Problem Set 2

This problem set will take you through some Stata commands to estimate simple regression equations with dummy variables. You will learn how to interpret the estimated coefficients and test some linear hypotheses. Interpretation of these coefficients will be useful when we do treatment evaluation models later in term 1.

The hypothesis tests discussed in this problem set include standard T-tests and F-tests, which is assumed undergraduate knowledge for this module.

You will need to download the dataset problem-set-2-data.dta, which is available on Moodle.

Conditional Expectation Function

Consider the Conditional Expectation Function (CEF), $E[Y_i|X_i]$. If X takes on discrete values: $X_i \in \{x_1, x_2, ..., x_m\}$, then

$$E[Y_i|X_i] = E[Y_i|X_i = x_1] \cdot \mathbf{1}\{X_i = x_1\} + ... + E[Y_i|X_i = x_m] \cdot \mathbf{1}\{X_i = x_m\}$$

where $\mathbf{1}\{X_i = x_m\}$ is a dummy variable, $= 1$ when $X_i = x_m$. Since the values of X_i are mutually exclusive there is no overlap of these dummy variables.

Note, we do not need to assume that X is a single random variable. It can be a vector of random variables that takes on discrete values.

We can re-arrange this expression using anyone of the values of X. The natural option is to choose the first, but this is arbitrary.

$$\begin{aligned} E[Y_i|X_i] =& E[Y_i|X_i = x_1] + (E[Y_i|X_i = x_2] - E[Y_i|X_i = x_1]) \cdot \mathbf{1}\{X_i = x_2\} + ... \\ &+ (E[Y_i|X_i = x_m] - E[Y_i|X_i = x_1]) \cdot \mathbf{1}\{X_i = x_m\} \end{aligned}$$

Since, $E[Y_i|X_i = x_m]$ is a constant (X_i is set to a specific value), we can express the CEF as a function that is linear in parameters.

$$E[Y_i|X_i] = \beta_1 + \beta_2 D_{i2} + ... + \beta_m D_{im}$$

where $D_{im} = \mathbf{1}\{X_i = x_m\}$.

Preamble

```
<IPython.core.display.HTML object>
```

Create a do-file for this problem set and include a preamble that sets the directory and opens the data. For example,

```
clear
//or, to remove all stored values (including macros, matrices, scalars, etc.)
*clear all

* Replace $rootdir with the relevant path to on your local harddrive.
cd "$rootdir/problem-sets/ps-2"

cap log close
log using problem-set-2-log.txt, replace

use problem-set-2-data.dta
```

Questions

1. Consider the $E[ln(Wage_i)|Gender_i]$, where $Gender_i \in \{1 \text{`` } Male'', 2 \text{`` } Female''\}$. Show that this CEF implies a linear model,

$$ln(Wage_i) = \beta_1 + \beta_2 D_{i2} + \varepsilon_i$$

What do the parameters β_1 and β_2 imply?

2. Regress `lwage` (log wage) on just a set of binary indicators that will enable you to test the hypothesis that males and females are on average, paid the same wage, ceteris paribus. Test this hypothesis.

3. Extend the specification in (2) that will enable you to test the hypothesis that there is no difference in the wages between the following gender-ethnicity groups. Begin by defining the following dummy variables:

- `female_black` = `female`×`black`

- `male_black` = (1-`female`)×`black`
- `female_nonblack` = `female`×(1-`black`)
- `male_nonblack` = (1-`female`)×(1-`black`)

Then estimate the following regressions:

a. `lwage` on `female_black`, `female_nonblack`, `male_black`, `male_nonblack` (without a constant: option `nocons`)

b. `lwage` on `female`, `black`, `female_black`

c. `lwage` on `female_black`, `female_nonblack`, `male_black`

For some of these exercises you may be able to use Stata's factor notation. However, in some instances you will need to manually create the above dummy-variable interactions.

In each case, identify the base category and write down the parameters of the (implied) model in terms of conditional expectations.

4. Compare the estimated coefficients with the sample average values for the lwage for the four subgroups. What do you see?

5. In each of the above models, describe the null hypothesis you would test to evaluate whether there is a significant earnings difference between the earnings of black and non-black females.

6. Very your solution to 4. by performing a test using the three set of regression output. You can use the post-estimation `test` command.

7. In each case, test equality across all four gender-ethnicity groups. Again, you should get the same result.

8. Try to replicate the F-statistic for one of the above models. Hint, the F-stat for these models is the same as that of the whole model.

9. Estimate the following model:

$$lwage = \beta_1 + \beta_2 F + \beta_3 B + \beta_4 F \times B + \beta_5 exp + \beta_6 exp^2 + \beta_7 educ + \varepsilon$$

- Interpret the estimated coeffiecients $\hat{\beta}_7$.
- Interpret the effect of experience variable `exp`. Use the median level of experience to make your calculation.
- A one unit incease in years of education is associated with an increase of 1.78% in expected wages, holding other regressors fixed.

10. Theoretically, how would you test the following restrictions for the model below?

a. $\beta_2 = \beta_3$

b. $\beta_4 + \beta_5 = 1$

c. $\beta_2 = \beta_3$ **and** $\beta_4 + \beta_5 = 1$

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Postamble

```
log close
```