Segment 9: Assumptions and Hypothesis Testing

Prof. Sharyn O’Halloran
Assumptions of the Linear Regression Model

I. Assumptions of the Linear Regression Model

A. Overview

- Develop techniques to test hypotheses about the parameters of our regression line.
  - For instance, one important question is whether the slope of the regression line is statistically distinguishable from zero?
  - As we showed last time, if the independent variable has no effect on the dependent variable then \( b \) should be close to zero.
Assumptions of the Linear Regression Model

A. Overview (cont.)

- So testing if $b=0$ is really a test to see if the independent variable has any effect on the dependent variable.
Assumptions of the Linear Regression Model

B. OLS Line

1. The predicted line

   The standard regression line that we had seen from last time is just

   $\hat{Y} = a + bX$

   \[ b = \frac{\sum xy}{\sum x^2} \]

   \[ a = \bar{Y} - b\bar{X} \]

   Think of this regression line as the expected value of $Y$ for a given value of $X$. 
Assumptions of the Linear Regression Model

2. Error term

- To include this in our analysis we use the equation

\[ Y = \alpha + \beta X_i + \epsilon_i \]

- where \( \epsilon_i \) is the error in observation \( i \).
- These epsilons (errors) are random variables and each have their own distribution.
Assumptions of the Linear Regression Model

C. Assumptions

To be explicit, the analysis that follows makes the following four assumptions.

1. Linearity
   - The true relation between Y and X is captured in the equation: \( Y = a + bX \)

2. Homoskedasticity (Homogeneous Variance)
   - Each of the \( \varepsilon_i \) has the same variance.
     \[
     E(\varepsilon_i^2) = \sigma^2 \quad \text{for all } i
     \]
Assumptions of the Linear Regression Model

3. Independence
   - Each of the $\epsilon_i$'s is independent from each other. That is, the value of one does not effect the value of any other observation i's error.
   \[ E(\epsilon_i \epsilon_j) = 0 \quad \text{for } i \neq j \]

4. Normality
   - Each $\epsilon_i$ is normally distributed
Assumptions of the Linear Regression Model

- Combine assumption 4 with assumption 2, this means that the error terms are normally distributed with mean = 0 and variance $\sigma^2$

We write this as $\varepsilon_i \sim N(0, \sigma^2)$
Assumptions of the Linear Regression Model

- Notice that $\sigma^2$, the variance of the $\varepsilon$'s, is the expected squared deviation of the Y's from the regression line.

$\sigma^2 = E(Y_i - \hat{Y}_i)^2$

$= E(Y_i - (a + bX_i))^2$

$= E(\varepsilon_i)^2$. 
Assumptions of the Linear Regression Model

D. Examples of Violations

1. Linearity

- The true relation between the independent and dependent variables may not be linear.
  - For example, Consider campaign fundraising and the probability of getting of winning an election.
Assumptions of the Linear Regression Model

2. Homoskedasticity
   - This assumption means that we do not expect to get larger errors in some cases than in others.
   - Of course, due to the luck of the draw, some errors will turn out to be larger than others. But homoskedasticity is violated only when this happens in a predictable manner.
Assumptions of the Linear Regression Model

2. Homoskedasticity (cont.)
   - Example: income and spending on certain goods.
     - People with higher incomes have more choices about what to buy.
     - We would expect that there consumption of certain goods is more variable than for families with lower incomes.
Assumptions of the Linear Regression Model
Assumptions of the Linear Regression Model

3. Independence
- The independence assumption means that two variables will not necessarily influence one another.
- The most common violation of this occurs with data that are collected over time or time series analysis.
  - Example: high tariff rates in one period are often associated with very high tariff rates in the next period.
  - Example: Nominal GNP

4. Normality
- Of all the assumptions, this is the one that we need to be least worried about violating. Why?
Assumptions of the Linear Regression Model

E. Summary of Assumptions

1. Algebra

These assumptions can be written as:

- a. The random variables $Y_1, Y_2, \ldots, Y_i$ are independent with

  \[
  \text{Mean} = \alpha + \beta X \\
  \text{Variance} = \sigma^2.
  \]

- Notice that when $\beta = 0$, the mean of $Y_i$ is just simply the intercept $\alpha$. 
Assumptions of the Linear Regression Model

b. Also, the deviation of $Y$ from its expected values is captured by the disturbance terms or the error $e$.

$$Y_i = \alpha + \beta X_i + e_i$$

where $e_1, e_2, \ldots, e_i$ are independent errors with
- Mean = 0
- Variance $= \sigma^2$. 
Assumptions of the Linear Regression Model

2. Recap

There are two things that we must be clear on:

- First, we never observe the true population regression line or the actual errors.
- Second, the only information that we know are Y observations and the resulting fitted regression line.
  - We observe only X and Y from which we estimate the regression and the distribution of the error terms around the observed data.
II. Discussion of Error term

A. Regression as an approximation

1. The Fitted Line

- The least squares method is the line that fits that data with the minimum amount of variation.
- That is, the line that minimizes the sum of squared deviations.

a. Remember the criterion:

\[ \text{Minimize } \sum d^2 = \sum (Y - \hat{Y})^2 \]
Discussion of Error term

1. The Fitted Line (cont.)

- b. Approximation
- But the regression equation is just an approximation, unless all of the observed data fall on the predicted line

\[ Y = a + bX \]
Discussion of Error term

2. The Fitted line with variance
   a. Analogy to point estimates
      i. With point estimates, we did not expect that all the observed data would fall exactly on the mean.
      ii. Rather, the mean represented the expected or the average value.
      iii. Similarly, then, we do not expect that the regression line be a perfect estimate of the true regression line. There will be some error.
Discussion of Error term

2. The Fitted line with variance (cont.)

![Graph showing the relationship between Yield and Fertilizer. The graph includes a line representing a fitted line and labels for change in Y and change in x. The error term (ei) is also indicated.]
b. Measurement of variance

- We need to developed a notion of dispersion or variance around the regression line.
- A measurement of the error of the regression line is

\[ e_i = Y_i - \hat{Y}_i \]
B. Causes of Error

This error term is the result of two components:

1. Measurement Error

- Error that results from collecting data on the dependent variable.
  - For example, there may be difficulty in measuring crop yields due to sloppy harvesting.
  - Or perhaps two scales are calibrated slightly differently.
Discussion of Error term

2. Inherent Variability
   - Variations due to other factors beyond the experimenter's control.
     - What would be some examples in our fertilizer problem?
       - Rain
       - Land conditions
C. Note:

- The error term is very important for regression analysis.
- Since the rest of the regression equation is completely deterministic, all unexpected variation is included in the error term.
- It is this variation that provides the basis of all estimation and hypothesis testing in regression analysis.
III. Sampling Variability

Are the parameters $a$ and $b$ good estimates of the true regression line?

$$E(Y) = \alpha + \beta x \quad \text{(true regression line)}$$

$$\hat{Y} = a + bx \quad \text{(estimated regression line)}$$
Sampling Variability

A. Sampling distribution of $b$

- We know from the central limit theorem that if we take a large enough sample, that the sample will be normally distributed, with
  
  Expected value of $b = \beta$
  
  Standard error of $b = \frac{\sigma}{\sqrt{\Sigma x^2}}$

- Here $\sigma$ represents the standard deviation of the $Y$ observations about the population line.
  
  - We define:
    
    $\sigma = \sqrt{\frac{\Sigma (Y - \bar{Y})^2}{n}}$
Sampling Variability

- **B. Standard Error of b**
  - **1. Graph**
    - So as before we can represent the estimate of the true population as follows:
    - \[ \text{SE}(b) = \frac{\sigma}{\sqrt{\sum x^2}} \]

\[ E(b) = \beta \]
2. Formulas

- We can restate the standard error as follows:
  \[
  \text{Standard Error} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{\sum x^2}{n}}
  \]

- Recall that \( \frac{\sum x^2}{n} \) is just the mean standard deviation.
  - Remember \( x^2 = (X-\bar{X})^2 \)
  - so the \( \frac{\sum x^2}{n} \) is approximately the variance \( s^2_x \).

- Thus,
  \[
  \text{the standard error} \approx \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{s_x}
  \]
Sampling Variability

3. Three ways to improve the accuracy of estimating $b$.

- There are three ways for the standard error to be reduced to improve the accuracy of our estimate $b$:
  - By reducing $\sigma$, the variability of the Y observations (controlling for outside effects).
  - By increasing $n$, the sample size.
  - By increasing $S_X$, the spread of the X values, which are determined by the experimenter.
Sampling Variability

- Example: Data with little spread

When $S_X$ is small then we have little information from which we can predict the value of $Y$.

Our regression line will be very sensitive to the exact location of each $X$ value.

So small errors may have a large effect on our estimates of the true regression line.
Sampling Variability

b. Example: data with large spread

- If the $x$ values take on a wide range of values ($s_x$, the spread, is large), then our estimates are not so sensitive to small errors.
Sampling Variability

C. Example: Ruler Analogy

- For example, think about the ruler analogy.
- If I want you to draw a line and I give you two points that are very close together then there are many lines that you can possible draw.
- On the other hand, if I give you two points that are at the ends of the ruler you will be more accurate in drawing your line.
- The spread of X has the same effect.
IV. Confidence Intervals and Tests for $\beta$

A. Estimating the Standard Error of $b$

1. $\sigma^2$ Known

- Before we can make inferences about our regression parameters, there is one more piece of information that we need. The standard error of $b$ is given by:
  \[
  \text{Standard Error of } b = \sqrt{\frac{\sigma^2}{\sum x^2}}
  \]

- Our formulas for the standard error includes $\sigma^2$, which is the variance of $Y$ about the true population regression line.

- $\sigma^2$ is the variance of each of the $\varepsilon_i$. 
Confidence Intervals and Tests for $\beta$

2. $\sigma^2$ unknown

- But as we know, most times we do not observe the underlying population.
- Since $\sigma^2$ is generally unknown, it must be estimated.
- Remember that for the $\epsilon_i$'s, $\sigma^2$ is the averaged squared distance from the mean.
- So we estimate $\sigma^2$ by the averaged squared distance from the observed data points from the estimated regression line.
An obvious candidate would just be the average of the $d^2$s. That is,

$$s^2 = \frac{\sum(Y - \bar{Y})^2}{n} = \frac{\sum d^2}{n}.$$
Confidence Intervals and Tests for $\beta$

3. N-2 Degrees of Freedom

- Correction
  - We need to divide by $n-2$ instead of $n$, so that we have an unbiased estimate of $\sigma^2$:
    
    $$s^2 = \frac{\sum d^2}{n-2} = \frac{1}{n-2} \sum (Y - \hat{Y})^2$$

  - Here, we have already used up two degrees of freedom in calculating the regression line, one for the $y$-intercept and one for the slope.
Correction (cont.)

- So for the estimated standard error of $b$, we have the formula:

  $$\text{estimated } SE(b) = \frac{s}{\sqrt{\sum x^2}}.$$ 

- Since we used the estimated standard error instead of the actual standard error, what distribution do you think we have to use?
Confidence Intervals and Tests for $\beta$

- **B. Confidence Intervals**
  - **1. Formula**
    - Finally, we can specify a confidence interval for the true slope of $\beta$ given our estimate $b$:
      \[
      \beta = b \pm t_{0.025} \cdot \text{SE}
      \]
      \[
      \beta = b \pm t_{0.025} \frac{s}{\sqrt{\sum x^2}}
      \]
  - **2. Distribution**
    - We use the t-distribution with n-2 degrees of freedom
Confidence Intervals and Tests for $\beta$

- 3. Example: Calculate a 95% Confidence Interval for the Slope

Let's go back to the fertilizer example from last time.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>$Y = 36.4 + 0.059X$</th>
<th>d = $Y - \bar{Y}$</th>
<th>$(Y - \bar{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>40</td>
<td>42.3</td>
<td>-2.3</td>
<td>5.29</td>
</tr>
<tr>
<td>200</td>
<td>50</td>
<td>48.2</td>
<td>1.8</td>
<td>3.24</td>
</tr>
<tr>
<td>300</td>
<td>50</td>
<td>54.1</td>
<td>-4.1</td>
<td>16.81</td>
</tr>
<tr>
<td>400</td>
<td>70</td>
<td>60.0</td>
<td>10.0</td>
<td>100.0</td>
</tr>
<tr>
<td>500</td>
<td>65</td>
<td>65.9</td>
<td>-0.9</td>
<td>.81</td>
</tr>
<tr>
<td>600</td>
<td>65</td>
<td>71.8</td>
<td>-6.8</td>
<td>46.24</td>
</tr>
<tr>
<td>700</td>
<td>80</td>
<td>77.7</td>
<td>2.3</td>
<td>5.29</td>
</tr>
</tbody>
</table>

$\bar{X} = 400 \quad \bar{Y} = 60 \quad 0 \quad \Sigma = 177.68$
Confidence Intervals and Tests for $\beta$

- a. Calculate regression line?
  - Remember that we calculated the regression line to be:
    $$\hat{Y} = 36.4 + 0.059X$$

- b. Calculate a 95% confidence interval:
  - So we can calculate $s^2$ as:
    $$s^2 = \frac{1}{n-2} \sum (Y - \hat{Y})^2$$
    $$s^2 = \frac{177.68}{7 - 2}$$
    $$s^2 = 35.5$$
Confidence Intervals and Tests for $\beta$

- b. Calculate a 95% confidence interval: (cont.)

$$
\beta = b \pm t_{0.025} \frac{s}{\sqrt{\sum x^2}}.
$$

$$
\beta = 0.059 \pm 2.57 \frac{\sqrt{35.5}}{\sqrt{280,000}}
$$

$$
\beta = 0.059 \pm 2.57 (0.0113)
\beta = 0.059 \pm 0.029
$$

$$
0.030 \leq \beta \leq 0.088
$$
Confidence Intervals and Tests for $\beta$

- If the slope were .030, what would the y-intercept be?
  \[ a = \bar{Y} + b\bar{X} \]
  \[ a = 60 - .030 (400) = 48 \]

- If the slope were .088, what would the y-intercept be?
  \[ a = 60 - .088 (400) = 24.8 \]
Confidence Intervals and Tests for $\beta$

- We can graph these two lines, then, like this:

![Graph showing yield vs. fertilizer with confidence intervals]

- This means that we are 95% confident that the true regression line falls within these lower and upper bounds.
C. Testing Hypotheses: Accept or Reject

Specifically, the hypothesis we're most interested in testing is that $\beta = 0$.

- $H_0: \beta = 0$
- $H_a: \beta \neq 0$

Against the alternative that there is no relation between the independent and dependent variables.

How can we perform this test at the $\alpha = 5\%$ level?
Since zero does **not** fall within this interval, we fail to accept the hypothesis that $\beta = 0$.

That is, according to our data, fertilizer really does help increase crop yield. Yay!
Confidence Intervals and Tests for $\beta$

- **D. p-Values**
  - We can also perform a p-value test of hypotheses
    - Similar to the ones we did before, we calculate:
      \[ t = \frac{b - b_0}{SE} \]
    - where $b_0$ is the value of $b$ under the null hypothesis.
    - Since we are interested in the hypothesis $H_0: b = 0$,
      \[ t = \frac{b}{SE} \]
Confidence Intervals and Tests for $\beta$

- Just like before, we can draw the distribution of $b$:

- If the observed value of $b$ falls within one of the tails, then we reject the null.
Confidence Intervals and Tests for $\beta$

- For the fertilizer data, we have:
  \[ t = \frac{.059}{.0113} = 5.2; \quad \text{d.f.} = 7-2=5 \]

- We go to our t-tables and find that for 5 degrees of freedom, 5.2 is greater than the value $t_{.0025} = 4.77$.

- These data are very significant. \( \Rightarrow \) We can reject the null hypothesis.
Confidence Intervals and Tests for $\beta$

E. Examples

Now we want to know do people save more money as their income increases?

Suppose we observed 4 individual's income and saving rates?

<table>
<thead>
<tr>
<th>Income (X)</th>
<th>Saving s (Y)</th>
<th>x</th>
<th>y</th>
<th>xy</th>
<th>x^2</th>
<th>d=Y-Ȳ</th>
<th>(Y- Ȳ)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>2.0</td>
<td>1</td>
<td>-0.2</td>
<td>-0.2</td>
<td>1</td>
<td>2.34</td>
<td>-0.34</td>
</tr>
<tr>
<td>18</td>
<td>2.0</td>
<td>-3</td>
<td>-0.2</td>
<td>0.6</td>
<td>9</td>
<td>1.77</td>
<td>0.23</td>
</tr>
<tr>
<td>17</td>
<td>1.6</td>
<td>-4</td>
<td>-0.6</td>
<td>2.4</td>
<td>16</td>
<td>1.63</td>
<td>-.03</td>
</tr>
<tr>
<td>27</td>
<td>3.2</td>
<td>6</td>
<td>1.0</td>
<td>6.0</td>
<td>36</td>
<td>3.05</td>
<td>0.15</td>
</tr>
</tbody>
</table>

$\bar{X} = 21$  $\bar{Y} = 2.2$  $\sum = 8.8$  $\sum = 62$  $\sum = 0.192$
1. Calculate the fitted line

\[ Y = a + bX \]

- Estimate \( b \)
  
  \[ b = \frac{\Sigma xy}{\Sigma x^2} = \frac{8.8}{62} = 0.142 \]

  - What does this mean?
  - On average, people save a little over 14% of every extra dollar they earn.

- Intercept \( a \)
  
  \[ a = \bar{Y} - b\bar{X} = 2.2 - 0.142 \times 21 = -0.782 \]

  - What does this mean?
  - With no income, people borrow.

So the regression equation is:

\[ Y = -0.78 + 0.142X \]
2. Calculate a 95% confidence interval

Now let's test the null hypothesis that $\beta = 0$. That is, the hypothesis that people do not tend to save any of the extra money they earn.

$H_0: \beta = 0$

$H_a: \beta \neq 0$;

at the 5% significance level

What do we need to calculate the confidence interval?

$s^2 = \Sigma d^2 / n-2 = .192 / 2 = 0.096$

$s = \sqrt{0.096} = .309$
Confidence Intervals and Tests for $\beta$

- What is the formula for the confidence interval?

$$\beta = b \pm t_{0.025} \frac{s}{\sqrt{\sum x^2}}.$$ 

$$\beta = .142 \pm 4.30 \cdot .309 / \sqrt{62}$$ 

$$\beta = .142 \pm .169$$

$$-.027 \leq \beta \leq .311$$
Confidence Intervals and Tests for $\beta$

3. Accept or reject the null hypothesis

- Since zero falls within this interval, we cannot reject the null hypothesis. This is probably due to the small sample size.
Confidence Intervals and Tests for $\beta$

F. Additional Examples:

1. How about the hypothesis that $\beta = 0.50$, so that people save half their extra income?
   - It is outside the confidence interval, so we can **reject** this hypothesis.

2. Let's say that it is well known that Japanese consumers save 20% of their income on average. Can we use these data (presumably from American families) to test the hypothesis that Japanese save at a higher rate than Americans?
   - Since 20% also falls within the confidence interval, we **cannot reject** the null hypothesis that Americans save at the same rate as Japanese.
V. Homework

- Part of your homework for next time is to run a regression on SPSS.
- This will be a quick exercise, just to get you used to the commands in regression analysis.
A. Variables

- Take the same two variables you used for the homework you turned in today.
- I used \textit{PARTYID} and \textit{INCOME91}, to see if one's income had a significant effect on the party you identify yourself with.
- The computer will automatically delete all missing data for regressions, so you don't have to recode the variables. (Usually it's a good idea to do this anyway, but this time you can be a little sloppy.)
B. Commands

- The command to run a regression is just `REGRESSION`.
- Next, you input the **variables** you'll be using in the analysis.
- You **enter all variables**, including the dependent and independent variables.
- Then tell the computer which variable is the **dependent** variable. In this case, it was `PARTYID`.
- Finally, you put in `/ METHOD ENTER`. Don't worry about this part of the command for now; I'll explain it later.
C. Results

- The bottom of the page shows the regression results.
- The slope was 0.039, which is positive. This means that as income rises, you are more likely to identify yourself as a Republican.
- The standard error was .0125, as reported in the next column.
- The t-statistic, then, is just .039 / .0125, which is 3.093.
- The p-value corresponding to this is given in the "Sig T" column, and it's .0020.
C. Results (cont.)

- So would we accept or reject the null hypothesis that $\beta = 0$ at the 5% level?
- That's right, since .0020 is less than .05, we would **reject** the null.
- So as your income rises, you're **more likely** to identify yourself as Republican.
VII. Predicting Y at a given level of X.

As we said before, regression analysis can be used to predict outcomes as well as to summarize data.

- For instance, after running a regression and getting the best fit line for some data, we can ask what is the most likely value for Y at a given level of X, say $X_0$.
- We know that this is the value $\hat{Y}_0 = a + bX_0$
- Now we might ask how confident we are of this prediction. That is, we want a confidence interval for $\hat{Y}_0$.
- It turns out that this confidence interval can be calculated in two ways, depending on exactly what we're interested in:
Predicting Y at a given level of X.

1. What is the true expected value of $Y_{x_0}$ over many different observations? That is, what is the expected value of Y when $X=X_0$?
   - For instance, we might ask on average how large the yield should be if we add 450 lbs of fertilizer per acre.

2. What will the value of Y be if we have only one observation of it for $X=X_0$?
   - For instance, say we want a 95% confidence interval for next year's crop only, given that we plan to add 450 lbs of fertilizer per acre.
Predicting Y at a given level of X.

- A. Confidence Interval for the Mean of Y at X
- B. Prediction Interval for a Single Observation
- C. Comparisons of the two