Stat 324 – Day 6
Basic Regression Model Assumptions (2.2, 2.3)

**Last Time:** $R^2$ and $s$ can be used as *descriptive* measures of how well the model fits the data

- $R^2$ indicates the percentage of variation in the response (e.g., $\text{Var}(Y) = s_Y^2$) that is explained by the regression model (e.g., $\text{Var}$ of residuals $= s^2$) relative to $\text{Var}(Y)$ (*model utility*)
  - $R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$
  - Bounded between 0 and 1 so can compare across data sets
- $s$ indicates the typical error in observations from the regression line (*degree of predictability*)
  - At 3 years of age, how likely is it for a boy to be taller than 90 cm?
  - $s^2 = \frac{\text{SSE}}{(n - 2)} = \frac{\sum (y_i - \hat{y}_i)^2}{(n - 2)}$

For statistical *inference*, we will need to make some explicit assumptions about our model.

**Example 1:** The Berkeley Guidance Study which monitored the height (cm) and weight of boys and girls born in Berkeley, California between January 1928 and June 1929.

Shapes:  
![Shape Diagram]

Spreads:  
![Spread Diagram]

Centers:  

The *Simple Linear Regression Model* specifies a mathematical, probabilistic relationship between the means of these subpopulations at each $x$ and the explanatory variable. We are saying the value of the mean of the response variable $Y$ depends on the value of $x$ and that the dependence is linear.

$$E(\text{height} \mid \text{age}) = \beta_0 + \beta_1 \text{age}$$

**Interpretation of slope** $\beta_1$ (vs. slope estimate $\hat{\beta}_1$): the change in average height for each additional year

- **Linearity**: The means of the conditional distributions at each $x$ value fall along a line.
- **Independence**: The observations at each $x$ value are independent and across the $x$ values (no serial correlation or clustering)
- **Normality**: We expect the observations at each $x$ to vary normally about the (conditional) mean
- **Equal variance**: We expect the (conditional) variability in the responses to be the same at each $x$ with standard deviation $\sigma_e$.

So another way to express this model is: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_e)$.

Notice we are conditioning on $x$ because we assume we “control” $x$ ($X$ is not a random variable). This is different from how we treated $X$ when we analyzed the correlation coefficient. (See Section 2.11)
Estimating $\sigma^2$:
We have already decided how we will estimate the parameters $\beta_0$ and $\beta_1$, but still need to be able to estimate $\sigma^2$. An unbiased (but model-dependent) estimator of $\sigma^2$ is found by “averaging” the squared residuals, i.e., the variance of the residuals.

$$\hat{\sigma}^2 = s^2 = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{(n-2)} = \frac{\text{SSE}}{n-2} = \text{MSE}$$

MSE is called the residual mean square or the mean square error. This (“model-based”) estimator is only useful if the model assumptions are met.

When these assumptions are reasonably met, we can make even more claims about our parameter estimators, e.g., $E(\hat{\beta}_1) = \beta_1$.

**Example 2:** Suppose we want to know whether students’ GPAs are related to how much they study. A student project group surveyed a random sample of 80 students from their large university and asked them their GPA and how many hours a week they study.

(a) Which variable do you consider the response variable? Do you think the relationship will be positive or negative? Strong or weak?

(b) Open the UOPgpa.txt data file and paste the two variables into the Two Quantitative Variables applet. Is the relationship positive or negative? Is it strong or weak? (How are you deciding?) What is the regression equation? What is the value of $s$?

(c) Because the students took the random sample from a larger population of students, they may be interested in deciding whether the relationship observed in the sample is strong enough to convince us that there is also a relationship in the population. Is it possible that there is no relationship between gpa and study hours in the population, but that they obtained a sample slope coefficient this large just by random chance? Explain.

So we need to determine whether “unlucky random sample” is a plausible explanation for the observed association in the sample (rather than there actually being a relationship in the population). We will explore this by replicating the randomness in the sampling process and seeing what kinds of sample slopes we find when we know there is no genuine association in the population.

First, let’s make a population similar to our sample that follows the basic model assumptions, but where we know there is no association between the two variables in the population.
(e) Determine the univariate statistics for each variable (scroll down in the applet and check the box that says Show descriptive). Record the following:

\[
\begin{align*}
y &= \text{gpa} & \text{mean } \bar{y} &= \quad \text{std dev } s_y = \\
x &= \text{study hrs} & \text{mean } \bar{x} &= \quad \text{std dev } s_x = 
\end{align*}
\]

(f) Now check the box that says Design Population and use the pull-down menu underneath to select Observed x. This will allow you to create a large population of 20,000 students with the same study hours, where there is no relationship gpa and study hours (\(\beta_1=0\)). To match the observed sample:

- The initial intercept has been set equal to the mean of the gpa variable. Because we are assuming \(\beta_1=0\), this says that the population regression line is constant at \(\bar{y}=3.24\).
- The \(x\) mean and \(x\) standard deviation are set to match the values found above.
- The value of \(\sigma\), representing the variability about the population regression line, is set to .45, the standard deviation of the residuals.
- Press the Create Population button to see the scatterplot of the population. (Play along, don’t worry about unrealistic values 😊.)
- Check the Show Regression Line box.

So the population we are creating here mostly matches the characteristics of the students’ sample data. The key difference here is that we are forcing the population slope to be equal to zero.

(g) Check the Show Sampling Options box, select the Slope as the statistic, and set the Sample Size from 20 to 80 to match our study. Click the Draw Samples button. Record the “Most Recent” sample regression line displayed in blue underneath the sample scatterplot. Did you obtain the same sample regression line as the students?

(h) Click the Draw Samples button again. Did you obtain the same sample regression line as in (g)?

Just as you saw in other courses that the value of a sample mean varies from sample to sample, now you see that the “value” of a sample regression line also varies from sample to sample (“sampling variability”) just due to random sampling from the population. The applet is collecting the different sample slopes on the right. Select the Intercept statistic and you will see the two sample intercepts as well. In order to make inferences about the population regression line, we need to analyze the long-run pattern of this variation, that is, the sampling distributions of the sample slope coefficients and the sample intercept coefficients.

You will now approximate these sampling distributions by taking a large number of random samples.

(i) Change the Number of Samples from 1 to 50. Press the Draw Samples button. Describe the pattern of variation that you see in the simulated regression lines. (Describe the lines as if to someone who can’t see the graph of regression lines you see.)
(j) To better see the long run pattern, ask for even more samples. Change the Number of Samples value from 50 to 948. This should bring you to 1000 total. Check the Total Samples output in the graph of the sampled slopes to make sure it is at least 1000.) Record the shape, center, and spread of each distribution (roughly) in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Distribution of $\hat{\beta}_1$</th>
<th>Distribution of $\hat{\beta}_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(k) Do the regression coefficients appear to be unbiased estimators of $\beta_0$ and $\beta_1$? (That is, do the sampling distributions center on the population values?)

(l) Now let’s consider factors that affect the sampling variability of the sample slopes. Make predictions:

1. If we decrease the number of students in the sample from 80 to 40. Do you think there will be more or less variability (or no change) in the distribution of the sample slopes?

2. If we decrease the variability about the regression line, $\sigma$, from .454 to .225, do you think there will be more or less variability (or no change) in the distribution of the sample slopes?

3. If we decrease the variability in the explanatory variable ($\sigma_x$) from 1.842 to .921, do you think there will be more or less variability (or no change) in the distribution of the sample slopes?

4. If we change the population mean of the explanatory variable ($\bar{X}$) from 3.928 to 1.964, do you think there will be more or less variability (or no change) in the distribution of the sample slopes?
(m) Now test each of your predictions by making the change in the applet (one at a time) and generating 1000 random samples. Each time you change a population setting, press the Create Population button. You should also note how (if) the population scatterplot changes each time. (You might want to sketch a graph without changing axis scaling. Change the value back and press Create Population before making the next change.) Focus on the summary statistics reported for each sampling distribution.

<table>
<thead>
<tr>
<th>Standard deviation of $\hat{\beta}_1$</th>
<th>Standard deviation of $\hat{\beta}_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{X} = 3.928, \sigma_x = 1.84, \sigma = .45, n = 40$</td>
<td></td>
</tr>
<tr>
<td>$\bar{X} = 3.928, \sigma_x = 1.84, \sigma = .225, n = 80$</td>
<td></td>
</tr>
<tr>
<td>$\bar{X} = 3.928, \sigma_x = .921, \sigma = .45, n = 80$</td>
<td></td>
</tr>
<tr>
<td>$\bar{X} = 1.964, \sigma_x = 1.84, \sigma = .45, n = 80$</td>
<td></td>
</tr>
</tbody>
</table>

(n) Discuss how your predictions compared to what actually happened.

So we can make the following observations:
- The distributions of sample slopes and intercepts are approximately normal.
- The means of the distributions of sample slopes and intercepts are $\beta_0$ and $\beta_1$ respectively.
- The variability in the sample slopes increases if we increase $\sigma$.
- The variability in the sample slopes increases if we decrease $\sigma_x$.
- The variability in the sample slopes decreases if we increase $n$.

Practice Problem (Questions o and p to be submitted in PolyLearn by 9am Wednesday)
- (o) Explain why each of the last 3 observations make intuitive sense (not algebraically, but considering how the population is changing and how that would affect variability in regression lines we might get from different samples).

(p) Regenerate the original sampling distribution. The students found a sample slope of ($\hat{\beta}_1 = 0.0894$). Where does the sample slope observed by the students fall in this simulated sampling distribution of slopes? Is it plausible that the population slope is really 0 and the students obtained a sample slope at least as large as .0894 just by chance? How often did such a sample slope occur in your 1000 samples?
(q) So if we were to derive a formula for the standard deviation of the sample slopes (the first column in the table above), which terms would be in the numerator and which in the denominator?

(r) You changed each term by a factor of 2, did the standard deviation always change by a factor of 2?

When all the model assumptions are met, it can be shown that:

\[ \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(n-1)s^2_x} \quad \text{and} \quad \text{Var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{x^2}{(n-1)s^2_x} \right) \]

(s) Use your calculator to calculate \( \text{Var}(\hat{\beta}_1) \) and compare the value to the result obtained in (j).

**Example 2 cont.**

(t) Change the population slope to .05 and press Create Population. How does the population scatterplot change? Conjecture how this will change the sampling distribution of sample slopes.

(u) Click Draw Samples and examine the sampling distribution of the sample slopes. Where is the center? Roughly how often did you get a sample slope as big as .089 or bigger? Does it seem plausible that the students’ regression line came from a population with \( \beta_1 = .05 \)?