Classwork 5: Model Assumptions

1. In classwork 4, we showed that $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$. Now, show that $\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}$. (Why am I asking you this? You'll see in the workbook chapter 6.) *(4 points)*

   Hint: note that the left hand side is the estimate for β_1 and the right hand side includes the true value of β_1. These will not be exactly equivalent except by chance. You should start this problem by making a substitution for y_i, since $y_i = \beta_0 + \beta_1 x_i + u_i$. This will get the true β_1 and u_i into the equation.

2. Models are really just sets of assumptions. Some of those assumptions are plausible, and others can be a little far-fetched. In this problem, we'll think about the six assumptions that comprise the OLS linear model.

   (a) We assume **the model is correctly specified and linear in parameters.** The models we can fit using OLS must be linear in *parameters* β, but don't have to be linear in *variables* x or y. For instance, we can fit this model: $log(y_i) = \beta_0 + \beta_1 x_i^2 + u_i$ but not this model: $y_i = \beta_0 + \beta_1 \beta_2 x_i + u_i$. To see why the second model would create issues, suppose we tried to fit it and got that $\hat{\beta}_0 = 5$ and $\hat{\beta}_1 \hat{\beta}_2 = 3$. Would it be possible to separately identify $\hat{\beta}_1$ from $\hat{\beta}_2$? Explain why or why not. *(1 point)*

   (b) We assume **there is some variation in** x. That is, `var(x)` can't be zero. Draw a scatterplot of some data where there is variation in y but no variation in x. Sketch a line of best fit and explain why it's important for there to be variation in x in order to get estimates for β_0 and β_1. *(1 point)*

   I'll talk about the third assumption in chapter 6 of the workbook, so for now I'll just present it without a detailed explanation.

   (c) We assume **exogeneity: the conditional expectation of** u **given** X **is zero:** $E[u_i|X] = 0$.

   The last three assumptions are not required for us to prove that OLS $\hat{\beta}_0$ or $\hat{\beta}_1$ are unbiased estimators for the parameters of the true model. These assumptions are required to use OLS standard errors with no adjustments. You'll study all of these in some depth when you take EC421, so I'll just present these here and leave it at that.

   (d) We assume u **is homoskedastic.**

   (e) We assume u_i **and** u_j **are independent for all** $i \neq j$**.**

   (f) We assume u_i **has a normal distribution for all observations** i**.**

3. Practice finding a regression's R^2: Open a new R script, make the first line `library(tidyverse)`, and make the second line this in order to read in the `students` dataset:

```
students <- read_csv("https://raw.githubusercontent.com/cobriant/students_dataset/main/students.csv")
```

Include a compiled R script that answers these questions:

3a) Is this data cross-sectional, time series, or panel data? *(1 point)*

3b) Fit the model $final\_grade_i = \beta_0 + \beta_1 failures_i + u_i$. Interpret the estimates for β_0 and β_1. *(1 point)*

3c) Pipe the lm object into the function `summary()` to read the regression R^2. It can be found in the second to last line under "Multiple R-squared". Interpret this number. *(1 point)*

3d) Fit some more models: instead of using `failures` as the explanatory variable, use `absences` or `grade1` or `grade2`. Which of these models yields the highest R-squared? Which yields the lowest R-squared? Does it make sense about why this would be the case when we think of higher R-squared indicating a stronger ability to predict the outcome `final_grade`? *(1 point)*