  13.9      0  69      122        72    86   78.3  171.3 26.7 100.0 20.5  1.2344306
2  73558   9.1      0  54      156        62    74   89.5  176.8 28.6 107.6 24.2  0.7429755
```

The variables are as follows. (i) seqn is the NHANES id number, (ii) lbxgh is a measure of glycohemoglobin, recorded as a percent, where high values, above 6.5, are indicative of diabetes, (iii) female = 1 for female, 0 for male, (iv) age is in years, (v) systolic and diastolic record blood pressure, (vi) pulse is pulse, (vii) weight is recorded in kilograms, (viii) height, waist circumference and sad are recorded in centimeters. sad is sagittal abdominal diameter, the height indicated below – it is thought to be an improvement on “bmi” = body mass index as a measure of health risk.



The final variable, y, is just a transformation of lbxgh obtained as follows. For your work, you can use y as instructed without considering yjPower. The yjPower function is extends the Box-Cox transformation to permit both positive and negative values and is described in Yeo and Johnson (2000); however, there is no need to consult this paper unless you want to.

```
> library(car)
> help("yjPower")
> y<-yjPower(lbxgh-6.5,0,jacobian.adjusted = TRUE)
```

I suggest doing the following plots to understand how y relates to lbxgh. Notice that y>0 if and only if lbxgh>6.5, so both are values consistent with diabetes.

```
> par(mfrow=c(1,2))
> boxplot(lbxgh)
> boxplot(y)
> plot(lbxgh,y)
> abline(h=0)
> abline(v=6.5)
```

Yeo, In-Kwon and Johnson, Richard (2000) A new family of power transformations to improve normality or symmetry.*Biometrika*, 87, 954-959.

PROBLEM SET #2 STATISTICS 500 FALL 2015: DATA PAGE 2

Due in class at noon in class on Tuesday 24 November 2015

This is an exam. Do not discuss it with anyone.

You do not have to look at the following web pages unless you want to. The first describes lbxgh and suggests 6.5% as distinguishing diabetes. The second provides general information about NHANES.
http://www.niddk.nih.gov/health-information/health-topics/diagnostic-tests/a1c-test-diabetes/Pages/index.aspx

http://www.cdc.gov/nchs/nhanes.htm

You will be comparing sad and bmi as predictors of lbxgh, adjusting for age and gender.

Model 1: lbxgh = $\beta_0 + \beta_1$ female + β_2 age + β_3 bmi + β_4 sad + ε where ε is iid $N(0, \sigma^2)$

Model 2: y = $\gamma_0 + \gamma_1$ female + γ_2 age + γ_3 bmi + γ_4 sad + ζ where ζ is iid $N(0, \omega^2)$

Model 1 has betas and model 2 has gammas so that different things have different symbols. It does not matter which Greek letter we use.

Question 1 asks you to compare plots of studentized residuals for models 1 and 2. The best way to compare the plots is to use `par(mfrow=c(1,2))` to put two plots right next to each other. When thinking about Normal plots, it is often helpful to add the qqline, and to do the Shapiro-Wilk test, so do that here.

Note!: If a test can be done either as a t-test or as an F-test with equivalent results, do it as a t-test.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. **This is an exam**. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Last name:_____________________ First Name: _________________ ID#:___________________

Statistics 500, Problem 2, Fall 2015, p1. **This problem set is an exam. Do not discuss it with anyone**.

1. Fit models 1 and 2 from the data page. In question 1, always use the studentized residuals from `rstudent(.)`	**CIRCLE**/FILL IN THE CORRECT ANSWER
1.1 Compare the Normal plots of studentized residuals from models 1 and 2, and do the Shapiro-Wilk test on the studentized residuals. Is this true? "Neither set of studentized residuals looks Normal, but the Normal plot is closer to a line for model 2 than for model 1."	TRUE FALSE
1.2 The Normal plot of studentized residuals from model 1 shows them to be skewed right compared to the Normal distribution.	TRUE FALSE
1.3 The Normal plot of studentized residuals from model 2 shows the residuals to have shorter tails than the Normal distribution.	TRUE FALSE
1.4 Compare the boxplots of studentized residuals from models 1 and 2. Is this true? "The studentized residuals from model 1 look more nearly symmetric about their median than the residuals from model 2."	TRUE FALSE
1.5 If you were to test for outliers in model 1 at the two-sided 0.05 level using the Bonferonni adjustment, what is the critical value that an absolute studentized residual needs to exceed to be declared an outlier? What are the degrees of freedom? How many outliers are there?	Model 1. Critical value: ____________ Degrees of freedom: __________ Number of outliers: ___________
1.6 If you were to test for outliers in model 2 at the two-sided 0.05 level using the Bonferonni adjustment, what is the critical value that an absolute studentized residual needs to exceed to be declared an outlier? What are the degrees of freedom? How many outliers are there?	Model 2. Critical value: ____________ Degrees of freedom: __________ Number of outliers: ___________

2. Use model 2 in part 2.	
2.1 In model 2, which observation has the largest leverage? Give the row #. What is the numerical value of the leverage?	Row #__________ Leverage value:___________
2.2 The person identified in 2.1 has large leverage because, despite having a high bmi and sad, the person does not have lbxgh>6.5 or y>0.	TRUE FALSE
2.3 How many individuals have large leverage by our standard rule of thumb? Give one number.	How many?______________
2.4 If you added 1 to the y for the person identified in question 2.1 and reran the regression with this new y, the predicted value or yhat for this person would increase by more than 0.5.	TRUE FALSE

Last name:_____________________ First Name: _________________ ID#:___________________

Statistics 500, Problem 2, Fall 2015, p2. **This problem set is an exam. Do not discuss it with anyone**.

3. Question 3 asks about model 2.	**CIRCLE**/FILL IN THE CORRECT ANSWER
3.1 Which observation has the largest absolute value of dffits? Give a row number. What is the value (with its sign) of the dffits?	Row #___________ dffits:____________
3.2 The individual in 3.1 has a large dffits because she has the lowest lbxgh and y in the data set and the highest sad.	TRUE FALSE
3.2 Using dfbetas, the individual in 3.1 is seen to move the coefficient of sad by half its standard error.	TRUE FALSE

:

4. Question 4 asks about model 2 and its extensions.	**CIRCLE**/FILL IN THE CORRECT ANSWER
4.1 In model 2, test for interaction between sad and female. What is the name of the test statistic? What is the numerical value of the test statistic? What is P-value? Is it plausible that there is no interaction between sad and female?	Name:____________ Value: ______________ P-value:______ Circle one PLAUSIBLE NOT PLAUSIBLE
4.2 In model 2, use a **centered** quadratic term in sad to test whether the relationship with sad is linear. What is the name of the test statistic? What is the numerical value of the test statistic? What is P-value? Is it plausible that relationship between y and sad is linear?	Name:____________ Value: ______________ P-value:______ Circle one PLAUSIBLE NOT PLAUSIBLE
4.3 Use Tukey's one-degree of freedom method to test the null hypothesis that no transformation of y is needed in model 2. What is the P-value? Is the null hypothesis plausible	P-value:______ Circle one PLAUSIBLE NOT PLAUSIBLE
4.4 The original question is whether sad is better than bmi at predicting diabetes. In the fit of model 2, it is clear that sad is not needed as predictor of y if you have bmi, age and female.	TRUE FALSE
4.5 Give the squared multiple correlation, R^2, for models 1 and 2.	Model 1 R^2:________ Model 2 R^2:________
4.6 If you plot residuals of model 2 (as y) against fitted values from model 2 (as x) and add a lowess smooth, you see a dramatic inverted-U shape.	TRUE FALSE

Answers

Statistics 500, Problem 2, Fall 2015, p1. **This problem set is an exam. Do not discuss it with anyone**.

1. Fit models 1 and 2 from the data page. In question 1, always use the studentized residuals from `rstudent(.) (5 points each)`	**CIRCLE**/FILL IN THE CORRECT ANSWER
1.1 Compare the Normal plots of studentized residuals from models 1 and 2, and do the Shapiro-Wilk test on the studentized residuals. Is this true? "Neither set of studentized residuals looks Normal, but the Normal plot is closer to a line for model 2 than for model 1."	[✓ TRUE] FALSE
1.2 The Normal plot of studentized residuals from model 1 shows them to be skewed right compared to the Normal distribution.	[✓ TRUE] FALSE
1.3 The Normal plot of studentized residuals from model 2 shows the residuals to have shorter tails than the Normal distribution.	TRUE [✓ FALSE]
1.4 Compare the boxplots of studentized residuals from models 1 and 2. Is this true? "The studentized residuals from model 1 look more nearly symmetric about their median than the residuals from model 2."	TRUE [✓ FALSE]
1.5 If you were to test for outliers in model 1 at the two-sided 0.05 level using the Bonferonni adjustment, what is the critical value that an absolute studentized residual needs to exceed to be declared an outlier? What are the degrees of freedom? How many outliers are there?	Model 1. Critical value: 4.402874 Degrees of freedom: 4570 Number of outliers: 64
1.6 If you were to test for outliers in model 2 at the two-sided 0.05 level using the Bonferonni adjustment, what is the critical value that an absolute studentized residual needs to exceed to be declared an outlier? What are the degrees of freedom? How many outliers are there?	Model 2. Critical value: 4.402874 Degrees of freedom: 4570 Number of outliers: 4

2. Use model 2 in part 2. (5 points each)	
2.1 In model 2, which observation has the largest leverage? Give the row #. What is the numerical value of the leverage?	Row #3972 Leverage value: 0.0158
2.2 The person identified in 2.1 has large leverage because, despite having a high bmi and sad, the person does not have lbxgh>6.5 or y>0.	TRUE [✓ FALSE]
2.3 How many individuals have large leverage by our standard rule of thumb? Give one number.	How many? 182
2.4 If you added 1 to the y for the person identified in question 2.1 and reran the regression with this new y, the predicted value or yhat for this person would increase by more than 0.5.	TRUE [✓ FALSE]

Answers

Statistics 500, Problem 2, Fall 2015, p2. **This problem set is an exam. Do not discuss it with anyone**.

3. Question 3 asks about model 2. (5 points each)	**CIRCLE**/FILL IN THE CORRECT ANSWER
3.1 Which observation has the largest absolute value of dffits? Give a row number. What is the value (with its sign) of the dffits?	Row #1785 dffits: -0.243
3.2 The individual in 3.1 has a large dffits because she has the lowest lbxgh and y in the data set and the highest sad.	TRUE [✓ FALSE]
3.3 Using dfbetas, the individual in 3.1 is seen to move the coefficient of sad by half its standard error.	TRUE [✓ FALSE]

:

4. Question 4 asks about model 2 and its extensions. (6 points each, except 5 for 4.5)	**CIRCLE**/FILL IN THE CORRECT ANSWER
4.1 In model 2, test for interaction between sad and female. What is the name of the test statistic? What is the numerical value of the test statistic? What is P-value? Is it plausible that there is no interaction between sad and female?	Name: t-test Value: 0.199 P-value: 0.84 Circle one [✓ PLAUSIBLE] NOT PLAUSIBLE
4.2 In model 2, use a **centered** quadratic term in sad to test whether the relationship with sad is linear. What is the name of the test statistic? What is the numerical value of the test statistic? What is P-value? Is it plausible that relationship between y and sad is linear?	Name: t-test Value: 2.562 P-value: 0.0104 Circle one PLAUSIBLE [✓ NOT PLAUSIBLE]
4.3 Use Tukey's one-degree of freedom method to test the null hypothesis that no transformation of y is needed in model 2. What is the P-value? Is the null hypothesis plausible	P-value: 0.997 Circle one [✓ PLAUSIBLE] NOT PLAUSIBLE
4.4 The original question is whether sad is better than bmi at predicting diabetes. In the fit of model 2, it is clear that sad is not needed as predictor of y if you have bmi, age and female.	TRUE [✓ FALSE]
4.5 Give the squared multiple correlation, R^2, for models 1 and 2. (5 points)	Model 1 R^2: 0.1444 Model 2 R^2: 0.2668
4.6 If you plot residuals of model 2 (as y) against fitted values from model 2 (as x) and add a lowess smooth, you see a dramatic inverted-U shape.	TRUE [✓ FALSE]

Fall 2015, Problem Set 2, Doing the Problem Set in R

Part 1.

```
> m1<-lm(lbxgh ~ age + female + sad + bmi)
> m2<-lm(y ~ age + female + sad + bmi)
> par(mfrow=c(1,2))
> boxplot(m1$residuals)
> boxplot(m2$residuals)
> qqnorm(rstudent(m1))
> qqline(rstudent(m1))
> qqnorm(rstudent(m2))
> qqline(rstudent(m2))
> shapiro.test(rstudent(m1))
        Shapiro-Wilk normality test
data:  rstudent(m1)
W = 0.63903, p-value < 2.2e-16
> shapiro.test(rstudent(m2))
        Shapiro-Wilk normality test
data:  rstudent(m2)
W = 0.97161, p-value < 2.2e-16
> boxplot(rstudent(m1))
> boxplot(rstudent(m2))
> qt(.025/4576,4570)
[1] -4.402874
> help("outlierTest")
> library(car)
> outlierTest(m1,n.max=2000)
      rstudent unadjusted p-value Bonferonni p
3292 11.826358         8.2646e-32   3.7819e-28
4313  9.537447         2.3109e-21   1.0575e-17
…
3967  4.414978         1.0335e-05   4.7294e-02
> length(outlierTest(m1,n.max=2000)$rstudent)
[1] 64
> sum(abs(rstudent(m1))>=4.402874)
[1] 64
> outlierTest(m2,n.max=2000)
      rstudent unadjusted p-value Bonferonni p
3362 -6.614237         4.1638e-11   1.9054e-07
1785 -5.760300         8.9470e-09   4.0942e-05
3988 -4.615171         4.0353e-06   1.8466e-02
898  -4.468362         8.0726e-06   3.6940e-02
> sum(abs(rstudent(m2))>=4.402874)
[1] 4
```

Part 2

```
> which.max(hatvalues(m2))
3972
3972
> nhanes14st500[3972,]
> summary(sad)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.10   19.30   22.20   22.62   25.40   40.00
> summary(bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.10   24.10   27.70   28.75   32.10   67.50
```

```
> max(hatvalues(m2))
[1] 0.0158098
> mean(hatvalues(m2))
[1] 0.001092657
> 2*mean(hatvalues(m2))
[1] 0.002185315
> 2*(5/4576)
[1] 0.002185315
> sum(hatvalues(m2)>=0.002185315)
[1] 182
```

Part 3

```
> which.max(abs(dffits(m2)))
1785
> nhanes14st500[1785,]
> dffits(m2)[1785]
-0.2430739
> dfbetas(m2)[1785,]
(Intercept)         age      female         sad         bmi
 0.10847980  0.06293416 -0.06076407  0.01112735 -0.08829110
```

Part 4

4.1

```
> fs<-female*sad
> m2a<-lm(y ~ age + female + sad + bmi + fs)
> summary(m2a)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.4558271  0.0660460 -37.184   <2e-16 ***
age          0.0139349  0.0005342  26.084   <2e-16 ***
female      -0.0092342  0.0872408  -0.106    0.916
sad          0.0476955  0.0050535   9.438   <2e-16 ***
bmi         -0.0046524  0.0032582  -1.428    0.153
fs           0.0007604  0.0038272   0.199    0.843
```

4.2

```
> s2<-(sad-mean(sad))^2
> m2b<-lm(y ~ age + female + sad + bmi + s2)
> summary(m2b)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.4394789  0.0481298 -50.685   <2e-16 ***
age          0.0140531  0.0005355  26.245   <2e-16 ***
female       0.0056009  0.0180846   0.310   0.7568
sad          0.0471909  0.0049120   9.607   <2e-16 ***
bmi         -0.0055487  0.0032213  -1.723   0.0850 .
s2           0.0008271  0.0003229   2.562   0.0104 *
```

4.3

```
> summary(lm(y ~ age + female + sad + bmi+tukey1df(m2)))
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.4650399  0.0473411 -52.070   <2e-16 ***
age           0.0139385  0.0005348  26.065   <2e-16 ***
female        0.0077219  0.0180860   0.427    0.669
sad           0.0479333  0.0049454   9.693   <2e-16 ***
bmi          -0.0045282  0.0032177  -1.407    0.159
tukey1df(m2) -0.0004076  0.1090079  -0.004    0.997
> plot(m2$fitted.values,m2$residuals)
> lines(lowess(m2$fitted.values,m2$residuals),col="red")
```

PROBLEM SET #3 STATISTICS 500 FALL 2015: DATA PAGE 1

Due in class at noon, Thursday, 17 Decmeber 2015

This is an exam. Do not discuss it with anyone.

For part 1, the data are the same as for Problem Set 2. These data are from the 2013-2014 NHANES. The data are in an object nhanes1314st500 in the course workspace.

```
> dim(nhanes14st500)
[1] 4576   13
> head(nhanes14st500)
    seqn lbxgh female age systolic diastolic pulse weight height  bmi waist  sad          y
1  73557  13.9      0  69      122        72    86   78.3  171.3 26.7 100.0 20.5  1.2344306
2  73558   9.1      0  54      156        62    74   89.5  176.8 28.6 107.6 24.2  0.7429755
```

The variables are as follows. (i) seqn is the NHANES id number, (ii) lbxgh is a measure of glycohemoglobin, recorded as a percent, where high values, above 6.5, are indicative of diabetes, (iii) female = 1 for female, 0 for male, (iv) age is in years, (v) systolic and diastolic record blood pressure, (vi) pulse is pulse, (vii) weight is recorded in kilograms, (viii) height, waist circumference and sad are recorded in centimeters. sad is sagittal abdominal diameter, the height indicated below – it is thought to be an improvement on “bmi” = body mass index as a measure of health risk.

The final variable, y, is just a transformation of lbxgh as in Problem Set 2.
Create the following x matrix.

```
> x<-nhanes14st500[,c(3,4,8,9,10,11,12)]
> head(x)
   female age weight height  bmi waist  sad
1       0  69   78.3  171.3 26.7 100.0 20.5
2       0  54   89.5  176.8 28.6 107.6 24.2
> dim(x)
[1] 4576    7
```

So you will be considering 7 predictors, female, age, weight, height, bmi, waist and sad. As in the second problem set, you will predict y, the transformed lbxgh, contained in `nhanes14st500` as its last column. You will use leaps. You should plot C_P against the size of the model and think about the plot before doing part 1.

Several questions say “C_P suggests xyz.” Remember C_P is an estimate of a population quantity J_P, so C_P has some sampling error and is not equal to J_P. “C_P suggests xyz” means “if we ignore the sampling error in the estimate, pretending that $J_P=C_P$ then xyz would be true.” Example: “C_P suggests the moon is made of green cheese” means “if the true value of J_P were equal to the observed value of C_P, then the moon would be made of green cheese.”

PROBLEM SET #3 STATISTICS 500 FALL 2015: DATA PAGE 2

Due in class at noon on Thursday 17 December 2015

This is an exam. Do not discuss it with anyone.

The second data set for part 2 is from NHANES 2011-2012. The data are in an object `smkdustfume` in the course workspace and are available on my web page as `data.csv`. It has 868 people in four groups of size 217, 868=4x217, recorded in the variables dfsmkf and dfsmki. Groups `dustfume` and `both` were exposed at work for at least 10 years to dusts or fumes. Groups `smoker` and `both` were daily smokers. Groups `neither` and `dustfumes` smoked fewer than 100 cigarettes in their lives. The variable `dustfumesy` is the total years of exposure to dusts and fumes, whereas `mineraly`, `organic, efumesy` and `ofumesy` record years of exposure to mineral dust, organic dust, exhaust fumes and other fumes. Other variables are `SEQN = NHANES id, age, female=1, fev1, fvc, and fvratio = fev1/fvc`, where fvratio is the Tiffeneau-Pinelli index, a measure of chronic obstructive lung disease. Lower values of `fvratio` indicate poor lung function.

You should plot the data in `boxplot(fvratio~dfsmkf)`. You should see how big the groups are in `table(dfsmkf)`. Question 2.4 asks for group names: the names are in dfsmkf.

The model for part 2, **MODEL A**, is $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where for i = 1, 2, 3, 4 groups defined by dfmskf and j = 1, 2, …, 217 people in group i, where $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$ and ε_{ij} are iid $N(0,\sigma^2)$, and y_{ij} is the fvratio for the jth person in group i.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. **This is an exam**. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

The exam is due in my office, 473 JMHH, Thursday, Dec 17, 2014 at noon. You may turn in the exam early by placing it in an envelope addressed to me and leaving it in my mail box in statistics, 4th floor, JMHH. If you prefer, give it to Noelle at the front desk in statistics. **Make and keep a photocopy of your answer page**. The answer key will be posted in the revised bulk pack on-line. You can compare your photocopy to the on-line answer page.

Have a great holiday!

Last name:______________________ First Name: _________________ ID#:___________________

Statistics 500, Problem 3, Fall 2015, p1. **This problem set is an exam. Do not discuss it with anyone**.

Use leaps and the 7 predictors in x for the nhanes14st500 data to predict y.	Fill in or CIRCLE the correct answer.
1.1 With 7 predictors, how many possible regressions can be formed as subsets of the 7 predictors? In your count, include the regression with 7 predictors and the regression with no predictors.	Number: _________________
1.2 Among regressions with just one of the seven predictors, which single predictor yields the smallest C_P? What is the "size" of this model? What is the value of C_P?	Which predictor? (variable name):___________ Size=_______ C_P = _____________
1.3 A model with fewer predictors may or may not produce better predictions than a model with more predictors. Does the value of C_P in question 1.2 suggest it produces better predictions than the best two-predictor model (as judged by C_P)?	CIRCLE ONE YES NO
1.4 Of the 7 models that use 6 of the 7 predictors, which one is worst in the opinion of C_P? Name the variable left out of this worst model. What is the "size" of this model? What is the value of C_P?	Which predictor is left out:___________ Size=_______ C_P = _____________
1.5 The value of C_P for the model in question 1.4 suggests that the total squared error of predictions from the model in 1.4 are more than 60 times larger than the total squared error of the predictions from the model with all 7 variables.	CIRCLE ONE TRUE FALSE
1.6 Give the value of C_P for the best 5 predictor model and for the best 6 predictor model. List predictors in the best 6 predictor model that are not in the best 5 predictor model. List predictors, if any, in the best 5 predictor model that are not in the best 6 predictor model; if none, write "none". "5" is short for the best regression with 5 predictors, and "6" is short for the best regression with 6 predictors.	5 predictor 6 predictor C_P= ______________ C_P= ______________ Names of predictors in: 5 but not in 6: ____________________________ 6 but not in 5: ____________________________
1.7 The 6 predictor model in question 1.6 is estimated to make worse predictions than the 5 predictor model in question 1.6.	CIRCLE ONE TRUE FALSE
1.8 At the 0.05 level, Spjotvoll's method rejects as inadequate every model that excludes female, and every model that excludes sad, but it rejects neither of the two models in question 1.6.	CIRCLE ONE TRUE FALSE
1.9 Give the variance inflation factors for age and weight in the model with all 7 predictors.	Age vif = _________ Weight vif = ___________
1.10 Give the value of press for the 7 predictor model and the (bad) 6 predictor model in question 1.4.	7-predictor press:________ bad-6- press:_______

Last name:______________________ First Name: _________________ ID#:___________________

Due December 17, Thursday, noon. **This problem set is an exam. Do not discuss it with anyone**.

Use the smkdustfume data and MODEL A to answer questions in part 2.	Fill in or CIRCLE the correct answer
2.1 Do a one-way analysis of variance to test the null hypothesis that the four groups do not differ in their fvratios. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible?	Name:___________ Value: _______________ P-value:____________ PLAUSIBLE NOT PLAUSIBLE
2.2 Use Holm's method to compare the four groups in pairs controlling the family-wise error rate despite doing many t-tests. How many tests are done when comparing 4 groups in all possible pairs? List the pairs of groups that do **not** differ significantly at the 0.05 level by Holm's method, listing pairs of groups **by name**, so if abc does not differ from xyz, write (abc, xyz). If none, write NONE. Do not use group #s, **use group names**.	How many tests?_______________ Which pairs of groups do **not** differ? List (name1, name2) or write NONE.
2.3 Holm's method controls the familywise error rate in the weak sense but not in the strong sense, but the Bonferroni method controls in the strong sense.	TRUE FALSE
2.4 If you used the Bonferroni method, not the Holm method, the P-value comparing (both,smoker) would be 3 times as large as with Holm's method.	TRUE FALSE

2.5 In the table below, create 3 orthogonal contrasts, one representing the main effect of smoking, one representing the main effect of dust/fumes and one representing their interaction. Use integer contrast weights, not fractions or decimals.

Group	Neither	Smoker	Dustfumes	Both
Contrast:Smoking Main effect				
Contrast:Dust/Fumes Main effect				
Contrast:Interaction				

2.6 Use the contrasts in 2.5 to fill in the following detailed anova table. DF= degrees of freedom.

	Sum of squares	DF	Mean Square	F	P-value
Between Groups					
Smoking Main					
Dust/Fume Main					
Interaction					
Residual within groups				XXXXXX XXXXXX	XXXXXX XXXXXX

Statistics 500, Problem 3, Fall 2015, **ANSWERS**

Use leaps and the 7 predictors in x for the nhanes14st500 data to predict y.	Fill in or CIRCLE the correct answer. (6 points each)
1.1 With 7 predictors, how many possible regressions can be formed as subsets of the 7 predictors? In your count, include the regression with 7 predictors and the regression with no predictors.	Number: $2^7 = 128$
1.2 Among regressions with just one of the seven predictors, which single predictor yields the smallest C_P? What is the "size" of this model? What is the value of C_P?	Which predictor? (variable name): age Size=2 $C_P = 529.284$
1.3 A model with fewer predictors may or may not produce better predictions that a model with more predictors. Does the value of C_P in question 1.2 suggest it produces better predictions than the best two-predictor model (as judged by C_P)?	CIRCLE ONE YES [✓ NO]
1.4 Of the 7 models that use 6 of the 7 predictors, which one is worst in the opinion of C_P? Name the variable left out of this worst model. What is the "size" of this model? What is the value of C_P?	Which predictor is left out: age Size=7 $C_P = 503.73$
1.5 The value of C_P for the model in question 1.4 suggests that the total squared error of predictions from the model in 1.4 are more than 60 times larger than the total squared error of the predictions from the model with all 7 variables.	CIRCLE ONE [✓ TRUE] FALSE
1.6 Give the value of C_P for the best 5 predictor model and for the best 6 predictor model. List predictors in the best 6 predictor model that are not in the best 5 predictor model. List predictors, if any, in the best 5 predictor model that are not in the best 6 predictor model; if none, write "none". "5" is short for the best regression with 5 predictors, and "6" is short for the best regression with 6 predictors.	5 predictor 6 predictor C_P= 6.38 C_P= 6.07 Names of predictors in: 5 but not in 6: none 6 but not in 5: weight
1.7 The 6 predictor model in question 1.6 is estimated to make worse predictions than the 5 predictor model in question 1.6.	CIRCLE ONE TRUE [✓ FALSE]
1.8 At the 0.05 level, Spjotvoll's method rejects as inadequate every model that excludes female, and every model that excludes sad, but it rejects neither of the two models in question 1.6.	CIRCLE ONE [✓ TRUE] FALSE
1.9 Give the variance inflation factors for age and weight in the model with all 7 predictors.	Age vif = 1.455 Weight vif = 89.232
1.10 Give the value of press for the 7 predictor model and the (bad) 6 predictor model in question 1.4.	7-predictor press: 1399 bad-6- press: 1551

Use the smkdustfume data and MODEL A to answer questions in part 2.	Fill in or CIRCLE the correct answer 2.1-2.4 are 6 points, 2.5 and 2.6 are 8 points
2.1 Do a one-way analysis of variance to test the null hypothesis that the four groups do not differ in their fvratios. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible?	Name: F-test Value: 22.74 P-value: 3.56 x 10^{-14} PLAUSIBLE [✓ NOT PLAUSIBLE]
2.2 Use Holm's method to compare the four groups in pairs controlling the family-wise error rate despite doing many t-tests. How many tests are done when comparing 4 groups in all possible pairs? List the pairs of groups that do **not** differ significantly at the 0.05 level by Holm's method, listing pairs of groups **by name**, so if abc does not differ from xyz, write (abs, xyz). If none, write NONE. Do not use group #s, **use group names**.	How many tests? 6 Which pairs of groups do **not** differ? List (name1, name2) or write NONE. (dustfumes, neither)
2.3 Holm's method controls the familywise error rate in the weak sense but not in the strong sense, but the Bonferroni method controls in the strong sense	TRUE [✓ FALSE]
2.4 If you used the Bonferroni method, not the Holm method, the P-value comparing (both,smoker) would be 3 times as large as with Holm's method.	[✓ TRUE] FALSE

2.5 In the table below, create 3 orthogonal contrasts, one representing the main effect of smoking, one representing the main effect of dust/fumes and one representing their interaction. Use integer contrast weights, not fractions or decimals.

Group	Neither	Smoker	Dustfumes	Both
Contrast:Smoking Main effect	-1	1	-1	1
Contrast:Dust/Fumes Main effect	-1	-1	1	1
Contrast:Interaction	1	-1	-1	1

2.6 Use the contrasts in 2.5 to fill in the following detailed anova table. DF= degrees of freedom.

	Sum of squares	DF	Mean Square	F	P-value
Between Groups	0.4665	3	0.1555	22.74	3.56 x 10^{-14}
Smoking Main	0.4126	1	0.4126	60.34	2.25 x 10^{-14}
Dust/Fume Main	0.0316	1	0.0316	4.63	0.03177
Interaction	0.0222	1	0.0222	3.25	0.07184
Residual within groups	5.9083	864	0.0068	XXXXXX XXXXXX	XXXXXX XXXXXX

PROBLEM SET #3 STATISTICS 500 FALL 2015   ANSWERS
DOING THE PROBLEM SET IN R

1.1

```
> 2^7
[1] 12
> attach(nhanes14st500)
> x<-nhanes14st500[,c(3,4,8,9,10,11,12)]
> nhl<-leaps(x=x,y=y,names=colnames(x))
> plot(nhleaps$size,nhleaps$Cp)
> abline(0,1)
> plot(nhl$size,nhl$Cp)
> abline(0,1)
> cbind(nhl$which,nhl$size,nhl$Cp)[35:55,]
  female age weight height bmi waist sad
4      0   1      0      0   1     1   1 5  46.273615
4      1   1      0      0   0     1   1 5  46.286152
4      1   1      0      0   1     0   1 5  51.533812
5      1   1      0      1   1     0   1 6   6.375960
5      1   1      1      1   0     0   1 6   9.553443
5      1   1      0      1   0     1   1 6  12.083396
…
5      1   1      0      0   1     1   1 6  48.271690
6      1   1      1      1   1     0   1 7   6.066912
6      1   1      0      1   1     1   1 7   8.375293
6      1   1      1      1   0     1   1 7  11.390767
6      1   1      1      0   1     1   1 7  19.194041
6      0   1      1      1   1     1   1 7  21.556761
6      1   1      1      1   1     1   0 7  90.768968
6      1   0      1      1   1     1   1 7 503.726589
7      1   1      1      1   1     1   1 8   8.000000
```

```
> nhsp<-spjotvoll(x,y)
> nhsp[!nhsp$inadequate,]
    p     Cp    Fp  pval adjusted.pval inadequate female age weight height bmi waist sad
99  6  6.376 1.188 0.305         0.305      FALSE      1   1      0      1   1     0   1
100 6  9.553 2.777 0.062         0.062      FALSE      1   1      1      1   0     0   1
120 7  6.067 0.067 0.796         0.796      FALSE      1   1      1      1   1     0   1
121 7  8.375 2.375 0.123         0.305      FALSE      1   1      0      1   1     1   1
122 7 11.391 5.391 0.020         0.062      FALSE      1   1      1      1   0     1   1
127 8  8.000    NA 1.000         1.000      FALSE      1   1      1      1   1     1   1
```

```
> library(DAAG)
> m<-lm(y~female+age+weight+height+bmi+waist+sad)
> vif(m)
 female     age  weight  height     bmi   waist     sad
 1.9853  1.4553 89.2320 20.8960 70.8950 14.7730 11.1800
> m<-lm(y~female+age+weight+height+bmi+waist+sad)
> press(m)
[1] 1399.262
> m2<-lm(y~female+weight+height+bmi+waist+sad)
> press(m2)
[1] 1551.129
```

PROBLEM SET #3 STATISTICS 500 FALL 2015   ANSWERS
DOING THE PROBLEM SET IN R

2.

```
> attach(smkdustfume)
> boxplot(fvratio~dfsmkf)
> table(dfsmkf)
```

2.1

```
> anova(lm(fvratio~dfsmkf))
Analysis of Variance Table
Response: fvratio
           Df Sum Sq  Mean Sq F value    Pr(>F)
dfsmkf      3 0.4665 0.155499  22.739 3.559e-14 ***
Residuals 864 5.9083 0.006838

> pairwise.t.test(fvratio,dfsmkf)
        Pairwise comparisons using t tests with pooled SD
          both    dustfumes neither
dustfumes 1.2e-10 -         -
neither   2.8e-11 0.80542   -
smoker    0.01060 0.00023   0.00011
P value adjustment method: holm

> pairwise.t.test(fvratio,dfsmkf,p.adj="bonf")
        Pairwise comparisons using t tests with pooled SD
          both    dustfumes neither
dustfumes 1.5e-10 -         -
neither   2.8e-11 1.00000   -
smoker    0.03180 0.00046   0.00016
P value adjustment method: Bonferroni
0.03180 = 3*0.01060

> contrasts(dfsmkf)
> smk<-c(1,-1,-1,1)
> dufu<-c(1,1,-1,-1)
> interact<-smk*dufu
> contrasts(dfsmkf)<-cbind(smk,dufu,interact)
> contrasts(dfsmkf)
          smk dufu interact
both        1    1        1
dustfumes  -1    1       -1
neither    -1   -1        1
smoker      1   -1       -1
> m<-model.matrix(lm(fvratio~dfsmkf))
> head(m)
> cor(m[,2:4])
> m<-as.data.frame(m)
> summary(lm(fvratio~m$dfsmkfsmk+m$dfsmkfdufu+m$dfsmkfinteract))
> anova(lm(fvratio~m$dfsmkfsmk+m$dfsmkfdufu+m$dfsmkfinteract))
Analysis of Variance Table
Response: fvratio
                  Df Sum Sq Mean Sq F value    Pr(>F)
m$dfsmkfsmk        1 0.4126 0.41265 60.3442 2.254e-14 ***
m$dfsmkfdufu       1 0.0316 0.03163  4.6260   0.03177 *
m$dfsmkfinteract   1 0.0222 0.02221  3.2483   0.07184 .
Residuals        864 5.9083 0.00684
```

PROBLEM SET #1 STATISTICS 500 FALL 2014: DATA PAGE 1

Due in class at noon on Tuesday, October 21, 2014.

This is an exam. Do not discuss it with anyone.

The data are from NHANES, the 2009-2010 National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm). There is no need to visit the webpage unless you want to. The file adultcal describes calories consumed on the first interview day for individuals 20 years old or older. Calories are deduced from a food interview.

The file is simplified in several ways; in particular, missing data have been removed.

SEQN nhanes sequence number or id
age – age in year 0-19
female – 1 if female, 0 if male
ed and edf record education. Type table(adultcal$edf) for categories.
income – ratio of family income to the poverty level, capped at 5 times.
married – 1 if married or living with partner, 0 otherwise
bmi – body mass index
waist – waist circumference in cm 1 cm = 0.393701 inches
calories – calories consumed on first interview day

```
> head(adultcal)
    SEQN age female ed                      edf income married   bmi waist calories
1  51624  34      0  3      High School Grad/GED   1.36       1 32.22 100.4     1844
5  51628  60      1  3      High School Grad/GED   0.69       0 42.39 118.2     1913
6  51629  26      0  2               9-11 grade   1.01       1 32.61 103.7     3123
7  51630  49      1  4 Some college or AA degree   1.91       1 30.57 107.8     1345
10 51633  80      0  4 Some college or AA degree   1.27       1 26.04  91.1     1565
12 51635  80      0  2               9-11 grade   1.69       0 27.62 113.7     1479
```

```
> dim(adultcal)
[1] 5000   10
```

The data are in the object `adultcal` in the course workspace at http://www-stat.wharton.upenn.edu/~rosenbap/ . You will have to download the workspace again to have the current version with `adultcal`. If you download the workspace and `adultcal` is not there, it probably means that you web browser remembers the last time you downloaded the file and thinks (incorrectly) that you do not need to download it again – in this case, clear the browser's memory and try again. There is a csv file `adultcal.csv` with the data at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` if you wish to use software other than R. The csv file should open in excel and other packages.

PROBLEM SET #1 STATISTICS 500 FALL 2014: DATA PAGE 2

This is an exam. Do not discuss it with anyone. Due Tuesday, October 21, 2014

Before you do anything else, plot the data in various ways. For example:

```
> attach(adultcal)
> boxplot(calories)
> plot(age,calories)
> lines(lowess(age,calories),col="red",lwd=2)
> boxplot(calories~female)
> boxplot(calories~female:married)
 etc
```

DO NOT TURN IN THE PLOTS.

Model #1

calories $= \beta_0 + \beta_1$ age + β_2 female + β_3 ed + β_4 income + ε where ε are iid $N(0,\sigma^2)$

Model #2

calories $= \gamma_0 + \gamma_1$ age + β_2 female + γ_3 income + η where η are iid $N(0,\omega^2)$

Model #3

calories $= \lambda_0 + \lambda_1$ age + λ_2 female + ζ where ζ are iid $N(0,\kappa^2)$

Model 1 has slopes β (beta), while model 2 has slopes γ (gamma), so that different things have different names. The choice of Greek letters is arbitrary. The same is true for model 3.

It is often useful to put two plots next to each other on the same page so you can see the same point in both plots. If you type

```
> par(mfrow=c(1,2))
```

then the next two plots will appear on the same page, the first on the left, the second on the right. For example, you can compare a boxplot and a Normal quantile plot in this way. The command sets a graphics parameter (that's the 'par'), and it says that there should be 1 row of graphs with 2 columns, filling in the first row first. By setting graph parameters, you can control many aspects of a graph. The free document R for Beginners by Paradis (http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf) contains lots of useful information about graph parameters (see page 43).

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2014: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due **Tuesday, October 21 noon**

Question (Part 1)	CIRCLE the correct answer
1.1 There is one person who consumed more than 10,000 calories.	TRUE FALSE
1.2 The lower quartile of calories for males is above the median for females.	TRUE FALSE
1.3 The median waist size for females is more than 37 inches = 93.98 cm.	TRUE FALSE
1.4 Of the five education categories, the lowest median of calories is for the <9th grade category.	TRUE FALSE

Fit model 1 from the data page.	Fill in or CIRCLE the correct answer.
2.1 Test the null hypothesis that the coefficient of income in model 1 is zero, $H_0:\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the two-sided P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name:________ Value:__________ P-value: ____________________ Circle one: Plausible Not Plausible
2.2 Test the null hypothesis that all four coefficients are zero, $H_0:\beta_1=\beta_2=\beta_3=\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name:________ Value:__________ P-value: ____________________ Circle one: Plausible Not Plausible
2.3 Two people have the same gender, the same education, and the same income, but one is 30 years old and the other is 40 years old. Using just the least squares estimate of the coefficient β_1 of age, the model would guess that the 40 year-old consumes 500 calories less than the 30 year old.	TRUE FALSE
2.4 Give the 95% confidence interval for the coefficient β_2 of female. If a man and a woman had the same age, education and income, the model would predict higher calories consumed for the woman.	95% CI: [,] TRUE FALSE

Name: ______________________________ ID# __________________________

PROBLEM SET #1 STATISTICS 500 FALL 2014: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due **Tuesday, October 21 noon**

Fit models 2 and 3 from the data page.	Fill in or CIRCLE the correct answer.
3.1 Assuming model 2 is true, test the null hypothesis that the coefficient of income in model 2 is zero, $H_0:\gamma_3=0$. What is the name of the test? What is the numerical value of the test statistic? What is the two-sided P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name:________ Value:__________ P-value: ____________________ Circle one: Plausible Not Plausible
3.2 Assuming model 1 is true, test the null hypothesis that model 3 is also true, that is, test $H_0:\beta_3=\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name:________ Value:__________ P-value: ____________________ Circle one: Plausible Not Plausible

Use the fit of model 1 to answer questions in part 4.	Fill in or CIRCLE the correct answer.
4.1 The Normal quantile plot of the residuals from model 1 gives the appearance of residuals that are Normally distributed.	TRUE FALSE
4.2 Test Normality of the residuals using the Shapiro-Wilk test. What is the P-value?	P-value: ____________________
4.3 The Normal plot of residuals suggests negative skewness, a long left-hand tail, with too many people consuming far fewer calories than the model predicts. This impression of negative skewness is reinforced by a boxplot of the residuals.	TRUE FALSE
4.4 Plot residuals as y against fitted values as x. Plot the absolute value of residuals as y against fitted values as x. Add a lowess curve (in red, so you can see it) in the second plot. The assumption of constant variance is clearly violated here, with larger absolute residuals being more common at low fitted calories (say 1500) than at higher fitted calories (say 2500), so the variance looks larger when the fitted values are smaller.	TRUE FALSE

PROBLEM SET #1 STATISTICS 500 FALL 2014: ANSWER PAGE 1: ANSWERS

This is an exam. Do not discuss it with anyone. Due noon in class

7 points each, except as noted

Question (Part 1)	CIRCLE the correct answer
1.1 There is one person who consumed more than 10,000 calories.	[✓ TRUE] FALSE
1.2 The lower quartile of calories for males is above the median for females.	[✓ TRUE] FALSE
1.3 The median waist size for females is more than 37 inches = 93.98 cm.	[✓ TRUE] FALSE
1.4 Of the five education categories, the lowest median of calories is the <9th grade category.	[✓ TRUE] FALSE

Fit model 1 from the data page.	Fill in or CIRCLE the correct answer.
2.1 Test the null hypothesis that the coefficient of income in model 1 is zero, $H_0:\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the two-sided P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name: t-test Value: 1.469 P-value: 0.1420 Circle one: [✓ Plausible] Not Plausible
2.2 Test the null hypothesis that all four coefficients are zero, $H_0:\beta_1=\beta_2=\beta_3=\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name: F-test Value: 258.5 P-value: $<2.2 \times 10^{-16}$ Circle one: Plausible [✓ Not Plausible]
2.3 Two people have the same gender, the same education, and the same income, but one is 30 years old and the other is 40 years old. Using just the least squares estimate of the coefficient β_1 of age, the model would guess that the 40 year-old consumes 500 calories less than the 30 year old.	TRUE [✓ FALSE] -12.1585 x 10 is not -500
2.4 Give the 95% confidence interval for the coefficient β_2 of female. If a man and a woman had the same age, education and income, the model would predict higher calories consumed for the woman.	95% CI: [-735.8, -637.6] Women consume less, maybe -686.7 TRUE [✓ FALSE]

PROBLEM SET #1 STATISTICS 500 FALL 2014: ANSWER PAGE 2: ANSWERS

This is an exam. Do not discuss it with anyone. Due in class on

Fit models 2 and 3 from the data page.	Fill in or CIRCLE the correct answer.
3.1 Assuming model 2 is true, test the null hypothesis that the coefficient of income in model 1 is zero, $H_0:\gamma_3=0$. What is the name of the test? What is the numerical value of the test statistic? What is the two-sided P-value? Is the null hypothesis plausible using the conventional 0.05 level standard?	Name: t-test Value: 2.53 P-value: 0.0115 Circle one: Plausible [✓ Not Plausible]
3.2 Assuming model 1 is true, test the null hypothesis that model 3 is also true, that is, test $H_0:\beta_3=\beta_4=0$. What is the name of the test? What is the numerical value of the test statistic? What is the P-value? Is the null hypothesis plausible using the conventional 0.05 level standard? (9 points)	Name: F-test Value: 4.658 P-value: 0.009526 Circle one: Plausible [✓ Not Plausible]

Use the fit of model 1 to answer questions in part 4.	Fill in or CIRCLE the correct answer.
4.1 The Normal quantile plot of the residuals from model 1 gives the appearance of residuals that are Normally distributed.	TRUE [✓ FALSE] Remember, we expect a straight line for Normal data, and this is curved.
4.2 Test Normality of the residuals using the Shapiro-Wilk test. What is the P-value?	P-value: 2.2×10^{-16}
4.3 The Normal plot of residuals suggests negative skewness, a long left-hand tail, with too many people consuming far fewer calories than the model predicts. This impression of negative skewness is reinforced by a boxplot of the residuals.	TRUE [✓ FALSE] It is definitely skewed, but skewed right, not left. Sometimes people consume a lot more calories than the model expects
4.4 Plot residuals as y against fitted values as x. Plot the absolute value of residuals as y against fitted values as x. Add a lowess curve (in red, so you can see it) in the second plot. The assumption of constant variance is clearly violated here, with larger absolute residuals being more common at low fitted calories (say 1500) than at higher fitted calories (say 2500), so the variance looks larger when the fitted values are smaller.	TRUE [✓ FALSE] The assumption of constant variance looks wrong here, but calories are more unstable when fitted calories are higher.

Problem Set 1, Fall 2014
DOING THE PROBLEM SET IN R

```
1.1
> max(calories)
[1] 10463
1.2
> summary(calories[female==1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     70    1281    1691    1777    2150    5814 
> summary(calories[female==0])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    300    1735    2275    2450    2964   10460 
1.3
> tapply(waist*.393701,female,summary)
$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  25.08   35.91   39.49   40.01   43.66   66.42 
$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  23.50   33.62   37.64   38.24   42.24   64.96 
1.4
> tapply(calories,edf,summary)
$`<9th grade`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    279    1252    1734    1891    2336    6892 
$`9-11 grade`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    305    1415    1891    2087    2581    8077 
$`High School Grad/GED`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    164    1475    1994    2174    2643    9315 
$`Some college or AA degree`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    274    1464    1948    2136    2597   10460 
$`BA degree+`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     70    1513    1986    2110    2533    7301 
2
> md<-lm(calories~age+female+ed+income)
> summary(md)
Call:
lm(formula = calories ~ age + female + ed + income)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2958.7537    53.0998  55.721   <2e-16 ***
age          -12.1585     0.7219 -16.842   <2e-16 ***
female      -686.7017    25.0221 -27.444   <2e-16 ***
ed            18.8074    11.0156   1.707   0.0878 .  
income        12.8205     8.7301   1.469   0.1420    
Residual standard error: 880.8 on 4995 degrees of freedom
Multiple R-squared:  0.1715,    Adjusted R-squared:  0.1708 
F-statistic: 258.5 on 4 and 4995 DF,  p-value: < 2.2e-16
```

```
> confint(md)
                  2.5 %     97.5 %
(Intercept) 2854.654814 3062.85258
age          -13.573750  -10.74320
female      -735.756123 -637.64735
ed            -2.788033   40.40274
income        -4.294376   29.93546
```

3.1

```
> md2<-lm(calories~age+female+income)
> summary(md2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3013.9077    42.1511   71.50   <2e-16 ***
age          -12.3652     0.7118  -17.37   <2e-16 ***
female      -684.8067    25.0023  -27.39   <2e-16 ***
income        19.6402     7.7643    2.53   0.0115 *
```

3.2

```
> md3<-lm(calories~age+female)
> anova(md3,md)
Analysis of Variance Table
Model 1: calories ~ age + female
Model 2: calories ~ age + female + ed + income
  Res.Df        RSS Df Sum of Sq      F   Pr(>F)
1   4997 3882746217
2   4995 3875518001  2   7228216 4.6581 0.009526 **
```

4.

```
> qqnorm(md$resid)
> shapiro.test(md$resid)
        Shapiro-Wilk normality test
data:  md$resid
W = 0.9276, p-value < 2.2e-16
> boxplot(md$resid)
> plot(md$fit,md$resid)
> plot(md$fit,abs(md$resid)
> lines(lowess(md$fit,abs(md$resid)),col="red")
```

PROBLEM SET #2 STATISTICS 500 FALL 2014: DATA PAGE 1

Due in class at noon in class on Tuesday November 25, 2014.

This is an exam. Do not discuss it with anyone.

The data are from a paper by Card, Chetty and Weber (CCW) (2007) Cash-on-hand and competing models of intertemporal behavior: new evidence from the labor market. *Quarterly Journal of Economics*, 1511-1560. In their Table 1, their estimation sample had 650,922 people. The data sample for this problem set is simpler and smaller. It began as a random sample of 5000 individuals. Then people with missing data on key variables were removed, leaving 3923 people. The study is in various ways more complex than the data for the problem set, sometimes in interesting ways, but this is a problem set about regression, not a comprehensive study.

In Austria, a person who is unemployed may receive various unemployment benefits depending upon past employment. A person who has 3 or more years of job tenure receives a severance payment equivalent to two-months of pretax salary – i.e., two months wages. There is no severance payment for a person with less than 3 years of job tenure. The variables in the object ccwSt500 are as follows.

Variable name here	Variable name in CCW	Meaning
sevpay	sevpay	Severence pay, 1=yes, 0 =no
age	age	Age
female	female	1=female, 0=male
highed	high_ed	post-compulsory schooling, 1=yes, 0=no
married	married	1=married, 0=not
austrian	austrian	1=Austrian, 0=other
bluecollar	bluecollar	1=lost blue collar job, 0=lost other job
priorwage	ann_wage	Wage before job loss, euros/year
tenure	tenure_mths	Tenure (ie time) in previous job in months
uduration	uduration_mths	Duration of nonemployment in months
wrk12	wrk12<-pmax(12-uduration,0)	Months worked in the 12 months following unemployment
nextwage	ne_wage0	Monthly wage in next job (sometimes missing)
wage12	wage12<-ne_wage0*wrk12 wage12[wrk12==0]<-0	(Approximately) total wages in the 12 months after job loss.
id	1:3923	1 to 3923

The original paper had various motivations. One was to ask whether severance pay caused people to stay out of work longer. Another was to ask whether staying out longer was, in a sense, a good thing, because it gave people the chance to find a better job. You will do various regression with the data, but if you want to take a serious look at the paper's questions, you might want to look at the paper itself.

PROBLEM SET #2 STATISTICS 500 FALL 2014: DATA PAGE 2

This is an exam. Do not discuss it with anyone. Due Tuesday November 25, 2014.

```
> dim(ccwSt500)
[1] 3923   13
> colnames(ccwSt500)
 [1] "id"         "sevpay"     "female"     "highed"    
"married"    "austrian"   "bluecollar" "priorwage" 
"tenure"     "uduration"  "wrk12"      "nextwage"  
"wage12"
```

The data are in the object `ccwSt500` in the course workspace at http://www-stat.wharton.upenn.edu/~rosenbap/ . You will have to download the workspace again to have the current version with `ccwSt500`. If you download the workspace and `ccwSt500` is not there, it probably means that you web browser remembers the last time you downloaded the file and thinks (incorrectly) that you do not need to download it again – in this case, clear the browser's memory and try again. There is a csv file `ccwSt500.csv` with the data at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` if you wish to use software other than R. The csv file should open in excel and other packages.

The variable wage12 is (approximately) what the person earned in the year after losing the job: it is the number of months work in the next 12 months times the salary at the new job. It is not perfect – the person might have gotten a raise in the middle of the year – but let's ignore its imperfections. Define a new variable wageloss as the difference between the annual wage prior to job loss minus wage12

```
> attach(ccwSt500)
> wageloss<-priorwage-wage12
```

so this what the individual would have earned at the salary of the job just lost minus what the individual earned in the year after job loss. A positive number is **bad** for the individual: it means his/her wage income went down quite a bit, by not working and perhaps by working for less. A negative number is **good** for the individual – despite the job loss, he/she earned more. How many people had a negative wage loss?

Model #1

$$\text{wageloss} = \beta_0 + \beta_1 \text{ sevpay} + \beta_2 \text{ age} + \beta_3 \text{ female} + \beta_4 \text{ highed} + \beta_5 \text{ married} + \beta_6 \text{ austrian} + \beta_7 \text{ bluecollar} + \varepsilon \quad \text{where} \quad \varepsilon \text{ are iid } N(0,\sigma^2)$$

After fitting model 1, you should plot the residuals in the usual ways, even if questions do not ask you to do this.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

Last name: ______________ **First** name:_____________ ID# __________________

PROBLEM SET #2 STATISTICS 500 FALL 2014: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due **Tuesday November 25, 2014.**

Fit model 1 on the data page. Use it for the questions in part 1.	Fill in or circle the correct answer.
1.1 Give the estimate and the 95% confidence interval for the β_1 the coefficient of sevpay.	Estimate:_______ CI: [,]
1.2 For two people who look the same in terms of all predictors in model 1 except sevpay, the model predicts 1081.78 euro less wage loss for the person who received the severance payment.	TRUE FALSE
1.3 Model 1 assumes that the relationship between wageloss and age is parallel for sevpay=1 and sevpay=0. In model 1, test this assumption. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), the two-sided p-value and state whether parallelism is plausible.	Name: __________ Value:____________ DF: _____________ P-value:__________ PLAUSIBLE NOT PLAUSIBLE
1.4 Model 1 assumes that the relationship between wage loss and age is linear, not quadratic. In model 1, test the assumption that the relationship is linear against the alternative that it is quadratic in age. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom, the two-sided p-value and state whether a linear relationship with age is plausible.	Name: __________ Value:____________ DF: _____________ P-value:__________ PLAUSIBLE NOT PLAUSIBLE
1.5 Use Tukey's method to test the null hypothesis that no transformation of wageloss is needed against the alternative that a power transformation would be helpful. Give the name, value, DF, two-sided p-value and state whether no transformation is needed.	Name: __________ Value:____________ DF: _____________ P-value:__________ The null hypothesis that no transformation is needed is: PLAUSIBLE NOT PLAUSIBLE
1.6 $(y^p-1)/p$ is very close to the base-10 log of y, that is $\log_{10}(y)$, for p very near 0.	TRUE FALSE
1.7 In model 1, test the hypothesis that there is no interaction between female and married. Give the name, value, DF, two-sided p-value and state whether the hypothesis of no interaction is plausible.	Name: __________ Value:____________ DF: _____________ P-value:__________ PLAUSIBLE NOT PLAUSIBLE

Last name: _______________ **First** name:_____________ ID# __________________

PROBLEM SET #2 STATISTICS 500 FALL 2014: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due **Tuesday November 25, 2014.**

	Fill in or circle the correct answer.
2.1 Which observation in model 1 has the largest leverage (i.e. hatvalue)? Give the id number in the last column of ccwSt500. What is the numerical value of this person's leverage? Is this leverage large by the rule judging the size of the leverages?	id = __________ leverage = ____________ Circle one LARGE NOT LARGE
2.2 The individual identified in 2.1 has the leverage he/she does because the wage12 is so much lower than the priorwage.	TRUE FALSE
2.3 Which observation in model 1 has the largest absolute studentized residual (rstudent)? Give the id number in the last column of ccwSt500. What is the numerical value of this person's studentized residual (with its sign + or –)? This individual went from a low priorwage to a much higher wage12 (True or false)?	id = __________ studentized residual = ____________ Circle one TRUE FALSE
2.4 Is the person in 2.3 an outlier at the 0.05 level? What absolute value of the studentized residual would just barely reject a person as an outlier at the 0.05 level in model 1? What are the degrees of freedom used in computing this cut-off value?	Circle one Outlier: YES NO Value: _______________ Degrees of freedom: __________
2.5 Testing the null hypothesis of no outliers at the 0.05 level using the Bonferroni inequality with studentized residuals means we expect only one out of every 20 people in a regression to be wrongly judged an outlier.	Circle one TRUE FALSE
2.6 Which person had the largest absolute dffits? Give the id#. What is the value of this person's dffits with its sign +/-. This individual moved his/her fitted value up by 3.1 times its standard error (T/F)	id = __________dffits = ____________ Circle one TRUE FALSE
2.7 Do a Normal quantile plot of residuals from model 1 and add a qqline. The person identified in 2.6 is recognizably off the line, but even without this person, the residuals look longer-tailed than Normal.	Circle one TRUE FALSE

PROBLEM SET #2 STATISTICS 500 FALL 2014: ANSWER PAGE 1**: ANSWERS**

Fit model 1 on the data page. Use it for the questions in part 1.	Fill in or circle the correct answer.
1.1 Give the estimate and the 95% confidence interval for the β_1 the coefficient of sevpay.	Estimate: 1081.78 CI: [510.2, 1653.4]
1.2 For two people who look the same in terms of all predictors in model 1 except sevpay, the model predicts 1081.78 euro less wage loss for the person who received the severance payment.	TRUE [✓ FALSE] Associated with more, not less, wage loss.
1.3 Model 1 assumes that the relationship between wageloss and age is parallel for sevpay=1 and sevpay=0. In model 1, test this assumption. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom (DF), the two-sided p-value and state whether parallelism is plausible.	Name: t-statistic Value: 1.316 DF: 3914 P-value: 0.188 [✓ PLAUSIBLE] NOT PLAUSIBLE
1.4 Model 1 assumes that the relationship between wage loss and age is linear, not quadratic. In model 1, test the assumption that the relationship is linear against the alternative that it is quadratic in age. Give the name of the test statistic, the numerical value of the test statistic, its degrees of freedom, the two-sided p-value and state whether a linear relationship with age is plausible.	Name: t-statistic Value: -2.225 DF: 3914 P-value: 0.026 PLAUSIBLE [✓ NOT PLAUSIBLE]
1.5 Use Tukey's method to test the null hypothesis that no transformation of wageloss is needed against the alternative that a power transformation would be helpful. Give the name, value, DF, two-sided p-value and state whether no transformation is needed.	Name: t-statistic Value: 6.14 DF: 3914 P-value: 9.31 x 10^{-10} The null hypothesis that no transformation is needed is: PLAUSIBLE [✓ NOT PLAUSIBLE]
1.6 $(y^p-1)/p$ is very close to the base-10 log of y, that is $\log_{10}(y)$, for p very near 0. *Base 10 and base e are different!*	TRUE [✓ FALSE] log10(3)= 0.4771213, log(3)=1.098612 ((3^0.001)-1)/0.001=1.099216
1.7 In model 1, test the hypothesis that there is no interaction between female and married. Give the name, value, DF, two-sided p-value and state whether the hypothesis of no interaction is plausible.	Name: t-statistic Value: 2.419 DF: 3914 P-value: 0.0156 PLAUSIBLE [✓ NOT PLAUSIBLE]

PROBLEM SET #2 STATISTICS 500 FALL 2014: ANSWER PAGE 2: ANSWERS

	Fill in or circle the correct answer.
2.1 Which observation in model 1 has the largest leverage (i.e. hatvalue)? Give the id number in the last column of ccwSt500. What is the numerical value of this person's leverage? Is this leverage large by the rule judging the size of the leverages?	id = 659 leverage = 0.00557 Circle one [✓ LARGE] NOT LARGE
2.2 The individual identified in 2.1 has the leverage he/she does because the wage12 is so much lower than the priorwage.	TRUE [✓ FALSE] Leverage is about x, not about y.
2.3 Which observation in model 1 has the largest absolute studentized residual (rstudent)? Give the id number in the last column of ccwSt500. What is the numerical value of this person's studentized residual (with its sign + or –)? This individual went from a low priorwage to a much higher wage12 (True or false)?	id = 393 studentized residual = -5.209 Circle one [✓ TRUE] FALSE
2.4 Is the person in 2.3 an outlier at the 0.05 level? What absolute value of the studentized residual would just barely reject a person as an outlier at the 0.05 level in model 1? What are the degrees of freedom used in computing this cut-off value?	Circle one Outlier: [✓ YES] NO Value: 4.370027 Degrees of freedom: 3914
2.5 Testing the null hypothesis of no outliers at the 0.05 level using the Bonferroni inequality with studentized residuals means we expect only one out of every 20 people in a regression to be wrongly judged an outlier.	Circle one TRUE [✓ FALSE] That would mean 3923*0.05 = 196 false outliers in the current data set!
2.6 Which person had the largest absolute dffits? Give the id#. What is the value of this person's dffits with its sign +/-. This individual moved his/her fitted value up by 3.1 times its standard error (T/F)	id = 393 dffits = -0.31 Circle one TRUE [✓ FALSE]
2.7 Do a Normal quantile plot of residuals from model 1 and add a qqline. The person identified in 2.6 is recognizably off the line, but even without this person, the residuals look longer-tailed than Normal.	Circle one [✓ TRUE] FALSE These data would be more appropriate for m-estimation or some other robust or nonparametric method.

Statistics 500 Fall 2014 Problem Set 2

Doing the Problem Set in R

```
> attach(ccwSt500)
> wageloss<-priorwage-wage12
Question 1.1
> mod<-
lm(wageloss~sevpay+age+female+highed+married+austrian+bluecollar)
> summary(mod)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3499.90     753.14   4.647 3.48e-06 ***
sevpay       1081.78     291.56   3.710 0.000210 ***
age           130.41      16.58   7.865 4.73e-15 ***
…
bluecollar  -2376.52     266.18  -8.928  < 2e-16 ***
---
> confint(mod)
                  2.5 %     97.5 %
(Intercept)  2023.31434  4976.4899
sevpay        510.15462  1653.4031
age            97.90418   162.9198
…
Question 1.3
> isevpayage<-sevpay*age
>summary(lm(wageloss~sevpay+age+female+highed+married+austrian+
    bluecollar+isevpayage))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3819.27     791.23   4.827 1.44e-06 ***
sevpay       -460.82    1208.27  -0.381 0.702935
age           120.34      18.26   6.590 4.97e-11 ***
…
isevpayage     47.79      36.33   1.316 0.188395
```

```
Question 1.4
> age2<-(age-mean(age))^2
> summary(lm(wageloss~sevpay+age+female+highed+married+austrian+
     bluecollar+age2))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3069.583    777.216   3.949 7.97e-05 ***
sevpay       1057.242    291.622   3.625 0.000292 ***
…
age2           -4.279      1.923  -2.225 0.026156 *
Question 1.5
> summary(lm(wageloss~sevpay+age+female+highed+married+austrian+
    bluecollar+tukey1df(mod)))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2856.3795   756.9425   3.774 0.000163 ***
sevpay          948.5399   291.0163   3.259 0.001126 **
…
tukey1df(mod)     1.8795     0.3063   6.136 9.31e-10 ***

Question 1.7
> fm<-female*married
> summary(lm(wageloss~sevpay+age+female+highed+married+austrian+
```

```
    bluecollar+fm))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3694.32     756.95   4.881 1.10e-06 ***
sevpay       1091.00     291.40   3.744 0.000184 ***
…
fm           1162.50     480.51   2.419 0.015596 *
Question 2.1
> which.max(hatvalues(mod))
659
659
> ccwSt500[659,]
    sevpay age female highed married austrian bluecollar priorwage
tenure uduration    wrk12 nextwage   wage12  id
834      0  49      1      0       1        0          1  19698.12
23.22581 0.6451613 11.35484 772.9453 8776.669 659
> hatvalues(mod)[659]
        659
0.005571402
Question 2.3
> rstudent(mod)[393]
      393
-5.209135
> ccwSt500[393,]
    sevpay age female highed married austrian bluecollar priorwage
tenure uduration    wrk12 nextwage   wage12  id
492      1  44      0      0       0        1          0  1348.568
51.48387 0.4516129 11.54839 2427.689 28035.89 393
Question 2.4
> dim(ccwSt500)
[1] 3923   14
> qt(0.025/3923,3914)
[1] -4.370027
Question 2.5
> which.max(abs(dffits(mod)))
393
393
> dffits(mod)[393]
       393
-0.3108647
Question 2.7
> qqnorm(mod$resid)
> which.min(mod$resid)
393
393
> qqline(mod$resid)
```

PROBLEM SET #3 STATISTICS 500 FALL 2014: DATA PAGE 1

Due at noon Thursday, December 18, in my office, 473 JMHH.

This is an exam. Do not discuss it with anyone.

Data set `antineoplastic` is adapted from an article by Kopjar and Garaj-Vrhovac in Mutagenesis (2001) vol 16, #1, pp 71-78, concerned with the possibility that working with antineoplastic drugs (i.e., cancer chemotherapies) might damage the DNA of the nurses and doctors. DNA damage is measured by the tail moment of the comet assay. In the comet assay, genetic material from a cell is placed in an electric field, which pulls DNA in one direction. DNA is a large and hard to move, but if some of the DNA is broken it moves more. The assay creates the appearance of a comet with a tail, intact DNA being in the comet's head, broken DNA in the tail. Large values of the tail moment y_{ij} are taken to mean greater DNA breakage, not a good thing. There are 3 groups of different people, i=1,2,3, each of size 20, j=1,…,20: (i) gloves = workers wearing gloves, (ii) hood = workers wearing gloves in a safety cabinet under a laminar hood in a that vents fumes upwards, (iii) control = individuals with no exposure to antineoplastic drugs. You should do `boxplot(tailmoment~group)` and think about what you see.

Model 1 says $y_{ij} = \mu + \gamma_i + \varepsilon_{ij}$ where ε_{ij} are iid $N(0,\sigma^2)$ and $\gamma_1 + \gamma_2 + \gamma_3 = 0$ **Hypothesis set A** has three hypotheses A = $\{H_{12}: \gamma_1 = \gamma_2, H_{13}: \gamma_1 = \gamma_3, H_{23}: \gamma_2 = \gamma_3\}$. Hypothesis set B is similar to hypothesis set A, but it is for the situation with 4 groups rather than the 3 groups in the antineoplastic data. **Hypothesis set B** has six hypotheses B = $\{H_{12}: \gamma_1 = \gamma_2, H_{13}: \gamma_1 = \gamma_3, H_{23}: \gamma_2 = \gamma_3, H_{14}: \gamma_1 = \gamma_4, H_{24}: \gamma_2 = \gamma_4, H_{34}: \gamma_3 = \gamma_4\}$. When a question asks about hypothesis set B, it is not a question about the antineoplastic data, but about some analogous situation with 4 groups instead of three.

Question 2.5 asks you to build two orthogonal contrasts, `exp.control` and `glove.hood`. It is an easy question, but if you mess it up, then you mess up several more questions.

The two contrasts you built in question 2.5 constitute 2 new variables. **Model 2** uses these two variables plus two more from the data set, namely `age` and `smoker`, which is 1 for a smoker, 0 for a nonsmoker. **Model 2** is:

tailmoment = $\beta_0 + \beta_1$ `exp.control` + β_2 `glove.hood` + β_3 `age` + β_4 `smoker` + ε_{ij} where ε_{ij} are iid $N(0,\sigma^2)$.

Model 2 has 2^4 submodels formed by deleting predictors, including the model with all 4 predictors and the model with no predictors (just the constant).

Model 3 has

tailmoment = $\beta_0 + \beta_1$ `exp.control` + β_2 `glove.hood` + ε_{ij} where ε_{ij} are iid $N(0,\sigma^2)$, so it is like model 2, but without age and smoker.

Question 2.1 mentions the experimentwise error rate and the familywise error rate. These are two names for the same thing.

PROBLEM SET #3 STATISTICS 500 FALL 2014: DATA PAGE 2

This is an exam. Do not discuss it with anyone.
Due at noon Thursday, December 18, in my office, 473 JMHH.

The data are in the object `antineoplastic` in the course workspace at http://www-stat.wharton.upenn.edu/~rosenbap/ . You will have to download the workspace again to have the current version with `antineoplastic`. If you download the workspace and `antineoplastic` is not there, it probably means that you web browser remembers the last time you downloaded the file and thinks (incorrectly) that you do not need to download it again – in this case, clear the browser's memory and try again. There is a csv file `antineoplastic.csv` with the data at `http://stat.wharton.upenn.edu/statweb/course/Fall-2008/stat500/` if you wish to use software other than R. The csv file should open in excel and other packages.

Follow instructions. **Write your name** on both sides of the answer page. If a question has several parts, **answer every part**. Turn in **only the answer page**. Do not turn in additional pages. Do not turn in graphs. **Brief answers suffice**. Do not circle TRUE adding a note explaining why it might be false instead. If a question asks you to circle an answer, then you are correct if you **circle the correct answer** and wrong if you circle the wrong answer. If you cross out an answer, no matter which answer you cross out, the answer is wrong. If a true/false question says A&B&C and if C is false, then A&B&C is false, even if A&B is true. This is an exam. **Do not discuss the exam with anyone**. If you discuss the exam, you have cheated on an exam. The single dumbest thing a PhD student at Penn can do is cheat on an exam.

The exam is due Thursday, Dec 18, 2014 at noon. You may turn in the exam early by placing it in an envelope addressed to me and leaving it in my mail box in statistics, 4th floor, JMHH. If you prefer, give it to Noel at the front desk in statistics. **Make and keep a photocopy of your answer page**. The answer key will be posted in the revised bulk pack on-line. You can compare your photocopy to the on-line answer page.

HAVE A GREAT HOLIDAY!

Last name: ______________ **First** name:_____________ ID# __________________

PROBLEM SET #3 STATISTICS 500 FALL 2014: ANSWER PAGE 1

This is an exam. Do not discuss it with anyone. Due noon Thursday, December 18.

Use the antineoplastic data and model 1 to answer the following questions.	Fill in or CIRCLE the correct answer
1.1 Assuming model 1 is true, test the null hypothesis H_0: $\gamma_1 = \gamma_2 = \gamma_3 = 0$. Give the name of the test, the value of the test statistics, the degrees of freedom (DF), the P-value and indicate whether the null hypothesis is plausible.	Name:_________ Value: ___________ DF:____________ P-value:__________ PLAUSIBLE NOT PLAUSIBLE
1.2 Assuming model 1 is true, use the Tukey method to build three simultaneous 95% confidence intervals for $\gamma_1 - \gamma_2$, $\gamma_1 - \gamma_3$, and $\gamma_2 - \gamma_3$, but REPORT HERE ONLY the interval for $\gamma_1 - \gamma_2 = \gamma_{glove} - \gamma_{hood}$, being careful to get the sign (+/-) correct for glove-minus-hood. The interval suggests that people working with a hood and gloves had lower tail moments than people working with just gloves. (T or F)	95% simultaneous interval: $\gamma_1 - \gamma_2 = \gamma_{glove} - \gamma_{hood}$ [,] TRUE FALSE
1.3 If model 1 were true, the three intervals in 1.2 would all cover their three parameters in at least 95% of experiments.	TRUE FALSE
1.4 Under model 1 for the antineoplastic data, if you used the t-test to compare two group means, but you used the three group pooled estimate of σ^2, then what would be the degrees of freedom (DF3) for the test? What would the degrees of freedom (DF2) be if you used just the data from the two groups being compared to estimate σ^2? What is the two-sided 95% critical value for |t| for a **single** t-test with the corresponding DF?	3 group estimate DF3: _____________ 2 group estimate DF2: _____________ 3 group critical |t|:________________ 2 group critical |t|:________________ (This question asks about doing **one** t-test with DF3 or DF2 degrees of freedom. It is **NOT** about testing **several** hypotheses.)
1.5 If you test k true null hypotheses, and there is probability λ that you reject each one, then you expect to reject λk of these true null hypotheses.	TRUE FALSE
1.6 Use Holm's method and the two-sided pairwise t-test with pooled estimate of σ^2 to test the three null hypothesis in hypothesis set A. Give the Holm-adjusted p-values.	Give 3 adjusted p-values H_{12}: $\gamma_1 = \gamma_2$ or $\gamma_{glove} = \gamma_{hood}$:___________ H_{13}: $\gamma_1 = \gamma_3$, or $\gamma_{glove} = \gamma_{control}$: _________

	H_{23}: $\gamma_2 = \gamma_3$}, or $\gamma_{hood} = \gamma_{control}$: __________

Due **Last** name: _______________ **First** name:_____________ ID#______________

PROBLEM SET #3 STATISTICS 500 FALL 2014: ANSWER PAGE 2

This is an exam. Do not discuss it with anyone. Due noon Thursday Dec 18.

Use model 1 and the antineoplastic data to answer the following questions.	Fill in or CIRCLE the correct answer
2.1 If pairwise t-tests adjusted by the Bonferroni inequality reject a particular hypothesis with an experimentwise or familywise error rate of 0.05, then Holm's method rejects that hypothesis also.	TRUE FALSE
2.2 Hypothesis set A could contain exactly 1 true hypothesis or exactly 2 true hypotheses or exactly 3 true hypotheses, which is why the R function defaults to Holm's method.	TRUE FALSE
2.3 Hypothesis set B could contain exactly 1 true and 5 false null hypotheses.	TRUE FALSE
2.4 In hypothesis set A for model 1, the contrasts for Hypotheses H_{12}: $\gamma_1 = \gamma_2$ and H_{23}: $\gamma_2 = \gamma_3$ are two orthogonal contrasts.	TRUE FALSE
2.5 Give two orthogonal contrasts with integer weights for (a) exposed to drugs versus control, (b) glove only vs gloves plus hood. Fill in 6 integers as contrast weights for 2 contrasts.	a ______ _______ ________ b ______ _______ ________ Glove Hood Control

3. Use model 1 and the contrasts in question 2.5 to fill in the following anova table.

	Sum of squares	DF	Mean Square	F
Between Groups				
Exposed versus Control				
Gloves versus Hood				
With Groups Residual				

Use models 2 and 3 for part 4.	Fill in or CIRCLE the correct answer
4.1 Which of the 16 submodels of model 2 has the smallest C_P? List the variables in this model, its size (1+#vars), the value of C_P.	Variable in this model (**list names**): Size= _________ C_P= ______________
4.2 What is the PRESS value for model 2 and for model 3?	Model 2 Model 3 PRESS= ___________ ____________
4.3 How many observations have large leverages or hatvalues in model 2? In model 3? Give two counts.	Model 2:________ Model 3:___________

4.4 Give the variance inflation factors for models 2 and 3.

Put in VIFs	exp.control	glove.hood	age	smoker
Model 2				
Model 3			XXXXXXXX	XXXXXXXX

Answers

PROBLEM SET #3 STATISTICS 500 FALL 2014: ANSWER PAGE 1

6 points each, except #3 which is 10 points

Use the antineoplastic data and model 1 to answer the following questions.	Fill in or CIRCLE the correct answer
1.1 Assuming model 1 is true, test the null hypothesis $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0$. Give the name of the test, the value of the test statistics, the degrees of freedom (DF), the P-value and indicate whether the null hypothesis is plausible.	Name: F-test Value: 56.62 DF: 2 and 57 P-value: 2.86×10^{-14} PLAUSIBLE [✓ NOT PLAUSIBLE]
1.2 Assuming model 1 is true, use the Tukey method to build three simultaneous 95% confidence intervals for $\gamma_1 - \gamma_2$, $\gamma_1 - \gamma_3$, and $\gamma_2 - \gamma_3$, but REPORT HERE ONLY the interval for $\gamma_1 - \gamma_2 = \gamma_{glove} - \gamma_{hood}$, being careful to get the sign (+/-) correct for glove-minus-hood. The interval suggests that people working with a hood and gloves had lower tail moments than people working with just gloves. (T or F)	95% simultaneous interval: $\gamma_1 - \gamma_2 = \gamma_{glove} - \gamma_{hood}$ [.969 , 3.481] [✓ TRUE] FALSE
1.3 If model 1 were true, the three intervals in 1.2 would all cover their three parameters in at least 95% of experiments.	[✓ TRUE] FALSE
1.4 Under model 1 for the antineoplastic data, if you used the t-test to compare two group means, but you used the three group pooled estimate of σ^2, then what would be the degrees of freedom (DF3) for the test? What would the degrees of freedom (DF2) be if you used just the data from the two groups being compared to estimate σ^2? What is the two-sided 95% critical value for $\lvert t\rvert$ for a **single** t-test with the corresponding DF?	3 group estimate DF3: 57 = 60-3 2 group estimate DF2: 38 = 40-2 3 group critical $\lvert t\rvert$: 2.002 2 group critical $\lvert t\rvert$: 2.024 Once you have 38 DF for error, going to 57 does not add much. And you have to assume the three groups have the same σ^2 to get the extra DF. The 2-group test doesn't assume anything about the third group.
1.5 If you test k true null hypotheses, and there is probability λ that you reject each one, then you expect to reject λk of these true null hypotheses.	[✓ TRUE] FALSE
1.6 Use Holm's method and the two-sided pairwise t-test with pooled estimate of σ^2 to test the three null hypothesis in hypothesis set A. Give the Holm-adjusted p-values.	Give 3 adjusted p-values $H_{12}: \gamma_1 = \gamma_2$ or $\gamma_{glove} = \gamma_{hood}$: 7.7×10^{-5} $H_{13}: \gamma_1 = \gamma_3$, or $\gamma_{glove} = \gamma_{control}$: 1.4×10^{-14}

	H_{23}: $\gamma_2 = \gamma_3$}, or $\gamma_{hood} = \gamma_{control}$: 8.8×10^{-8}

Answers

PROBLEM SET #3 STATISTICS 500 FALL 2014: ANSWER PAGE 2

Use model 1 and the antineoplastic data to answer the following questions.	Fill in or CIRCLE the correct answer
2.1 If pairwise t-tests adjusted by the Bonferroni inequality reject a particular hypothesis with an experimentwise or familywise error rate of 0.05, then Holm's method rejects that hypothesis also.	[✓ TRUE] FALSE
2.2 Hypothesis set A could contain exactly 1 true hypothesis or exactly 2 true hypotheses or exactly 3 true hypotheses, which is why the R function defaults to Holm's method.	TRUE [✓ FALSE]
2.3 Hypothesis set B could contain exactly 1 true and 5 false null hypotheses.	[✓ TRUE] FALSE
2.4 In hypothesis set A for model 1, the contrasts for Hypotheses H_{12}: $\gamma_1 = \gamma_2$ and H_{23}: $\gamma_2 = \gamma_3$ are two orthogonal contrasts.	TRUE [✓ FALSE]
2.5 Give two orthogonal contrasts with integer weights for (a) exposed to drugs versus control, (b) glove only vs gloves plus hood. Fill in 6 integers as contrast weights for 2 contrasts.	a 1 1 -2 b 1 -1 0 Glove Hood Control

3. Use model 1 and the contrasts in question 2.5 to fill in the following anova table.

	Sum of squares	DF	Mean Square	F
Between Groups	308.6	2	154.3	56.6
Exposed versus Control	259.1	1	259.1	95.1
Gloves versus Hood	49.5	1	49.5	18.2
With Groups Residual	155.3	57	2.72	

Use models 2 and 3 for part 4.	Fill in or CIRCLE the correct answer
4.1 Which of the 16 submodels of model 2 has the smallest C_P? List the variables in this model, its size (1+#vars), the value of C_P.	Variable in this model (**list names**): exp.control and glove.hood Size= 3 C_P= 3.447
4.2 What is the PRESS value for model 2 and for model 3?	Model 2 Model 3 PRESS= 177.8 172.1
4.3 How many observations have large leverages or hatvalues in model 2? In model 3? Give two counts.	Model 2: 1 Model 3: 0 Without age and smoker, the design is balanced, constant leverage.

4.4 Give the variance inflation factors for models 2 and 3.

Put in VIFs	exp.control	glove.hood	age	smoker

Model 2	1.458	1.057	1.170	1.432
Model 3	1.000	1.000	XXXXXXXX	XXXXXXXX

DOING THE PROBLEM SET IN R

Problem Set 3, Fall 2014, Statistics 500, Answers

1.1

```
> summary(aov(tailmoment~group))
            Df Sum Sq Mean Sq F value   Pr(>F)
group        2  308.6  154.29   56.62 2.86e-14 ***
Residuals   57  155.3    2.72
```

1.2

```
> TukeyHSD(aov(tailmoment~group))
Tukey multiple comparisons of means 95% family-wise confidence
level
Fit: aov(formula = tailmoment ~ group)
$group
                 diff       lwr        upr    p adj
hood-glove    -2.2250 -3.481165 -0.9688347 2.24e-04
control-glove -5.5205 -6.776665 -4.2643347 0.00e+00
control-hood  -3.2955 -4.551665 -2.0393347 1.00e-07
```

1.4

```
> qt(.025,57)
[1] -2.002465
> qt(.025,38)
[1] -2.024394
```

1.5 (Holm method is the default)

```
> pairwise.t.test(tailmoment,group)
        Pairwise comparisons using t tests with pooled SD
        glove   hood
hood    7.7e-05 -
control 1.4e-14 8.8e-08
P value adjustment method: holm
```

3.

```
> exp.control<-c(1,1,-2)
> glove.hood<-c(1,-1,0)
> contrasts(group)<-cbind(exp.control,glove.hood)
> contrasts(group)
        exp.control glove.hood
glove             1          1
hood              1         -1
control          -2          0
> mm<-model.matrix(aov(tailmoment~group))
> head(mm)
  (Intercept) groupexp.control groupglove.hood
1           1                1               1
2           1                1              -1
3           1               -2               0
> is.data.frame(mm)
[1] FALSE
```

Problem Set 3, Fall 2014, Statistics 500, Answers continued

```
> mm<-as.data.frame(mm)
> attach(mm)
> anova(lm(tailmoment~groupexp.control+groupglove.hood))
                 Df  Sum Sq Mean Sq F value    Pr(>F)    
groupexp.control  1 259.073 259.073  95.076 9.408e-14 ***
groupglove.hood   1  49.506  49.506  18.168 7.678e-05 ***
Residuals        57 155.320   2.725                      
```

Because contrasts are orthogonal, order does not matter:

```
> anova(lm(tailmoment~groupglove.hood+groupexp.control))
                 Df  Sum Sq Mean Sq F value    Pr(>F)    
groupglove.hood   1  49.506  49.506  18.168 7.678e-05 ***
groupexp.control  1 259.073 259.073  95.076 9.408e-14 ***
Residuals        57 155.320   2.725                      
> library(leaps)
> x<-cbind(groupexp.control,groupglove.hood,age,smoke)
> mod<-leaps(x=x,y=tailmoment,names=colnames(x))
> cbind(mod$which,mod$size,mod$Cp)
  groupexp.control groupglove.hood age smoke           
1                0               0   1     0 2 114.906221
2                1               1   0     0 3   3.446797
2                1               0   1     0 3  17.478437
2                1               0   0     1 3  20.867206
2                0               1   1     0 3  96.002119
3                1               1   1     0 4   4.204977
3                1               1   0     1 4   4.556732
4                1               1   1     1 5   5.000000
> modfull<-
lm(tailmoment~groupexp.control+groupglove.hood+age+smoke)
> library(DAAG)
> press(modfull)
[1] 177.8312
> press(lm(tailmoment~groupexp.control+groupglove.hood))
[1] 172.0993
> vif(modfull)
groupexp.control  groupglove.hood              age 
smoke 
          1.4584           1.0568           1.1699 
1.4316 
> vif(lm(tailmoment~groupexp.control+groupglove.hood))
groupexp.control  groupglove.hood 
               1                1 
> summary(hatvalues(modfull))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.05007 0.07058 0.08064 0.08333 0.09565 0.17330 
> 2*0.08333
[1] 0.16666
> summary(hatvalues(lm(tailmoment~groupexp.control+groupglove.hood)))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.05    0.05    0.05    0.05    0.05    0.05 
```