Classwork 15: Dummy Variables

1. Log Transformations and Dummy Variables

Consider this model, where D is a dummy variable that takes on either 0 or 1.

$$log(y) = \beta_0 + \beta_1 log(x) + \beta_2 D + u$$

We know that we can interpret β_1 as an elasticity, so if we estimate β_1 to be 0.2, that means a 1% increase in x can be associated with a 0.2% increase in y.

How do you interpret β_2, the coefficient on the dummy variable? Suppose we estimate β_2 to be 0.04. When D goes from 0 to 1, how much is y expected to increase by? Hint: start with

$$y = ax^{\beta_1} e^{D\beta_2} e^u$$

Then let y_1 be the expected outcome when $D = 0$, let y_2 be the expected outcome when $D = 1$, and find $\frac{y_2 - y_1}{y_1}$ as the percent by which y is expected to change.

Show your work above and also fill in the blank: When we estimate β_2 to be 0.04, that means that when D goes from 0 to 1, we can expect y to increase by ___ %.

2. Take `students` and estimate the model $final\_grade = \beta_0 + \beta_1 romantic + u$.

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

2a) What is the reference category and what is the dummy variable in your estimation?

2b) Conduct a hypothesis test: is there evidence that being in a romantic relationship affects a person's final grade?

3. Take `students` and estimate the model $final\_grade = \beta_0 + \beta_1 alcohol + u$.

3a) Why does `lm` create a set of four dummy variables when `alcohol` takes on five values: "very low", "moderately low", "medium", "moderately high", "very high"?

3b) Use `factor` to set "very low" to be the reference category. Estimate the model again and interpret the coefficients by completing these sentences:

All the coefficients are in comparison to the reference category "very low". So we'd expect that someone with very low alcohol consumption gets an average final grade of (___).

We'd expect that someone with "moderately low" alcohol consumption earns a final grade that is (higher/lower) than someone with very low alcohol consumption, by (___) points. That estimate (is/is not) statistically significant.

We'd expect that someone with "medium" alcohol consumption earns a final grade that is (higher/lower) than someone with very low alcohol consumption, by (___) points. That estimate (is/is not) statistically significant.

We'd expect that someone with "moderately high" alcohol consumption earns a final grade that is (higher/lower) than someone with very low alcohol consumption, by (___) points. That estimate (is/is not) statistically significant.

We'd expect that someone with "very high" alcohol consumption earns a final grade that is (higher/lower) than someone with very low alcohol consumption, by (___) points. That estimate (is/is not) statistically significant.

4) Estimate the larger model `final_grade ~ alcohol + sex + study_time + failures + romantic + absences` and interpret the coefficients by completing the sentences below.

Use `factor()` on alcohol and also on `study_time` to set the reference categories for those variables to be "very low" and "less than 2H". Do all your hypothesis tests at the .05 significance level.

The interpretation for $\hat{\beta}_0$ is that we'd expect a (male/female) student with () **alcohol consumption who studies (**) hours per week, who failed () **courses the previous year, who (is/is not) in a romantic relationship, and who has had (**) absences, will earn a (___) for a final grade.

The coefficients on the `alcohol` dummies all have the sign we'd expect: compared to someone with (___) alcohol consumption, anyone who drinks more than that is expected to earn a (higher/lower) final grade.

A (male/female) student is expected to earn a final grade that is (___) points (higher/lower), and that estimate (is/is not) statistically significant.

The coefficients on the `study_time` dummies all have the sign we'd expect: compared to someone who studies (___), anyone who studies more than that is expected to earn a (higher/lower) final grade.

Every failure is expected to (increase/decrease) a student's final grade by (___) points, and that estimate (is/is not) statistically significant.

Being in a romantic relationship is expected to (increase/decrease) a student's final grade by (___) points, and that estimate (is/is not) statistically significant.

Every absence is expected to (increase/decrease) a student's final grade by (___) points, and that estimate (is/is not) statistically significant.

5. The model we estimated in question 4 was linear in variables. Let's conduct the Ramsey RESET test to see whether there are any squared or interaction terms that would really improve the model predictions.

Show your work and also complete this sentence: The Ramsey RESET test (points to / does not point to) the existence of squared or interaction terms which should certainly be added to this model.