Some Review and Hypothesis Testing
Outline

• Discussing the homework
  – questions from Joey and Phoebe
• Review of Statistical Inference
• Properties of OLS under the normality assumption
• Confidence Intervals, T test, p value
• your thoughts on the course so far
The Homework

• Thinking about missing values
• cap drop vs. drop
• why aren’t the coefficients and standard errors the same when I switch to a regression using z-scores?
• But shouldn’t the t-stats be the same?
  – What was the population when you set to z-scores?
The Homework

• Does the DV and/or the IVs need to be normally distributed for OLS to produce an unbiased result?
  – Short answer: No and No
  – Long answer: It helps.
Linear Regression

- Regression is based on the concept of a simple proportional relationship
- We want to estimate the PRF
  \[ Y_i = B_1 + B_2 X_i + u_i \]
- On the basis of the SRF

\[ Y_i = \hat{B}_1 + \hat{B}_2 X_i + \hat{u}_i \]
Method of OLS

• Ordinary Least Squares (OLS) is a method of finding the linear model which minimizes the sum of the squared errors.

\[
\text{USS} = \sum u_i^2
\]

\[
\text{USS} = \sum (Y_i - \hat{Y}_i)^2
\]

\[
\text{USS} = \sum (Y_i - \hat{B}_1 - \hat{B}_2 X)^2
\]
More Review

• To find the minimum of a function, we can use calculus

• We minimize the sum of the squared errors for OLS

• This gives us (and gives software) equations to estimate $B_1$ and $B_2$. 
Today

• We know how to estimate our parameters.
• What are the properties of the OLS Estimators?
• How precise are the estimates?
• How close are the estimates to the true relationship in the population?
• How can we use that to test hypotheses
Properties of OLS Estimators

• Recall the Estimators

$$\hat{B}_1 = \bar{Y} - \hat{B}_2 \bar{X}$$

$$\hat{B}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

• Property 1: They are Linear Estimators
More Properties

• Property 2: They are Unbiased

\[ E \left( \hat{B}_1 \right) = B_1 \]
\[ E \left( \hat{B}_2 \right) = B_2 \]
Property 3: Efficiency

- OLS produces efficient estimates
- An efficient estimator is one with the least variance compared to any other estimator of the parameter
Gauss-Markov Theorem

• Given the assumptions of the classical linear regression model, OLS, in the class of unbiased linear estimators is BLUE (Best Linear Unbiased Estimator), where best is minimum variance

• Property 4: OLS estimates are consistent:
  – As sample size, N, grows the OLS estimators converge to the true population value
Next Topic

• We can estimate our unknown parameters
• We know that the OLS estimators are the Best Linear Unbiased Estimators
• But, how precise are the estimators in a given sample?
How precise are our estimates?

• Standard error is a measure of the precision of our estimator.

• Variance of our estimate of beta_hat is directly proportional to $\sigma^2$ (i.e., the variance of the error term).

• Variance of both estimates is inversely proportional to $\Sigma(X_i - X\text{bar})$ (i.e., the variance of $x$).

• As $N$ increases, the variability of our estimates will go down.
Statistical Inference

• A methodology for learning about populations and their probability distributions by sampling from them
Review: The Normal Distribution

• For many random variables, the probability distribution is a bell curve, or a normal curve
• Use normal distribution to calculate probability beyond a given value of z
Review: The Sample Mean

\[ \bar{X} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

• How good of an estimate of the population mean is the sample mean?

• The central limit theorem gives an answer
Central Limit Theorem

• Let $X_1, X_2, ..., X_n$ denote $n$ independent random variables, all of which have the same pdf with mean, $u$, and variance, $\sigma^2$.

• The sample mean is denoted as:

$$\overline{X} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

• As $n$ increases indefinitely,

$$\overline{X} \xrightarrow{n\to\infty} N(u, \frac{\sigma^2}{n})$$

• English-The sample mean approaches the normal distribution with a mean of $u$, and a variance $\sigma^2/n$. 
Properties of the Sample Mean

- A linear function of the Random Variable $X_i$
- It is a normal Random Variable
- The expected value is the population mean
- The standard deviation decreases as sample size increases
Standard Normal Distribution

• A special type of normal distribution, with mean 0 and standard deviation 1

• Z Formula: Can use to convert normal distribution to standard normal:

\[ Z = \frac{X - \mu}{\sigma} \]
Review of Hypotheses

• A hypothesis is a statement about some characteristic of a variable or a collection of variables.

• The null hypothesis $H_0$ is the hypothesis that is directly tested. It states that the parameter tested has no effect.

• The alternative hypothesis $H_A$ is a hypothesis that contradicts the null hypothesis. This is usually a statement that the researcher believes to be true.

• Proof by contradiction: reject the null hypothesis and support the alternative hypothesis.
Significance Tests

• A significance test is a way of statistically testing a hypothesis by comparing the data to values predicted by the hypothesis.
Review of Type I and Type II Error

<table>
<thead>
<tr>
<th>State of the World</th>
<th>( H_0 ) acceptable</th>
<th>( H_0 ) Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 ) is true</td>
<td>Correct Decision</td>
<td>Type I error</td>
</tr>
<tr>
<td></td>
<td>Confidence level</td>
<td>Probability = ( \alpha )</td>
</tr>
<tr>
<td></td>
<td>= ( 1 - \alpha )</td>
<td>(level of test)</td>
</tr>
<tr>
<td>( H_0 ) is false</td>
<td>Type II error</td>
<td>Correction decision</td>
</tr>
<tr>
<td>(( H_a ) is true)</td>
<td>Probability = ( \beta )</td>
<td>Probability = ( 1 - \beta )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(power of test)</td>
</tr>
</tbody>
</table>
Hypothesis Testing

• Recall that there are two types of hypothesis tests, significance tests and confidence intervals, but both will yield the same results.

• To do hypothesis tests, we must first transform a normal RV to a standard normal RV:

\[ Z = \frac{X - \mu}{\sigma_x} \]
Example: Is America Moderate?

<table>
<thead>
<tr>
<th>Responses</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Extremely liberal</td>
<td>12</td>
</tr>
<tr>
<td>2 Liberal</td>
<td>66</td>
</tr>
<tr>
<td>3 Slightly liberal</td>
<td>109</td>
</tr>
<tr>
<td>4 Moderate, middle of road</td>
<td>239</td>
</tr>
<tr>
<td>5 Slightly conservative</td>
<td>116</td>
</tr>
<tr>
<td>6 Conservative</td>
<td>74</td>
</tr>
<tr>
<td>7 Extremely conservative</td>
<td>11</td>
</tr>
</tbody>
</table>
Using the Test Statistic

- $H_0 : \mu = 4.0$ (middle of road)
- $H_A : \mu \neq 4.0$
- Test Statistic

$$\bar{x} = 4.032; s = 1.257; SE = \frac{s}{\sqrt{n}} = \frac{1.257}{627} = 0.050$$

$$z = \frac{\bar{x} - \mu_0}{SE} = \frac{4.032 - 4.0}{0.050} = 0.64$$

- P-value
- $Pr[(Z>0.64) \text{ or } (Z< -0.64)] = 0.2611 + 0.2611 = 0.52$
- Do not reject the null hypothesis
Using the Confidence Interval

• Or we can use the confidence interval
  • $4.032 \pm 1.96 \text{ SE}$
  • $4.032 \pm 1.96(.05)$
  • $4.032 \pm 0.098$
  • $(3.93, 4.13)$

• $4.0$ is contained within the 95% confidence interval, so we can not reject the null hypothesis that America is moderate
Types of Tests

• Two-tailed tests
  – Use both the right and left tail
  – $H_0 : \mu = 4.0$ (middle of road)
  – $H_A : \mu \neq 4.0$

• One-tailed tests
  – Only use probability in one tail
  – $H_0 : \mu \leq 4.0$ (middle of road)
  – $H_A : \mu > 4.0$
Review

• If the sampling distribution of a RV is normal, then we can make inferences about how close our sample is to the true population.

• To be able to use test statistics with OLS, we need the sampling distribution of $B_2$ to be normal.
Classical Linear Regression Model

- Since $B_1$, $B_2$ and $\sigma^2$, are random variables, we need to know if their probability distributions are normal.
Classical Normal Linear Regression Model

• The CLRM does not make assumptions about the probability distribution of $u$
• To see how close our estimated estimated $B_2$ is to the true $B_2$, we need to know the probability distribution of our estimate
  – which depends on the probability distribution of $u$.
• The Classical Normal Linear Regression Model solves this problem
Normality Assumption of $U_i$ under CNLRM

- The CNLRM assumes that each $u_i$ is distributed normally with
  - Mean $E(u_i)=0$
  - Variance $E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma^2$
  - Covariance $(u_i, u_j) E\{[u_i - E(u_i)][u_j - E(u_j)]\} = E(u_i, u_j) = 0, i \neq j$
  - Or, $u_i \sim N(0, \sigma^2)$ or $u_i \sim \text{NID}(0, \sigma^2)$
    - $u_i$ is normally and independently distributed with mean zero and variance sigma squared
Why do we make the normality assumption?

• Recall that the error term is a random variable.
• From central limit theorem, know distribution of the sampling mean of a random variable tends to be normal in the limit.
• Since $B_1$ and $B_2$ are linear functions of $u_i$, they are normally distributed and we can use statistical tests such as t test.
Review of Properties of OLS Estimators

• Recall our OLS estimators

\[ \hat{B}_1 = \bar{Y} - \hat{B}_2 \bar{X} \]

\[ \hat{B}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \]

• Earlier, we showed that they are:
  – Linear
  – Unbiased
  – Efficient
  – They are BLUE
Now we added another property

- $B_1$ and $B_2$ are normally distributed

$$\hat{B}_1 \sim N\left( \beta_1, \frac{\sum X_i^2}{n\sum x_i^2} \sigma^2 \right)$$

$$\hat{B}_2 \sim N\left( \beta_2, \frac{\sigma^2}{\sum x_i^2} \right)$$

- This will allow us to do hypothesis tests
So what are t-statistics?

- $B_1\hat{\text{hat}}$ and $B_2\hat{\text{hat}}$ are normally distributed
- So is $\sigma_{\hat{\beta}_1}$
- But the ratio of these two is NOT (quite) normal
  - it follows a t distribution, which is a lot like a normal distribution, but not quite.

$$(\hat{\beta}_1)/\sigma_{\hat{\beta}} \sim t_{(n-k-1)}$$

- where $k$ is the # of parameters to be estimated
t-statistics (cont)

• Note the addition of a “degrees of freedom” constraint

• Thus the more data points we have relative to the number of parameters we are trying to estimate, the more the t distribution looks like the z distribution.

• When n>100 the difference is negligible
Statistical vs. substantive significance

• If $n$ is large, our estimated may be statistically significant but substantively trivial.

• If $n$ is small (or the standard errors are large for other reasons), we may know the effect is not zero but have NO IDEA how big it is.
How to Present Results?

• This could be half a class. But let's talk a bit.