

Name: \_\_\_\_\_

Check one: Section 1 (Mon-Wed. 10:30-noon): \_\_\_\_\_

Section 2 (Mon-Wed. 1:30-3:00): \_\_\_\_\_

## Statistics 431 Final Exam

May 4, 2004 4-6pm

Prof. Zanutto

- This exam is closed book.
- You may use a calculator.
- You must write the exam using pen (not pencil).
- A table of formulas and statistical tables have been provided for you at the end of this exam.
- Show all your work.
- Be sure to state the null hypothesis, test statistic, conclusion, interpretation, etc for any statistical test that you use.
- Check assumptions only when I ask.

Question	Total Points	Points Received
1	10	
2	22	
3	6	
4	8	
5	15	
6	4	
Total	65	

1) [ 2 points each] The following is a list of some of the statistical methods that you have learned about in this course.

- One sample t/Z-test of a population mean
- One sample Z-test of a population proportion
- Two sample t/Z-test of population means
- Two sample Z-test of population proportions
- Matched pairs t/Z-test.
- One-way ANOVA
- Two-way ANOVA
- Randomized Block Design ANOVA
- Simple linear regression
- Multiple Linear Regression

For each of the situations described below, state the technique (from the list above) that you believe is appropriate. **If none are appropriate, state “none of the above”.** **No calculations are required. If more than one procedure is appropriate, pick the simplest one.**

- a) A nutritionist claims that the proportion of females who consume too much saturated fat is lower than the proportion of males who consume too much saturated fat. As evidence, she reports the results of her study where she interviewed 950 randomly selected people (525 women and 425 men) and asked them about how many grams of saturated fat they consumed, on average per day.
- b) The insurance Institute for Highway Safety conducts experiments in which cars are crashed into a fixed barrier at 40mph. The impact of this crash on the crash test dummy is measured in terms of the chest compression (in mm) of the dummy. The test is conducted for a sample of large family cars, passenger vans, and midsize utility vehicles. You want to know if average chest compression differs among these three types of vehicles.
- c) The Gallup poll conducted a series of surveys to try to find out if the percentage of Americans who smoke is declining. In particular, they conducted a survey in November 1990 asking 1028 adults about their smoking habits. The question of most interest was whether they had smoked at least one cigarette in the past week. They repeated this survey to a separate random sample of 1028 adults in November 2000, asking the same questions.

- d)** Two drugs, A and B, used for the treatment of glaucoma (an eye disease) were tested for effectiveness on 10 diseased dogs. Drug A was administered to one eye (chosen randomly) of each dog and drug B was administered to the other eye. Eye pressure measurements were taken 1 hour later on both eyes of each dog. The larger the pressure measurement the more serious the eye disease.
- e)** A car owner wants to estimate the depreciation on his Chevy Camaro. To gather some data, he surveys used-car dealers and gets data on the selling price and age of 15 used Chevy Camaros (of various ages) in his local area. He wants to use this data to predict the value of his Chevy Camaro based on its age.

2) [1 point each] Multiple Choice. Circle ONE answer (the best answer) for each question below.

1) Under which of the following circumstances is it impossible to construct a confidence interval for the population mean using the methods from this course?

- a) A non-normal population with a large sample and an unknown population variance
- b) A normal population with a large sample and a known population variance
- c) Non-normal population with a small sample and an unknown population variance
- d) A normal population with a small sample and an unknown population variance

2) The lower limit of a confidence interval at the 95% level of confidence for the population proportion if a sample of size 200 had 40 successes is:

- a) 0.2554
- b) 0.1446
- c) 0.2465
- d) 0.554

3) A 99% confidence interval for the mean  $\mu$  of a normal population when the standard deviation  $\sigma$  is known is found to be 98.6 to 118.4. If the confidence level is reduced to .95, the confidence interval for  $\mu$

- a) becomes wider
- b) becomes narrower
- c) remains unchanged
- d) None of the above answers are correct.

4) Suppose that an investigator believes that virtually all values in the population are between 38 and 70. The appropriate sample size for estimating the true population mean  $\mu$  within 2 units with 95% confidence level is approximately

- a) 61
- b) 62
- c) 15
- d) 16
- e) None of the above answers are correct.

5) In testing the hypotheses

$$H_0 : p = 0.40$$

$$H_1 : p > 0.40$$

at the 5% significance level, if the sample proportion is .45 and the sample size is 100, the appropriate conclusion would be:

- a) to reject  $H_0$
- b) not to reject  $H_0$
- c) to reject  $H_1$
- d) to reject both  $H_0$  and  $H_1$

- 6) Suppose that a  $t$  test of  $H_o : \mu = 250$  versus  $H_a : \mu \neq 250$  is based on 12 degrees of freedom. If the calculated value of the test statistic is 2.8, then the  $P$ -value is
- .008
  - .992
  - .016
  - .492
  - .496
- 7) The Student  $t$  distribution approaches the normal distribution as the:
- degrees of freedom increase
  - degrees of freedom decrease
  - sample size decreases
  - population size increases
- 8) For testing the difference between two population proportions, the pooled proportion estimate should be used to compute the value of the test statistic when the:
- populations are normally distributed
  - sample sizes are small
  - samples are independently drawn from the populations
  - null hypothesis states that the two population proportions are equal
- 9) When the effect of a level for one factor depends on which level of another factor is present, the most appropriate ANOVA design to use in this situation is the:
- matched pairs design
  - one-way ANOVA
  - two-way ANOVA
  - randomized block design
- 10) The  $F$ -statistic in a one-way ANOVA represents the variation:
- between the treatments plus the variation within the treatments
  - within the treatments minus the variation between the treatments
  - between the treatments divided by the variation within the treatments
  - variation within the treatments divided by the variation between the treatments
- 11) In single-factor analysis of variance, if large differences exist among the sample means, it is then reasonable to
- reject the null hypothesis
  - reject the alternative hypothesis
  - fail to reject the null hypothesis
  - none of the above is correct

- 12) The randomized block design with exactly two treatments is equivalent to a two-tail:
- independent samples  $z$ -test
  - independent samples equal-variances  $t$ -test
  - independent samples unequal-variances  $t$ -test
  - matched pairs  $t$ -test
- 13) A professor of statistics in Wayne State University wants to determine whether the average starting salaries among graduates of the 15 universities in Michigan are equal. A sample of 25 recent graduates from each university was randomly taken. The appropriate critical value for the ANOVA test is obtained from the  $F$ -distribution with degrees of freedom equal:
- 15 and 25
  - 14 and 360
  - 360 and 14
  - 14 and 375
- 14) A randomized block design with 4 treatments and 5 blocks produced the following sum of squares values:  $SS_{Total} = 1951$ ,  $SSTr = 349$ ,  $SSE = 188$ . The value of  $SSB$  must be:
- 1414
  - 537
  - 1763
  - 1602
- 15) In order to estimate with 95% confidence the average value of  $y$  in a simple linear regression problem, a random sample of 10 observations is taken. Which of the following  $t$ -table values listed below would be used?
- 2.228
  - 2.306
  - 1.860
  - 2.262
- 16) In simple linear regression, the coefficient of correlation  $r$  (measuring the correlation between  $y$  and  $x$ ) and the least squares estimate  $\hat{\beta}_1$  of the population slope  $\beta_1$ :
- must be equal
  - must have opposite signs
  - must have the same sign
  - may have opposite signs or the same sign
- 17) Which of the following statements are not true if  $y = -3x + 7$  ?
- The  $y$ -intercept is 7
  - $y$  decreases by 3 when  $x$  increases by 4
  - $y$  decreases by 3 when  $x$  increases by 1
  - The slope of the line is -3
  - All of the above statements are not true.

- 18) In order to test the validity of a multiple regression model involving 5 independent (X) variables and 30 observations, the numerator and denominator degrees of freedom (respectively) for the overall F-test are:
- a) 5 and 30
  - b) 6 and 29
  - c) 5 and 24
  - d) 4 and 25
- 19) In multiple regression analysis involving 10 independent (X) variables and 100 observations, the critical value of  $t$  for testing individual coefficients in the model will have:
- a) 99 degrees of freedom
  - b) 10 degrees of freedom
  - c) 89 degrees of freedom
  - d) 9 degrees of freedom
- 20) If  $R^2$  is 0.975 in a simple linear regression, then the slope of the regression line:
- a) must be positive
  - b) must be negative
  - c) could be either positive or negative
  - d) None of the above answers is correct
- 21) In a simple linear regression predicting Y from X, when testing  $H_0 : \beta_1 = 0$  versus  $H_a : \beta_1 \neq 0$ , using a sample of 22 observations, the test statistic value is found to be  $t = -2.528$ . The approximate P-value of the test is
- a) .01
  - b) .02
  - c) .025
  - d) .05
  - e) .99
- 22) In general, to represent a categorical independent (X) variable that has  $m$  possible categories in a linear regression model, we must create:
- a)  $(m + 1)$  dummy variables
  - b)  $m$  dummy variables
  - c)  $(m - 1)$  dummy variables
  - d)  $(m - 2)$  dummy variables

3) Every April, Americans fill out their tax return forms. Many turn to tax preparation companies to do this tedious job. The question arises, “Are there differences between companies?” In an experiment, two of the largest companies were asked to prepare the tax returns of a sample of 55 taxpayers. The amounts of tax people owed were recorded and analyzes (this sample deals only with people who owed taxes, no one was due a refund). Descriptive statistics for this data is shown below (Diff= taxes calculated by company 1 – taxes calculated by company 2).

Distributions				
Company 1				
	Quantiles		Moments	
	100.0% maximum	13944	Mean	9073.8545
99.5%	13944	Std Dev	2276.2373	
97.5%	13858	Std Err Mean	306.92777	
90.0%	11940	upper 95% Mean	9689.2077	
75.0% quartile	10501	low er 95% Mean	8458.5014	
50.0% median	9287	N	55	
25.0% quartile	7291			
10.0%	5994			
2.5%	4427			
0.5%	4051			
0.0% minimum	4051			
Company 2				
	Quantiles		Moments	
	100.0% maximum	15481	Mean	8553
99.5%	15481	Std Dev	2853.5132	
97.5%	15449	Std Err Mean	384.76764	
90.0%	13092	upper 95% Mean	9324.4127	
75.0% quartile	10150	low er 95% Mean	7781.5873	
50.0% median	8422	N	55	
25.0% quartile	7170			
10.0%	4943			
2.5%	2193			
0.5%	1069			
0.0% minimum	1069			
Diff				
	Quantiles		Moments	
	100.0% maximum	5206	Mean	520.85455
99.5%	5206	Std Dev	1854.9234	
97.5%	4692	Std Err Mean	250.11783	
90.0%	3105	upper 95% Mean	1022.3106	
75.0% quartile	1634	low er 95% Mean	19.398497	
50.0% median	744	N	55	
25.0% quartile	-972			
10.0%	-1652			
2.5%	-3677			
0.5%	-3862			
0.0% minimum	-3862			

- a) [3 points] Can we conclude that there is a difference between the two companies? Use a 2-sided test. Check any necessary assumptions.

- b) [3 points] Suppose for a different analysis, using just data from Company 1 (ignoring data from Company 2), you wanted to test whether the average amount of tax owed is less than \$10,000. Using the sample data from Company 1, how much power would you have to reject  $H_0 : \mu = 10,000$  in favor of  $H_A : \mu < 10,000$  if the true average amount of tax owed is \$9,000? (use  $\alpha = 0.05$ )

- 4) A study investigated whether a practice exam helps students prepare for a final exam. In this experiment, students in an introductory psychology class at Penn State were initially divided into 3 groups based on their class standing (based on their most recent tests and homework): Low, Medium, High. Within each group, students were randomly assigned to either attend a review session or take a practice exam prior to the final exam. There were 22 students in each of the  $2 \times 3 = 6$  treatment groups. After completing the final exam, each student rated their exam preparation on an 11-point scale ranging from 0 (not helpful at all) to 10 (extremely helpful). These data are analyzed on the next page.

Prep=practice exam or review session  
Standing=low, medium, high

a) [3 points] Does the effect of the preparation (practice exam or review session) depend on a student's standing in the class?

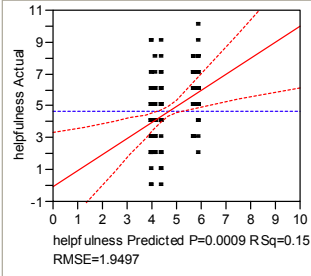
b) [4 points] Is there a difference in average helpfulness (as rated by the students) between the practice exam and the review session? If yes, which is more helpful? If you can't answer this question, say why.

c) [1 point] This design allows us to find out if different preparations (review session or practice exam) work differently for different types of students. Suppose we could only administer one preparation (either everyone gets the practice exam or everyone gets the review session), so we wanted to know which preparation was the best overall. Would you change the analysis on the next page? How?

**Response helpfulness**

**Whole Model**

**Actual by Predicted Plot**



**Summary of Fit**

RSquare	0.150276
RSquare Adj	0.116557
Root Mean Square Error	1.949673
Mean of Response	4.659091
Observations (or Sum Wgts)	132

**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	84.70455	16.9409	4.4567
Error	126	478.95455	3.8012	Prob > F
C. Total	131	563.65909		0.0009

**Parameter Estimates**

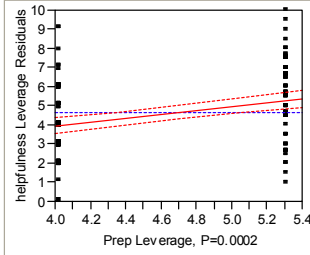
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	4.6590909	0.169697	27.46	<.0001
Prep[Practice]	0.6439394	0.169697	3.79	0.0002
standing[High]	0.25	0.239988	1.04	0.2995
standing[Low]	0.25	0.239988	1.04	0.2995
Prep[Practice]*standing[High]	0.3106061	0.239988	1.29	0.1979
Prep[Practice]*standing[Low]	0.1287879	0.239988	0.54	0.5925

**Effect Tests**

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Prep	1	1	54.734848	14.3993	0.0002
standing	2	2	16.500000	2.1704	0.1184
Prep*standing	2	2	13.469697	1.7718	0.1742

**Prep**

**Leverage Plot**



**Least Squares Means Table**

Level	Least Sq Mean	Std Error	Mean
Practice	5.3030303	0.23998821	5.30303
Review	4.0151515	0.23998821	4.01515

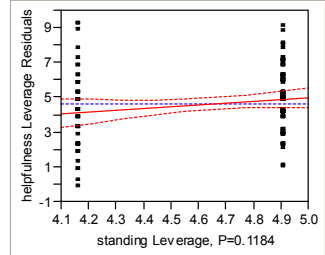
**LSMeans Differences Tukey HSD**

Alpha= 0.050 Q= 1.97897

		LSMean[j]	
Mean[i]-Mean[j]	Std Err Dif	Practice	Review
Practice	0	1.28788	0
	Lower CL Dif	0.33939	0
	Upper CL Dif	0.61623	0
Review	0	-1.2879	0
	Lower CL Dif	-0.33939	0
	Upper CL Dif	-0.6162	0

**standing**

**Leverage Plot**



**Least Squares Means Table**

Level	Least Sq Mean	Std Error	Mean
High	4.9090909	0.29392433	4.90909
Low	4.9090909	0.29392433	4.90909
Medium	4.1590909	0.29392433	4.15909

**LSMeans Differences Tukey HSD**

Alpha= 0.050 Q= 2.37176

		LSMean[j]		
Mean[i]-Mean[j]	Std Err Dif	High	Low	Medium
High	0	0	0	0.75
	Lower CL Dif	0.41567	0.41567	0
	Upper CL Dif	0.9859	0.9859	0
Low	0	0	0	0.75
	Lower CL Dif	0.41567	0.41567	0
	Upper CL Dif	0.98588	0.98588	0
Medium	-0.75	-0.75	-0.75	0
	Lower CL Dif	-1.7359	-1.7359	0
	Upper CL Dif	-0.23588	-0.23588	0

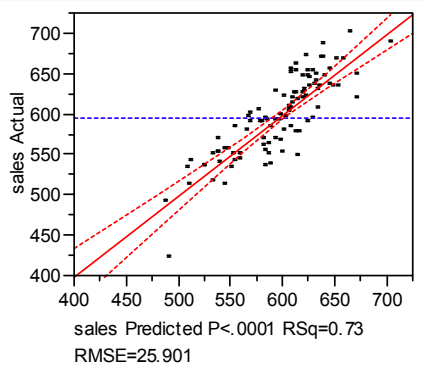
- 5) Market research was conducted for a national retail company to compare the relationship between sales and advertising expenditures during the warm Spring and Summer seasons (season=0) as compared with the cool Fall and Winter seasons (season=1). The data analyzed below were collected over 50 (!) years.
- a) [3 points] Three regression models were fitted to this data. Find the model that specifies parallel lines for warm season and cool season sales as a function of advertising expenditures (ad.exp=millions of dollars). Write out the estimated regression line for the warm seasons, and separately, the estimated regression line for cool seasons.
- b) [2 points] Does a model that allows for different slopes for warm and cool seasons fit the data better? Report and interpret the appropriate test (give null hypothesis, test statistic, distribution, interpretation, etc.)

- c) [1 point] Interpret the  $R^2$  for model #2.
- d) [2 points] Interpret the coefficient of ad.exp for model #2.
- e) [2 points] Comment on the residual plot for model #2.
- f) [2 points] What other diagnostics would you want to see for model #2? Give a complete list.
- g) [3 points] Calculate, by hand, an approximate 95% prediction interval for the sales in an individual warm season when advertising expenditures are 10 (million dollars) using Model #2.

## Model #1

### Whole Model

#### Actual by Predicted Plot



#### Summary of Fit

RSquare	0.73216
RSquare Adj	0.72379
Root Mean Square Error	25.90073
Mean of Response	596.688
Observations (or Sum Wgts)	100

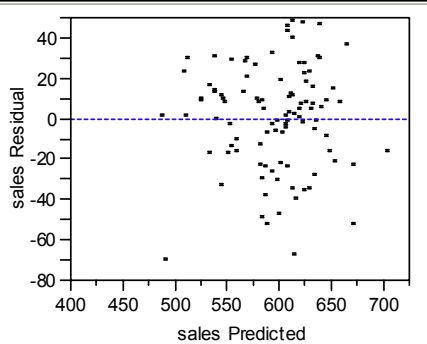
#### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	176046.16	58682.1	87.4745
Error	96	64401.38	670.8	Prob > F
C. Total	99	240447.55		<.0001

#### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	27.370564	46.31393	0.59	0.5559
season	110.24079	64.84197	1.70	0.0923
ad.exp	3.9302748	0.322864	12.17	<.0001
season*ad.exp	-0.704345	0.449597	-1.57	0.1205

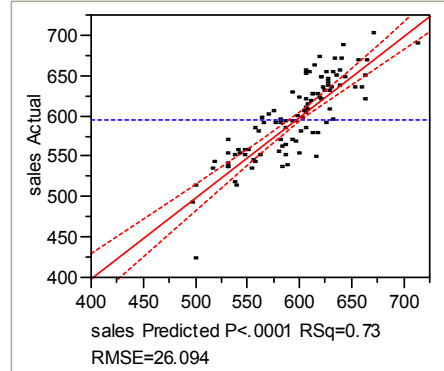
#### Residual by Predicted Plot



## Model #2

### Whole Model

#### Actual by Predicted Plot



#### Summary of Fit

RSquare	0.725313
RSquare Adj	0.719649
Root Mean Square Error	26.09417
Mean of Response	596.688
Observations (or Sum Wgts)	100

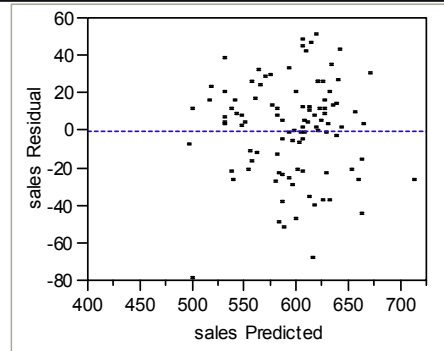
#### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	174399.71	87199.9	128.0646
Error	97	66047.83	680.9	Prob > F
C. Total	99	240447.55		<.0001

#### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	79.311186	32.57952	2.43	0.0167
season	8.9845996	5.230951	1.72	0.0891
ad.exp	3.5670486	0.226366	15.76	<.0001

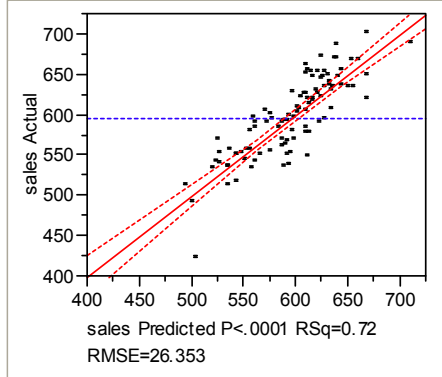
#### Residual by Predicted Plot



### Model #3

#### Whole Model

##### Actual by Predicted Plot



##### Summary of Fit

RSquare	0.716959
RSquare Adj	0.714071
Root Mean Square Error	26.35251
Mean of Response	596.688
Observations (or Sum Wgts)	100

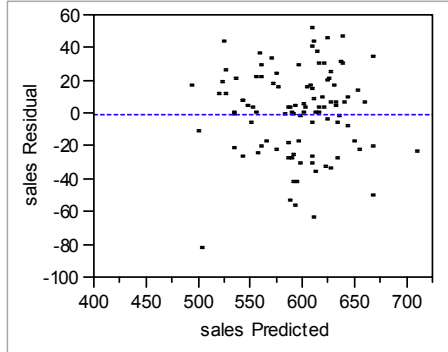
##### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	172390.98	172391	248.2393
Error	98	68056.57	694	Prob > F
C. Total	99	240447.55		<.0001

##### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	80.000523	32.89958	2.43	0.0168
ad.exp	3.5934977	0.228077	15.76	<.0001

##### Residual by Predicted Plot



- 6) [ 4 points] Suppose you have a sample of  $n$  observations, and you want to apply the following regression model to these data

$$y_i = \beta_1 x_i + \varepsilon_i$$

where  $\varepsilon$  is random error term with mean zero and variance  $\sigma^2$ . (Notice that there is no intercept coefficient in this model).

Recall that the least squares regression estimates minimize the sum of squared errors

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Derive the least squares estimate of  $\beta_1$  for this regression model (e.g. derive a formula that tells us how to calculate  $\hat{\beta}_1$  from the  $n$  observations).