Quantitative Methods – Assignment #2

Winter 2021/2022

Due date: December 2, 2021

This assignment is due Thursday, December 2, 2021, by 6pm. Submit your assignment through Slack to Felix Hagemeister or Timm Betz, depending on your tutorial. You are encouraged to work in groups. However, you must hand in your own work, written in your own words. All answers must be in complete sentences. You do not have to provide R code. Providing R code only, without written answers, is not sufficient. Make sure to write your name on top of your homework. **Please us the following file name for your assignment:** `lastname_QM_HW2`.

1. The file `QM_HW2_data.csv` contains data from the California Standardized Testing and Reporting data set (STAR). The data set contains variables on the performance on standardized tests, school characteristics, and demographic information on the students. The file contains data for 420 school districts, for 1998 and 1999. The variable `testscr` is an average of reading and math scores on a standardized test (also included separately as `read_scr` and `math_scr`, respectively). Observations are uniquely identified by district code, `dist_cod`. School districts are (usually) smaller than counties; counties are roughly the equivalent of a German 'Landkreis'. School characteristics and demographic information are averaged across schools in each school districts.

- Estimate a linear regression model to evaluate whether the student teacher ratio, `str` (the number of students divided by number of teachers), predicts test scores. Report the results.
- Interpret the results for the slope estimate $\widehat{\beta}$: What is the (not necessarily causal) effect of a one-unit increase in the student-teacher ratio? Would you consider this effect substantively important or meaningful?
- What is the predicted difference in test scores between the district with the lowest student-teacher ratio and the district with the highest student-teacher ratio? Put differently, what is the effect of increasing the student-teacher ratio from the sample minimum to the sample maximum?
- Moving from the sample minimum to the maximum is a bit extreme. And often, there are only few observations near the tails of the distributions. For many continuous variables, we instead often think of the 'typical' change of the variable across observations in our data set as a one standard deviation increase. That is, if a variable has a standard deviation of 2.5 in our data set, we could report the effect of a one standard deviation (and thus, in this case, a 2.5 unit) increase in the variable. In our data set, what is the effect of a one standard deviation increase in the student-teacher ratio?[*]

[*] This interpretation in terms of standard deviations also has the advantage that the reported effects are independent of scale: if you multiply your variable by any constant (say, you report income in $1,000 or in $; more generally, you could also use any linear transformation, such as using Fahrenheit instead of Celsius), the interpretation of a one-standard deviation increase remains invariant to this scaling or the unit of measurement.

- Create a new variable, `str_norm`, as `str` divided by its standard deviation. This is the student-teacher ratio normalized by its standard deviation. Then, use this new variable as predictor of test scores. Interpret the results and explain how the new results compare to the previous finding for the effect of a one standard deviation increase.

- Using the formula

  $$\widehat{\beta} = \beta + \frac{\widehat{\text{Cov}}(x_i, \epsilon_i)}{\widehat{\text{Var}}(x_i)}, \tag{1}$$

  where the 'hats' indicate that these are the sample covariance and variance, and re-arranging we can write our bias as

  $$\text{bias}(\widehat{\beta}) = \frac{\text{Cov}(x_i, \epsilon_i)}{\text{Var}(x_i)} \tag{2}$$

  where now we use the true, or population, covariance and variance.

  We know from prior work that the average income of parents is an important predictor of student success in schools and thus of standardized test scores. Under what conditions is not accounting for the average income of parents a source of endogeneity in our model, resulting in biased estimates of the effect of the student-teacher ratio?

  Using the above formula, what do you think this bias is likely doing to your estimates – is it resulting in estimates that are too low (i.e., our estimates were more negative than what they 'should have been' because of the bias) or in estimates that are too high (i.e., our estimates were less negative because of the bias)? Put differently, if we were able to somehow account for differences in average income, do you expect your estimate would become more negative or less negative?

2. Answer the following multiple choice questions:

a. Even if we know that our independent variable x_i is endogenous, we can use OLS for making predictions.

- true
- false

b. x_i is the slope coefficient in the following model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

- true
- false

c. For an unbiased estimator of some parameter β_1, we know that our estimate $\widehat{\beta_1}$ will be exactly identical to the true β_1.

- true
- false

d. An predictor variable is endogenous if

- it is correlated with another variable
  that is included in the model
- it is correlated with the error term of the outcome
- it is correlated with the outcome
- it is not correlated with anything

e. Consider the follwing model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Using OLS, we estimate that $\widehat{\beta_1} = 10$. Therefore, on average, a one unit increase in x is expected to lead to a

- hundred unit increase in y
- one unit increase in y
- one tenth of a unit increase in y
- ten unit increase in y