Econ 322: Econometrics

3rd Assignment: Multiple Linear Regression

Due Date: 23:00, 10/31/2023, EST

Read the following instructions carefully.

- You can submit your reports unlimited times before the deadline. Do wait until the last minute and tell me that you have an internet problem.
- Report the required results in a word or PDF document and submit it **together with your code**. In the report, make your answers concise and to the point. My TA has a lot of submissions to grade. So make sure that your answers are easy to spot, in case my TA may miss them.
- Name your submitted files in the following format: initial first_full last name_A3 (so that we know this is your answer to the second assignment).
- You can refer to the sample codes I gave in my lecture notes and the uploaded Python codes.

This assignment is designed to help you transit from the simple linear regression model to the multiple one. You will get a more concrete idea on the omitted variable bias by comparing these two regression results. In addition, you will how to avoid the multicolinearity problem when adding dummy variables into the regression.

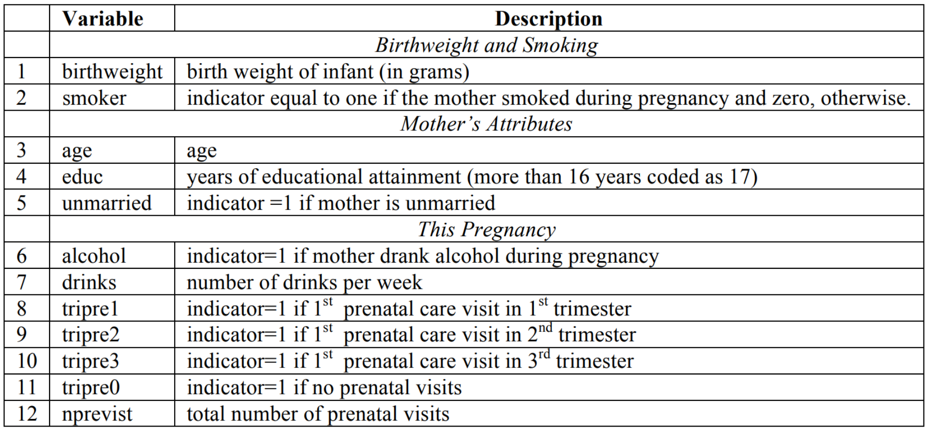
In the dataset **Birthweight_Smoking.csv**, which contains data for a random sample of babies born in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy. A detailed description is given in the table below.

You can continue to work on the .py file in Spyder or the .ipynb file in Juptyer Notebook (or Jupyter Lab).

Unless instructed otherwise, round up the numbers to the 2nd decimal point. For small numbers like 0.0123, you can round it up to the 3rd decimal point.

Q1 **(50 points)**. Regress *Birthweight* on *Smoker* (make sure you find the correct column names in the data). What is the estimated effect of smoking on birth weight? (This is a question that suppose to give some "free" points given what you have learned so far. Make sure you can get the points.)

Q2 **(40 points)**. Regress *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*.



	Variable	**Description**
		Birthweight and Smoking
1	birthweight	birth weight of infant (in grams)
2	smoker	indicator equal to one if the mother smoked during pregnancy and zero, otherwise.
		Mother's Attributes
3	age	age
4	educ	years of educational attainment (more than 16 years coded as 17)
5	unmarried	indicator =1 if mother is unmarried
		This Pregnancy
6	alcohol	indicator=1 if mother drank alcohol during pregnancy
7	drinks	number of drinks per week
8	tripre1	indicator=1 if 1^{st} prenatal care visit in 1^{st} trimester
9	tripre2	indicator=1 if 1^{st} prenatal care visit in 2^{nd} trimester
10	tripre3	indicator=1 if 1^{st} prenatal care visit in 3^{rd} trimester
11	tripre0	indicator=1 if no prenatal visits
12	nprevist	total number of prenatal visits

(a) Explain why the exclusion of *Alcohol* and *Nprevist* could lead to omitted variable bias in the regression estimated in Q1.

(b) Is the estimated effect of smoking on birth weight substantially different from the regression that excludes *Alcohol* and *Nprevist*? Does the regression in Q1 seem to suffer from omitted variable bias?

(c) Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child. Round your answer to an integer, which is already precise enough.

(d) Find R^2 and $\bar{R}^2$ (adjusted R-squared). Why are they so similar?

(e) How should you interpret the coefficient (i.e., slope) on Nprevist? Does the coefficient measure a causal effect of prenatal visits on birth weight? If not, what does it measure?

Q3 **(10 points)**. An alternative way to control for prenatal visits is to use the binary variables Tripre0 through Tripre3. Regress *Birthweight* on *Smoker*, *Alcohol*, *Tripre0*, *Tripre2*, and *Tripre3*. For your information, there are only three trimesters (note this word has the prefix tri-). Therefore, there are only four possibilities: no visits at all, started visit in the 1st trimester, stared in the 2nd, and started in the 3rd.

(a) Why is *Tripre1* excluded from the regression? What would happen if you included it in the regression? (Hint: Collinearity)

(b) Does the regression in Q3 explain a larger fraction of the variance in birth weight than the regression in Q2?