Assumptions in Regression Modeling

Preliminaries

Preparing for class…

– Read the casebook prior to class
…Pace in class is too fast to absorb without some prior reading

– Identify questions, find places where you did not follow the analysis.

– Focus on *managerial questions* that motivate and guide the analysis.

– What conclusion should be drawn from the analysis?
Under what conditions?

Once in class…

– Take notes directly in the casebook rather than redrawing figures.

– Participate in the discussion.

– Everything will not make sense the first time you see it!

Feedback

– Please fill out feedback forms, based on last digit of student ID number.

– Contact either
- Quality circle for the class
- Cohort academic reps (please identify yourselves)

Office hours

– "After class" on Monday and Wednesday (3-5 p.m.)
3014 SH-DH in the Statistics Department.

– Check e-mail in the evening; use the FAQ from my web page
www-stat.wharton.upenn.edu/~bob/stat621

Teaching assistants

– Room 3009 SH-DH (as in Stat 603/604)

– Hours for TAs appear in the course syllabus

Discussion of Assignment #1

Due on Friday

– Drop off printed copy in Statistics Department.

– One copy per team.

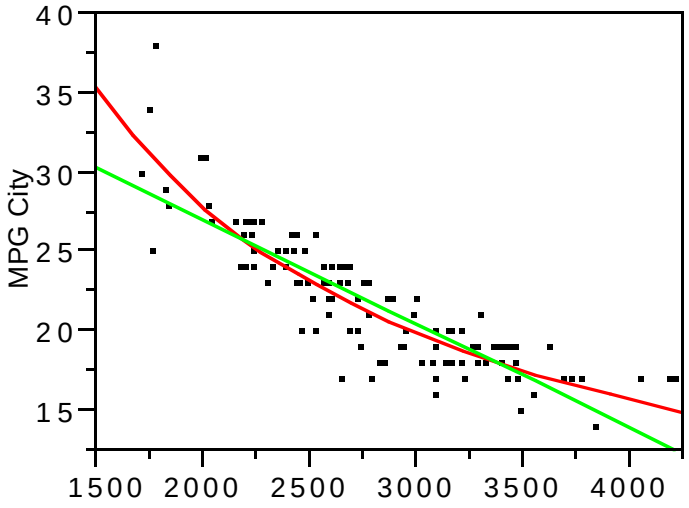
– Do *not* send electronically.

Questions about the assignment

– Get the data from the web.

– Question #1: Regression estimates of the *beta* of a stock.
Subset the data either by manual selection of the rows and using the exclude rows, or by using the tables menu to build a new data set.

– Question #2: Transformations and fuel consumption.
Use a single plot with three models for the data. For example, this plot compares a log-log model (red) to the linear model (green).



In the assignment, discuss which model is best in the context of some question that can be answered with any of the models, such as prediction for cars of some weight.

Discussion from Class 1 (Fitting Equations)

Key points

– Smoothing (using splines) produces a summary of how the *average* of the response (Y) depends upon the predictor (X).

– Regression produces an equation for this summary.
The equation gives the average (or "expected") value of the response at a value of a predictor.

– The equation can be linear (easy interpretation) or nonlinear (harder).

Interpreting a linear regression equation

– Intercept and slope have units (unlike a correlation)
Intercept takes units from response, slope from response/predictor.

– Dangers of extrapolating outside the range of data:
e.g., prices for very small or very large diamonds

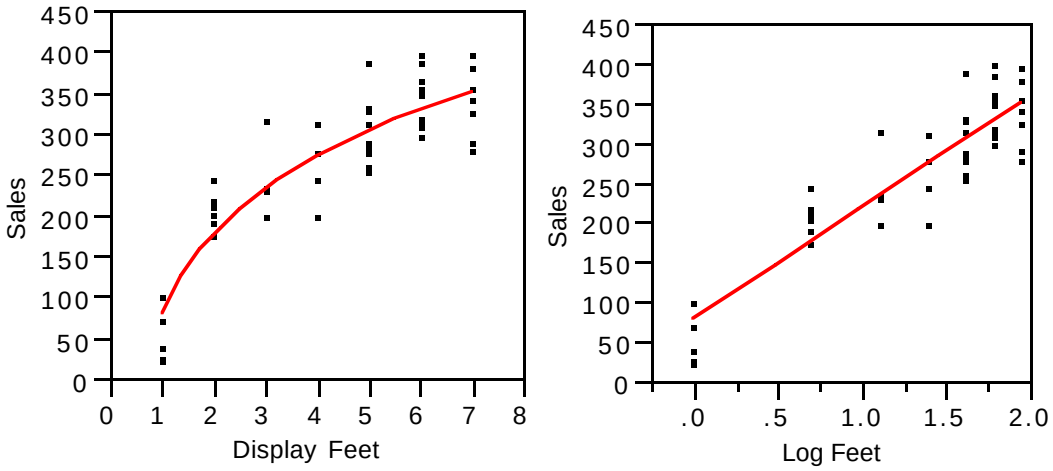
Transformations

– Intercept and slope when variables are transformed
• Heuristic for logs and percentage change (pages 19-22)

– Transform original measurements to new scales (e.g. log or reciprocal)

– Understanding curvature in data (e.g., diminishing returns to scale)

– *Subtle*: Curved model on original scales is linear on new scales. (p 32)
Note the different spacing of values on the horizontal axis.



JMP-IN Notes

Fit Y by X view

– Behavior changes when have "categorical" or "continuous" predictor
  • categorical – you get a t-test when you have two groups (Stat603)
  • continuous – you get a regression modeling tool

– Options for this view include the ability to fit models with
  • transformations ("Fit special")
  • polynomials ("Fit polynomial")
  • smooth curves ("Splines")

– Get residual plot in this view by using the button associated with the specific fitted model (each fitted model in the view has its own residual plot).

Subsets

– Manual selection of desired observations, then use "Exclude" from the "Rows" menu.

– Use the subset command from the "Tables" menu.

New Application for Today

Providing a margin for error

– Confidence interval for optimal shelf space?

– What is the probable accuracy of a forecast or prediction?

Confidence intervals provide the best answers to these types questions. To get these from regression, we need to embellish the regression equation with assumptions that complete the statistical model.

Regression Model

Four key elements of the standard regression model

0. *Equation*, stated either in terms of the average of the response

$$ave(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$

or for the values of the response as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

1. *Independent* observations (independent errors)

2. *Constant variance* observations (equal error variance σ^2)

3. *Normally distributed around* the regression line (errors ε_i are $N(0, \sigma^2)$)

Standard model is a utopian model

– Idealized "data generating process" or model.

– Closer to this ideal model, the more reliable inference becomes.

– Model indexed by 3 unknown parameters that define the specific characteristics of the process:

Intercept β_0, slope β_1, and SD of errors σ.

Supporting Concepts and Terminology

Discussion: "What are the errors in a regression model?"

– Represent the net effect of omitted factors.

– Model is a statement only about averages

$$ave(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$

and the errors represent the deviations from these averages.

– Errors combine all of the omitted factors that determine the response.

Problem with checking the assumptions

– Assumptions 1-3 are statements about the errors, but

– The errors are not directly observed.

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

Residual

– Estimate or proxy for the unobservable error term

– Observed vertical deviation from a fitted line

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

– RMSE "root mean squared error" is SD of the residuals

Transformations and residuals

– Assumptions presume you know how to transform data.

– Check the residuals on the associated transformed scales.

Least squares

"Best" fitted line is chosen to minimize the sum of squared residuals.

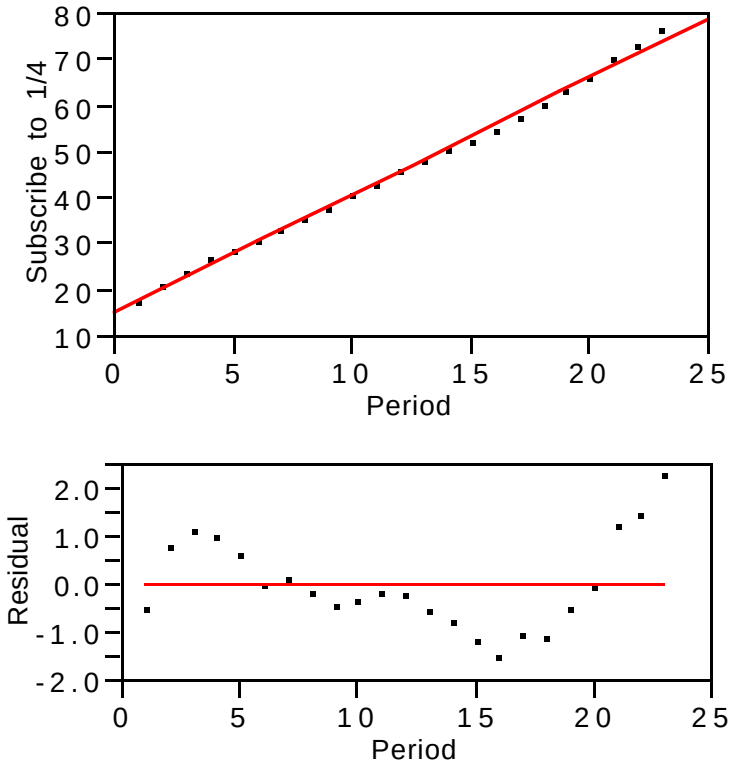
Assumption	**Diagnostic(graphical)**	**Problem**
Linearity	Scatterplots(data/residuals)	nonlinear
Independence	Plots showing trend	autocorrelation
Constant variance	Residual plots	heteroscedasticity
Normality	Quantile plots	outliers, skewness

Autocorrelation

– Occurs most clearly with time series data, in which case the residuals appear to "track" over time

– Cellular case study (various places in case book, p. 29, 53, 304)
  • Transformed response (uninterpretable $Y^{1/4}$)

– Linear model (on transformed scales) and residuals shown on next…



Residuals and transformation

– Subtle point: Residual plots are more useful if you plot the data *after* applying the transformation to the column, rather than looking at the data on the "raw scales"

– The view of the fitted regression model on the original scale stresses the qualitative features such as diminishing returns to scale.

– However, when looking for problems in the residuals, it's often best to look at the data on scales that make the equation linear. The assumptions live on the linear scale.

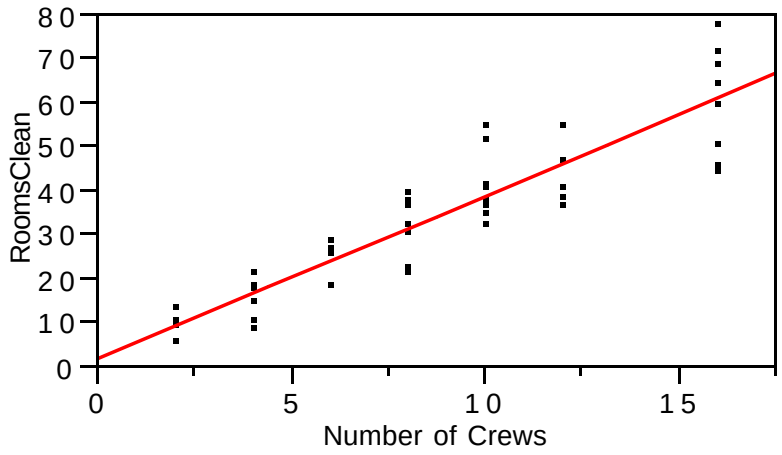
Heteroscedasticity

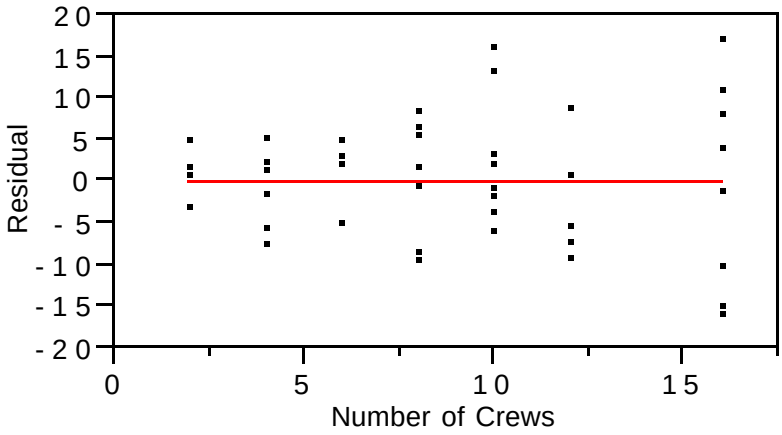
– Lack of constant variance implies not using data efficiently.

– Shows up as residuals that do not have constant variation.

– Remedied by a weighted analysis (we'll not explore this in class)

– Room cleaning example (p. 7, 57)





Outliers (unusually small/large values of X or Y)

– Distinguish what makes an observation unusual:
Unusual in X = **leveraged**
Unusual in Y = residual of large magnitude (negative or positive)

– Large absolute residual + leveraged location –> **influential** observation.

– Least squares fit is VERY sensitive to influential observations.

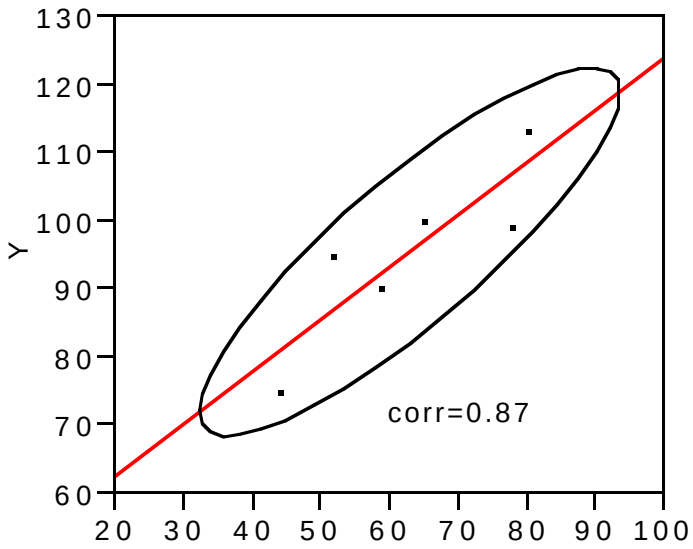
– Goal: don't just remove, understand why a point is unusual.

JMP script illustrating leverage, outliers

– See the impact of moving points

– Data ellipse quantifies the size of the correlation
(more on the correlation next time)

– Squared error of large residuals dominate least squares procedure



Examples for Today (and into next class)

Lots of examples to illustrate issues

Utopian data	What happens when all goes well?
Cellular phones	Dependence apparent in residuals.
Cleaning crews	Variance changes for larger operations.
Philadelphia housing	Unusual points and their effects.
Building profits	" "

The ideal regression model **Utopia.jmp, page 47**

"What do the plots look like if everything is working properly?"

– Visual training, getting comfortable with the graphical displays.

– Simulated data for which we know the "true" model.

– Note the interpretation of RMSE (p 48)

Discussed above

Predicting cellular phone use **Cellular.jmp, page 57**

"How many subscribers do you predict by the end of this year?"

– Nonlinear on original scale; linear on the transformed scale. (p 54)

– RMSE of fit is small compared to scale of response (p 54)

– Fit appears quite good, until one "zooms in" with residual plots. (p 55)

– Residuals show autocorrelation, tracking over time.

– Predictions will likely be too small due to local trend at end of data.

– Revisited in Class 12 (see also Class 1, p 29-37)

Discussed above

Efficiency of cleaning crews **Cleaning.jmp, page 57**

"How well can we predict the number of rooms cleaned?"

– Variability in residuals increases with size, heteroscedastic (p 58)

– Not a big deal for estimating the slope, but

– Lack of constant variance is very important when assessing predictions.

Philadelphia housing prices **Phila.jmp, page 62**

"How do crime rates impact the average selling price of houses?"

– Initial plot shows that Center City is "leveraged" (unusual in X).

– Initial analysis with all data finds $577 impact per crime (p 64).

– Residuals show lack of normality (p 65).

– Without CC, regression has much steeper decay, $2289/crime (p 66).

– Residuals remain non-normal (p 67).

– Why is CC an outlier? What do we learn from this point?

– Alternative analysis with transformation suggests may be not so unusual. (see pages 68-70)

Housing construction **Cottages.jmp, page 78**

"Should the company pursue the market for larger homes?"

– Whole analysis hinges on how we treat one very large "cottage".

– With this large cottage, find a steep, positive slope. (p 80)
… as though we only have two "points".

– Without this cottage, we find a much less clear relationship.

– Continue the analysis of this data in Class 3 on Thursday and discover the impact of this single point on prediction and the statistical properties of our model.

Key Take-Away Points

Regression model adds assumptions to the equation

– Independent observations

– Constant variance around regression line

– Ideally, normally distributed around regression line

Residuals provide the key to checking the assumptions

– Poor equation: Zoom in with the plot of the residuals

– Dependence: Plots of residuals over "time"

– Constant variance: Plots of residuals versus predictor

– Normality: Quantile plots of residuals

One outlier can exert much influence on a regression

– Ideas of leverage and influence

– Difficult choice of what to do about an outlier

Next Time

Build on the assumptions we have developed today.

– Confidence intervals

– Prediction intervals

– General inference (making statements about the underlying population)