Plan for Today

- Issues from Problem Set 9
- Lecture 12 Review
- Discuss Readings
- Problem Set 10
Issues from Problem Set 9

• Significant Digits!
• Interpreting results in plain English
  • Make sure you look at the units of the variables you’re using
  • Slope is change in y over change in x (When x goes up by 1, y goes up by b)
  • Intercept is value of y when x is zero
• Interpreting meaning of standard error
  • Standard error of the residuals tells you about the spread of the observed values (y) compared to the expected values (y-hat, i.e. the values on the best-fit line)
  • Since we know (assume) the residuals are normally distributed, we can say that 68% of the observed values lie within one standard error (plus or minus) of the expected values (the line of best fit)
  • NORM.S.DIST(1,1) – NORM.S.DIST(-1,1) = 0.68
Lecture 12 Review: Regression Analysis 2

- Prediction and Standard Error
- Analysis of Residuals
  - Assumptions
  - Outliers
  - Non-normality
  - Non-linearity
  - Heteroscedasticity
  - Autocorrelation
- Transformations
  - Linear
  - Exponential
  - Power
  - Logarithmic
Prediction and Standard Error

• The value of y predicted by the best-fit line for a given x is
  \[ y^\hat{} = a + bx \]

• What are two reasons that this prediction is uncertain?
  • The estimated regression line isn’t the “true” regression line \( (a \neq \alpha, b \neq \beta) \); this uncertainty is reduced as the sample size, n, is increased
    • Can’t get the “true” line unless we have every observation from the full population
  • There is natural variability in y for a given value of x. We model this variability with a normal distribution with constant standard deviation \( \sigma \approx se \)
    • X and Y aren’t perfectly correlated (unless it’s a trivial case like temperature in Celsius and Fahrenheit)
Prediction and Standard Error

- The uncertainty in the mean value (y-hat) that arises from the uncertainty in \( a \) and \( b \) (the error in the location of the best-fit line) is given by the equation:

\[
s_y = s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n - 1)s_x^2}} \approx \frac{s_e}{\sqrt{n}} \quad (\text{if } x \approx \bar{x})
\]

- If we knew the exact values of \( \alpha \) and \( \beta \), the uncertainty in any individual value of \( y \) would be given by \( \sigma \approx se \) (the standard error of the residuals), regardless of the value of \( x \). The overall uncertainty is then given by the equation:

\[
s_{\hat{y}-y} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n - 1)s_x^2}} \approx s_e \quad (\text{if } x \approx \bar{x})
\]

Error grows as \( x \) moves away from the middle of the data. Extrapolation (predicting \( y \) for \( x \) outside of range of original data) is frowned upon.
Prediction and Standard Error

• What happens to the error in individual predicted value as x moves away from the middle of the data (x gets farther away from x-bar)?

\[ s_{\hat{y}-y} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n - 1)s_x^2}} \approx s_e \text{ (if } x \approx \bar{x}) \]

• The error increases.

• This is why extrapolation (predicting y for x outside the range of original data (x far from x-bar) is frowned upon.
Analysis of Residuals: Assumptions

• Least-squares regression assumes residuals are normally distributed and that they are independent, random variables.

• What do these assumptions mean for the distribution of the residuals?

• Normally Distributed
  • mean = 0 regardless of x
  • stdev = σ regardless of x
  • Normally distributed: P(e > 2σ) = 0.05, P(e > 3σ) = 0.003, etc/

• Independent, Random Variable
  • Residuals not be correlated with each other
Analysis of Residuals: Assumptions

- What problems occur if these assumptions are not met?
  - Normally Distributed
    - mean = 0 regardless of x
    - If not, you get non-linearity

  - stdev = σ regardless of x
    - If not, you get heteroscedasticity

- Normally distributed: P(e > 2σ) = 0.05, P(e > 3σ) = 0.003, etc
  - If not, you likely have outliers that distort the distribution

- Independent, Random Variable
  - Residuals not be correlated with each other
  - If not, you get autocorrelation
Analysis of Residuals: Non-Linearity

• Non-linearity occurs when which assumption isn’t met?
  • The mean of the residuals is zero (i.e., non-linearity occurs when the mean of the residuals is not zero)

• If you have non-linearity, what pattern would you expect to see in your data? In a graph of the residuals?
  • The data are curved with respect to the line of best fit
  • The data are curved with respect to the x-axis
Analysis of Residuals: Non-Linearity

- What are two ways to test for non-linearity?
  - Visual inspection (usually sufficient)
  - Add a polynomial (quadratic) trendline to the residual plot; if $R^2$ is large (> 4/n), then curvature is significant, and you should consider a non-linear transformation.
Analysis of Residuals: Heteroscedasticity

- Heteroscedasticity occurs when which assumption isn’t met?
  - The standard deviation of residuals, $s_e$, is constant for all x’s
  - (i.e., heteroscedasticity occurs when $s_e$ is not constant for all x’s)
- If you have heteroscedasticity, what pattern would you expect to see in your data? In a graph of the residuals?
  - The spread of the residuals should not change systematically with x.
Analysis of Residuals: Heteroscedasticity

- How can you detect heteroscedasticity? What can you do about it?
  - Visual inspection (usually sufficient)
  - Some statistical programs (e.g. STATA), use White or Cook-Weisburg tests to detect
  - Robust standard errors can be used to deal with heteroscedasticity
Analysis of Residuals: Non-normality

- Non-normality occurs when which assumption isn’t met?
  - The assumption that there are no outliers in the data
  - (i.e. non-normality occurs when there are outliers in the data)

- If you have non-normality, what pattern would you expect to see in your data? In a graph of the residuals?
  - You would see residuals outside $+/-3s_e$
Analysis of Residuals: Non-Normality

- What are two ways to test for non-normality? What can you do about it?
  - Compute and inspect standardized residuals
  - Delete the outlier and see if it affects the coefficients
    - If not, it was not an important outlier
    - If yes, discover whether there is a reasonable basis for removing the observations
  - Make sure that a histogram of the residuals is approximately bell-shaped and symmetrical
  - Carry out a formal test using chi-square analysis
Analysis of Residuals: Autocorrelation

- Autocorrelation occurs when which assumption isn’t met?
  - The assumption that residuals are random and uncorrelated
  - i.e. Autocorrelation occurs when the residuals are correlated

- If you have autocorrelation, what pattern would you expect to see in your data? In a graph of the residuals?
  - There is a regular pattern of residuals (common in time-series data)
Analysis of Residuals: Autocorrelation

- What are two ways to test for autocorrelation? What can you do about it?
  - To test for first-order autocorrelation, test for correlation between $e_t$ and $e_{t-1}$
  - To test for second-order autocorrelation, test for correlation between $e_t$ and $e_{t-2}$, etc.
  - Use the Durbin-Watson Test
    - $D \approx 2 - 2r$
    - If no autocorrelation, $D \approx 2$
      - If strong positive autocorrelation, $D \approx 0$
    - If strong negative autocorrelation, $D \approx 4$
    - Critical value of $D$ depends on $n$, $\alpha$, $k$
Transformations

• What are three reasons you might use a transformation?
  • To linearize non-linear relationships between independent and dependent variables
  • To produce residuals that are normally distributed with constant standard deviation
  • To remove autocorrelation from a time series

• What are four types of transformations?
  • Linear
  • Exponential
  • Power
  • Logarithmic
Transformations

- People learn by doing, and get better at their jobs over time. We assume people improve by the same amount at their specific job (e.g. teachers become better teachers, investment bankers become better investment bankers).

- Two people graduate college, one becomes a high school teacher and one goes to Wall Street and becomes an investment banker. After two years, they both get raises. How large do you expect these increases to be?
  - You could try to answer in absolute dollars – i.e., they’ll both get wage increases of $5,000, OR
  - You could try to answer in percentages – i.e., they’ll both get wage increases of 5%
  - Based on what we know about how raises are given out, it seems much more likely that the raise will be relative to your base pay (a percentage)
  - Therefore, in a regression of years of experience and wages, it would be better to look at the percentage changes in y (wages) for a unit change in x (years of experience) – This is a log transformation.
Transformations: Linear

- If the relationship between x and y is linear, the equation characterizing this relationship is:

  \[ y = \alpha + \beta x \]

- This relationship is already linear

- In a simple linear regression, the slope, \( \beta \), is the average change in y for a unit change in x:

  \[ \frac{\Delta y}{\Delta x} \]

- What change in y results from a unit change in x?
  - A unit change in x produces a \( \beta \) change in y

- If \( \beta = 0.1 \), how much will y increase when x increases by 1?
  - If \( \beta = 0.1 \), a unit increase in x produces a 0.1 increase in y
  - i.e., If x goes up by 1, y will go up by 0.1
Transformations: Exponential

- If the relationship between $x$ and $y$ is exponential, the equation characterizing this relationship is:

$$y = \alpha e^{\beta x}$$

- This can be linearized by taking the log of both sides

$$\log(y) = a + \beta x$$

$$a = \log(\alpha)$$
Transformations: Exponential

- In an exponential transformation (an exponential relationship that is linearized by regressing \( \log(y) \) on \( x \)), the slope, \( \beta \), is \( \frac{\Delta y}{y} / \Delta x \).

- What change in \( y \) results from a unit change in \( x \)?
  - A unit change in \( x \) produces a \( 100\beta \% \) change in \( y \).

- If \( \beta = 0.1 \), how much will \( y \) increase when \( x \) increases by 1?
  - If \( \beta = 0.1 \), a unit increase in \( x \) produces a 10% increase in \( y \).

\[
y = \alpha e^{\beta x}
\]

\[
\log(y) = a + \beta x
\]

\[
a = \log(\alpha)
\]
Transformations: Power

• If the relationship between $x$ and $y$ is a power function, the equation characterizing this relationship is:

$$y = \alpha x^\beta$$

• This can be linearized by taking the log of both sides

$$y = \alpha x^\beta \quad \log(y) = \log(\alpha) + \beta \log(x)$$
Transformations: Power

- In a power transformation (a power function relationship that is linearized by regressing log(y) on log(x)), the slope, $\beta$, is
  \[
  \frac{\Delta y}{\Delta x} = \frac{\Delta y}{y} \frac{y}{\Delta y} = \beta
  \]

- What change in $y$ results from a one percent change in $x$?
  - A one percent change in $x$ produces a $\beta\%$ change in $y$

- If $\beta=0.1$, how much will $y$ increase when $x$ increases by 1%?
  - If $\beta=0.1$, a one percent increase in $x$ produces a 0.1% increase in $y$
Transformations: Logarithmic

- If the relationship between x and y is a logarithmic, the equation characterizing this relationship is:

  \[ y = \alpha + \beta \log(x) \]

- This equation has already been linearized
Transformations: Logarithmic

- In a logarithmic relationship (a function that regresses y on log(x)), the slope, $\beta$, is $\frac{\Delta y}{\Delta x/x}$.

- What change in y results from a one percent change in x?
  - A one percent change in x produces a $0.01\beta$ change in y.

- If $\beta=0.1$, how much will y increase when x increases by 1%?
  - If $\beta=0.1$, a one percent increase in x produces a 0.001 increase in y.

$$y = \alpha + \beta \log(x)$$
# Transformation Summary

<table>
<thead>
<tr>
<th>Type</th>
<th>Explanatory</th>
<th>Response</th>
<th>Slope ( \beta )</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>x</td>
<td>y</td>
<td>( \frac{\Delta y}{\Delta x} )</td>
<td>( \beta ) change in y for 1 unit change in x</td>
</tr>
<tr>
<td>Exponential</td>
<td>x</td>
<td>( \ln(y) )</td>
<td>( \frac{\Delta y}{\Delta x} )</td>
<td>100( \beta )% change in y for 1 unit change in x</td>
</tr>
<tr>
<td>Power</td>
<td>( \ln(x) )</td>
<td>( \ln(y) )</td>
<td>( \frac{\Delta y}{\Delta x/x} )</td>
<td>( \beta )% change in y for 1% change in x</td>
</tr>
<tr>
<td>Logarithmic</td>
<td>( \ln(x) )</td>
<td>y</td>
<td>( \frac{\Delta y}{\Delta x/x} )</td>
<td>0.01( \beta ) change in y for 1% change in x</td>
</tr>
</tbody>
</table>
Excel Functions

- Scatterplot:
  - Chart/Add Trendline/Linear/Display equation, R2

- Menu-driven tools:
  - Tools/Data Analysis/Regression

- Excel functions:
  - \( a = \text{INTERCEPT}([y\text{-range}],[x\text{-range}]) \)
  - \( b = \text{SLOPE}([y\text{-range}],[x\text{-range}]) \)
  - \( se = \text{STECX}() \)
  - \( \text{LINEST}([y\text{-range}],[x\text{-range}],1,1) \) gives all of above

- Y-hat (expected y) = \( \text{FORECAST}(x,[y\text{-range}],[x\text{-range}]) \)
Readings

• Kahneman
• Chapter 20: The Illusion of Validity
• Chapter 21: Intuitions vs. Formulas
The Illusion of Validity

• “The amount of evidence and its quality do not count for much, because poor evidence can make a very good story.”

• Evaluating Soldiers for Officer Training
  • Watch them complete a “leaderless group challenge”
  • Make comments and rate whether they would do well in officer training
  • Feedback from training showed that ratings were not good predictors
  • Continued practice and felt confident in predictions
The Illusion of Validity

• Which line is longer?
  • Answer if you’ve never seen this before
  • Answer if you have seen this before

• This is a visual illusion; the book discusses cognitive illusions.
• A major difference is that with the visual illusion once you know both lines are equal length, you will give that answer in the future.
• With a cognitive illusion, you will continue to will not want to change your behavior (i.e. admit your evaluations of soldiers are not predictive of performance in officer training)
The Illusion of Validity

• Other cognitive Illusions
  • Stock-picking
    • Results are no better than chance – doing well or poorly is luck
    • However, lots of professional training and work goes into choices
    • When shown evidence of this, firm chose to ignore it
    • When we were done, one of the executives I had dined with the previous evening drove me to the airport. He told me, with a trace of defensiveness, “I have done very well for the firm and no one can take that away from me.” I smiled and said nothing. But I thought, “Well, I took it away from you this morning. If your success was due mostly to chance, how much credit are you entitled to take for it?”
  • Political Pundits
    • Being able to explain things in retrospect makes it seem like you should have been able to predict it
    • Experts did worse than non-experts – experts have an enhanced illusion of skill and become overconfident
Intuitions vs. Formulas

• Paul Meehl looked at 20 studies that analyzed whether clinical predictions based on subjective impressions of trained professionals were more accurate than statistical predictions made by combining a few scores or ratings according to a rule.

• E.g. Counselors interview students and predict their grades at the end of the school year vs. formula using high school grades and one aptitude test; formula wins.
Intuitions vs. Formulas

- Formulas do better than people in “low-validity environments” – high uncertainty and unpredictability
  - Diagnosis of cardiac disease
  - Susceptibility of babies to sudden infant death syndrome
  - Evaluation of credit risks by banks
  - Suitability of foster parents
  - Odds of recidivism among juvenile offenders
  - Winners of football games
  - Future prices of Bordeaux wine
Intuitions vs. Formulas

• Why does this happen?
  • “Experts try to be clever, think outside the box, and consider complex combinations of features in making their predictions. Complexity may work in the odd case, but more often than not it reduces validity.”
  • “When asked to evaluate the same information twice, they frequently give different answers.
    • The extent of the inconsistency is often a matter of real concern. Experienced radiologists who evaluate chest X-rays as “normal” or “abnormal” contradict themselves 20% of the time when they see the same picture on separate occasions.”
  • The research suggests a surprising conclusion: to maximize predictive accuracy, final decisions should be left to formulas, especially in low-validity environments.
Intuitions vs. Formulas

- Idea of using formulas instead of expert opinion generally not welcomed
- Some acceptance of this idea – in baseball, for example
Intuitions vs. Formulas

• The prejudice against algorithms is magnified when the decisions are consequential.

• Meehl remarked, “I do not quite know how to alleviate the horror some clinicians seem to experience when they envisage a treatable case being denied treatment because a ‘blind, mechanical’ equation misclassifies him.”

• In contrast, Meehl and other proponents of algorithms have argued strongly that it is unethical to rely on intuitive judgments for important decisions if an algorithm is available that will make fewer mistakes.

Will Robots Steal Your Job?
You’re highly educated. You make a lot of money. You should still be afraid.
By Farhad Manjoo | Updated Monday, Sept. 26, 2011, at 4:20 AM ET

Join Farhad Manjoo in Washington, D.C., on Thursday, Sept. 29, for a Future Tense event on robots and the workforce. Manjoo will be joined by Tyler Cowen, author of The Great Stagnation and blogger at Marginal Revolution; Robbie Allen, whose company StatSheet could put sportswriters on unemployment lines; and others. To RSVP for a free ticket, click here.

If you’re taking a break from work to read this article, I’ve got one question for you: Are you crazy? I know you think no one will notice, and I know that everyone else does it. Perhaps your boss even approves of your Web surfing; maybe she’s one of those new-age managers who believes the studies showing that short breaks improve workers’ focus. But those studies shouldn’t make you feel good about yourself. The fact that you need regular breaks only highlights how flawed you are as a worker. I don’t mean to offend. It’s just that I’ve seen your competition. Let me tell you: You are in peril.

At this moment, there’s someone training for your job. He may not be as smart as you are—in fact, he could be quite stupid—but what he lacks in intelligence he makes up for in drive.
Problem Set 10

1. In question 1 of problem set 9, we used the UMCPsalaries.xlsx data set to explore a linear relationship between salary and years of experience (yrdeg, or years after highest degree), with yrdeg as the independent (x-axis) and salary as the dependent (y-axis) variable.

1. Add a new trendline to the scatterplot you created last week. Select “exponential” for the trendline type, and display the equation and R-squared value on the chart. The best-fit equation has the form \( y = ae^{bx} \). Explain the meanings of the values of a and b, and the value of R-squared, in plain English.

2. In a new column in the data set, create a new variable, LNsalary, that is the natural logarithm of salary. (Use the formula “=LN(x)” to calculate the log of salary.) Create a new scatterplot with yrdeg on the x-axis and LNsalary on the y-axis. Add a linear trendline and display the equation and R-squared value on the chart. The best-fit equation has the form \( y=a+bx \). Compare these values of a and b to those in part 1. What do you notice?

3. Now use the data analysis toolpack (Mac users should use StatPlus or the reg.xlsx worksheet) to do a regression analysis of yrdeg and LNsalary. What is the standard error? Explain the meaning of the standard error in plain English. (If the logarithm of salary changes by a certain amount, what happens to salary?)

4. Plot the residuals, and compare to last week’s plot. Has the heteroscedasticity been corrected?

5. Compare last week’s regression with this week’s regression. Compare the values of R-squared, the t-statistic corresponding to the test of the null hypothesis that the response and explanatory variables are uncorrelated, the standard error of the regression, and the analyses of the residuals. Which model is best overall?
Problem Set 10

2. The file USpopulation.xlsx contains Census Bureau estimates of U.S. population (including armed forces overseas) from 1900-2012, on July 1st of each year.

1. Make a scatterplot of the data from 1965 to 2012. Using the "trendline" tool, fit an exponential curve to the data, displaying the regression equation and R-squared on the plot. Is the equation a good fit to the data? The equation is of the form $y = ae^{bx}$. Explain the value of $b$ in plain English.

2. Using the best-fit equation, calculate the estimated population in 1970, 1980, 1990, 2000, and 2010 (i.e., the values of $y$-hat for these values of $x$). (Use at least six significant digits for $a$ and $b$ to get accurate values of $y$-hat.) Compare these values to the actual population. Compute both the error, and the percentage error. How accurate are the estimates produced by the best-fit equation?

3. In a new column in the data set, create a new variable, LNpop, that is the natural logarithm of population. Using the Data Analysis Toolpack, StatPlus, reg.xlsx, or LINEST, do a regression analysis of LNpop and year. What is the standard error (i.e., the standard deviation of the residuals)? Compare this to the errors you computed in part 2. What do you find?
Problem Set 10

3. The file CigarettesPrice.xlsx contains data on the average price and per-capita annual sales in 48 continental states from 1985 to 1995. (This type of data set, in which observations for a particular entity occur multiple times for different years, is known as “panel” data.)

1. Create a scatterplot of the data. Add a trendline to the chart, select “power,” and display the equation and R-squared on the chart. The best-fit equation has the form \( y = ax^b \). Interpret the values of b and R-squared in plain English.

2. In a new column in the data set, create two new variables, LNpacks and LNprice, by taking the natural logarithm of packs and price. Create a new scatterplot with these variables. Add a linear trendline and display the equation and R-squared value on the chart. The best-fit equation has the form \( y=a+bx \). Compare the values of a and b to those in part 1. What do you notice?

3. Now use the data analysis toolpack, StatPlus, reg.xlsx, or LINEST to do a regression analysis of LNpacks and LNprice. What is the standard error? Explain the meaning of the standard error in plain English. (If the logarithm of price changes by a certain amount, what happens to price?)
Questions?