

Econ 31 Problem Set 5

Due on Wednesday, November 22nd, by 5:00pm

1. Suppose that a random sample of 200 twenty-year old men is selected from a population and that these men's height and weight are recorded. A regression of weight on height yields the following output:

Variable	Coef. Estimate	Std. Error	t-statistic	95% CI
Constant	-99.41	10.2	----	---- to ----
Height	3.94	1.62	----	---- to ----

- a) Fill in the blanks above.
2. The Wallingford-Swarthmore School District superintendent must decide whether to hire additional elementary school teachers (thus lowering the student-teacher ratio), and he wants your advice. He must compare the costs of hiring more teachers to the benefits, in terms of improving students' academic performance (as measured by test scores). You have a sample of 420 observations, with a student-teacher ratio and test score for each observation, and you plan to use OLS to provide him with an estimate of the "effect" of a change in the student-teacher ratio on students' test scores. You are willing to assume the following population-level regression model, where STR_i is the student-to-teacher ratio for individual student i :

$$Test\ Score_i = \beta_0 + \beta_1 \times STR_i + \epsilon_i$$

- a) What sign do you expect for your estimate of β_1 ? Why?
 - b) Now suppose Your OLS regression of test scores yields the following results: $b_0 = 698.91$, $b_1 = -2.28$, $se_{b_0} = 10.4$, $se_{b_1} = 0.52$, and $R^2 = 0.051$. Calculate the t-statistic for b_1 .
 - c) How would you expect average test scores to change if a district reduced its student-to-teacher ratio by 2 students?
 - d) Test the null hypothesis that the student-teacher ratio has no effect on $Test\ Score$, at a confidence level of 95% using a two-sided test (show the set-up of the test, too).
 - e) Calculate a 99% two-sided confidence interval for β_1 .
3. Consider the following linear regression model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

Explain exactly how you would test the following hypotheses:

- a.) $\beta_1 = 0$.
 - b.) $\beta_1 = 0$ and $\beta_4 = \beta_5$.
 - c.) $\beta_1 = 0$ and $\beta_3 = 2$ and $\beta_4 = \beta_5$.
4. Suppose we are interested in understanding how the number of skipped classes affects Econ 31 grades (measured as points out of 100). You collect data for a sample of 500 current and past Econ 31 students, and you run a multiple regression based on the following model:

$$Grade_i = \beta_0 + \beta_1 HS_GPA_i + \beta_2 Skipped_i + \beta_3 Stat11_i + \epsilon_i$$

where HS_GPA_i is a student's high school GPA, $Skipped_i$ is the number of classes a student skipped over the semester, and $Stat11$ is a dummy variable that equals 1 if the student had a previous statistics course, like Stat 11 (and equals 0 otherwise).

a) For each of your OLS estimates, b_1 , b_2 , and b_3 , explain how the estimate should be interpreted and whether you expect coefficient to be positive or negative. (Be careful not to offend your professor!)

Next, you modify your regression model slightly to include a control for gender and an interaction term between the *Skipped* and *Stat11* variables:

$$Grade_i = \beta_0 + \beta_1 HS_GPA_i + \beta_2 Skipped_i + \beta_3 Stat11_i + \beta_4 Skipped \times Stat11_i + \beta_5 Female_i + \epsilon_i$$

where all variables are defined as they were above.

You obtain the following results:

Parameter	OLS Estimate	Standard Error
Intercept	30.0	12.0
HS_GPA	15.0	6.0
Skipped	-1.5	0.2
Stat11	10.0	4.0
SkippedXStat11	0.8	0.25
Female	1.0	0.75

- b) For each slope estimate, b_1 through b_5 , test whether the coefficient is statistically significant.
- c) Calculate the predicted Econ 31 grade for a female student who skipped 3 classes, had a high school GPA of 3.5, and had not taken Stat 11.
- d) What does the estimate of β_4 tell us? (Precisely interpret it.) Hint: What is the predicted “effect” of an additional skipped class on Econ 31 grade?
- e) Suppose for the initial model (in part a), $R^2 = 0.32$, and for the expanded model, $R^2 = 0.40$. Show how you can test whether the three additional two variables *Skipped11* and *Female* are jointly significant.

Applications with Stata

For this work, please download the data set on Moodle, called MarchCPSworkers2020.dta. Then write a .do file that creates a .log file that does the following:

5. First, let's start with some data management tasks:

- a. Create a variable called **hourly_wage** by dividing the individual's total wage and salary income by the product of their usual hours worked per week last year and their number of weeks worked last year.
 - b. Create a variable called **lnwage**, which equals the natural log of the individual's hourly wage.
 - c. Generate a dummy variable called **nonwhite**, which equals 1 if the individual's reported race is anything other than "White."
 - d. Create a dummy variable called **college**, which equals 1 if the individual has at least a bachelor's degree.
6. Estimate a simple regression of **hourly_wage** on the individual's age (measured in years).
 - a. Interpret your coefficient estimate, b_1 , in words.
 - b. Do your results suggest the relationship between *age* and *hourlywage* is statistically significant (using $\alpha = 0.01$)? Explain how you know.
7. Estimate a simple regression of **lnwage** on the individual's age (measured in years).
 - a. Interpret your coefficient estimate, b_1 , in words.
 - b. Do your results suggest the relationship between *age* and *lnwage* is statistically significant (using $\alpha = 0.01$)? Explain how you know.
8. To assess the relationship between education and wage, you are asked to develop a regression model using log hourly wage as the dependent variable: $\ln wage = \beta_0 + \beta_1 Age + \beta_2 Female + \beta_3 Education + \epsilon_i$
 - a. Briefly explain why it is NOT appropriate to use *educ* directly as an independent variable in the regression model specified above.
 - b. How many dummy variables would you need if you would like to analyze how the different levels of education affect wage? How many could be included in your regression?
 - c. Instead of using a full set of dummy variables, use the variable *college* as a rough indicator for education. Estimate the regression model specified above, but replace *Education* with a dummy variable for college degree or more. *gender* (female) and the dummy variable for education level (*college*). Write out the estimated sample regression equation.
 - d. Interpret the estimated regression coefficients. Can you interpret the constant term in a meaningful way? (Note that the dependent variable is in logs.)
 - e. Based on your analysis, can you conclude that there is a difference between the (log of) hourly earnings of male and female employees, controlling for their age and education levels?
 - f. What is the R^2 value for this regression? Interpret it.
9. Next, add dummy variables for race to the model above and estimate this multiple linear regression. The variables representing four mutually exclusive, broad race categories in this sample are **white**, **black**, **asian**, and **other**.
 - a. Why can't you include all 4 race dummy variables to your model? What happens if you try it? Leave out the variable **white** and re-estimate the model.
 - b. Run an F test of whether the race variables are jointly significant. Interpret your result.

Note: If we don't get to F-tests by next Monday, you may omit this last part!