Econ 322: Econometrics

4th Assignment: Dummy Variable

Due Date: 23:00, 11/29/2023, 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_A4 (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 understand more about dummy variable: how to interpret their regression slopes when they are used as regressors, and what models you can use when the dependent variable is a dummy.

Q1 **(50 points)**. In the dataset **CASchools.csv**, you will find some collected data of schools in California. We are interested in whether smaller class size can improve students' test performance.

First, create a dummy variable *NS* (not small) based on the student to teacher ratio *str_s*. More specifically, let *NS* be 1 if the str ratio is greater than 25.

Next create other dummy variables using the following code (if you directly copy this into Spyder, the quotation marks may be transform into something else, hence can not be recognized. You need to replace them with the correct ones.).

```
BayAreaCounties = ["San Francisco", "San Mateo", "Santa Clara",
                   "Contra Costa", "Marin", "Alamda", "Solano",
                   "Sonoma", "Napa", "Santa Cruz", "San Benito"]

df['BayArea'] = df['countyname'].apply(lambda text: text in BayAreaCounties)
df['NS_Bay'] = df['NS'] * df['BayArea']
df['NS_NonBay'] = df['NS'] * (1-df['BayArea'])
df['S_Bay'] = (1-df['NS']) * df['BayArea']
df['S_NonBay'] = (1-df['NS']) * (1-df['BayArea'])
```

(a) In the sample, What is the average value of *testscore* for all schools? For those that with a relatively large class size (NS = 1)? For those with small class size (NS = 0)? What is the difference in the averages of these two subgroups (large-class minus small-class)?

(b) Run a regression of *testscore* on the binary variable *NS*. Explain how the estimated slope and intercept are related to your answers in part (a). Is the estimated slope statistically significant at 5% level? Construct a 95% confidence interval for the effect of large class size on test score. Round the numbers up to the 3rd decimal point.

(c) Regress *testscore* on *S_Bay*, *NS_NonBay*, *NS_Bay*. Does (large) class size have a significant effect for schools that are not in the Bay Area? What about for schools that are in the Bay Area? Report t statistics (in absolute value) and p-values to justify your answer.

Q2 **(50 points)** The dataset **Employment_08_09.csv** contains a random sample of 5440 workers who were surveyed in April 2008 and reported that they were employed full-time. These workers were surveyed one year later, in April 2009, and asked about their employment status (employed, unemployed, or out of the labor force). The data set also includes various demographic measures for each individual. Use these data to answer the following questions.

Variable Name	Description
	Variables from the 2009 Survey
employed	indicator =1 if employed in 2009
unemployed	indicator =1 if unemployed in 2009
	Variables from the 2008 Survey
age	age
female	indicator =1 if female
married	indicator =1 if
race	= 1 if self-identified race = white (only) = 2 if self-identified race = black (only) = 3 if self-identified race was not white (only) or black (only)
union	indicator =1 if a member of a union
ne_states	indicator =1 if from a northeastern state
so_states	indicator =1 if from a southern state
ce_states	indicator =1 if from a central state
we_states	indicator =1 if from a western state
private	indicator =1 if employed in a private firm
government	indicator =1 if employed by the government
self	indicator =1 if self-employed
educ_lths	indicator =1 if highest level of education is less than a high school graduate
educ_hs	indicator =1 if highest level of education is a high school graduate
educ_somecol	indicator =1 if highest level of education is some college
educ_aa	indicator =1 if highest level of education is AA degree
educ_ba	indicator =1 if highest level of education is BA or BS degree
educ_adv	indicator =1 if highest level of education is advanced degree
earnwke	average weekly earnings

(a) Regress *employed* on *age* and age^2 (you need to create this variable), using a linear probability model. Based on this regression, was age a statistically significant determinant of employment in April 2009? Perform a formal F-test and report the F statistics and p-value (up to 3rd decimal point). Is there evidence of a nonlinear effect of age on the probability of being employed? Compute the predicted probability of

employment for a 20-year-old worker, a 40-year-old worker, and a 60-year-old worker. Round the predicted probability to 2nd decimal point.

(b) Repeat Part (a) using a logit regression and answer the same questions.

Hint: Suppose you defined square age as a new column *age2*. You can predict with new values using the following command with a given estimation result:

```
print(results.predict(exog={'age':20, 'age2':400}))
```

This will be extremely helpful when you use the logit model.