ECON 113 DI A/B

Dizhi Wang

OH : Thursday 8:30am – 10:30 pm E2 403B

1. What is wrong with the following STATA code? How would you correct it?

```
postfile buffer beta_0_hat beta_1_hat using mcs, replace

forvalues i=1/100 {
        quietly drop _all
        set seed 12345
        quietly set obs 1000
        quietly generate y = 3 + 5*x + u
        quietly generate x = rchi2(1)
        quietly generate u = rnormal()
        quietly reg y x
        scalar beta_1= _b[x]
        scalar beta_0= _b[_cons]
        post buffer (beta_0) (beta_1)
}

postclose buffer
use mcs, clear
```

1. What is wrong with the following STATA code? How would you correct it?

```
postfile buffer beta_0_hat beta_1_hat using mcs, replace

forvalues i=1/100 {
        quietly drop _all
        set seed 12345
        quietly set obs 1000
        quietly generate y = 3 + 5*x + u
        quietly generate x = rchi2(1)
        quietly generate u = rnormal()
        quietly reg y x
        scalar beta_1= _b[x]
        scalar beta_0= _b[_cons]
        post buffer (beta_0) (beta_1)
}

postclose buffer
use mcs, clear
```

You are creating a different population every time instead of drawing different samples from the same population

Stata has no idea what x or u is

Answer:

Set seed 12345 should be put outside of the loop
Quitetly generate y = 3+5*x+u should be put after x and u were generated

2. Let $Y = \beta_0 + \beta_1 X + U$. Explain why it does not make sense to compute neither $E[\beta_0]$ nor $Var[\beta_0]$.

What is β_0 or β_1 here? Are they random variables which have distributions?

2. Let $Y = \beta_0 + \beta_1 X + U$. Explain why it does not make sense to compute neither $E[\beta_0]$ nor $Var[\beta_0]$.

What is β_0 or β_1 here? Are they random variables which have distributions?

Answer:

Because β_0 and β_2 are fixed (non-random) parameters, i.e. constants, they do not have a probability distribution and it is not meaningful to compute these.

3. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS estimators of a regression of y on x. That is, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Suppose that you run a regression of y on x^*, where $x^* = \frac{x}{a}$ for a some constant. Call these OLS estimators $\tilde{\beta}_0$ and $\tilde{\beta}_1$. How do they relate to $\hat{\beta}_0$ and $\hat{\beta}_1$? Explain why the R^2 does not change.

Important formula: $\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{\bar{x}})^2}$ and $\bar{y} = \widehat{\beta_0} + \widehat{\beta_1}\bar{x}$

Sub any new defined variables to these formula (for this question, the only new defined var is x*

3. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS estimators of a regression of y on x. That is, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Suppose that you run a regression of y on x^*, where $x^* = \frac{x}{a}$ for a some constant. Call these OLS estimators $\tilde{\beta}_0$ and $\tilde{\beta}_1$. How do they relate to $\hat{\beta}_0$ and $\hat{\beta}_1$? Explain why the R^2 does not change.

Given $x^* = \frac{x}{a}$, one has: $\bar{x^*} = \frac{1}{n}\sum_i x_i^* = \frac{1}{n}\sum_i \frac{x_i}{a} = \frac{1}{a}\frac{1}{n}\sum_i x_i = \frac{\bar{x}}{a}$

The OLS estimator $\tilde{\beta_1}$ for the regression y on x^* will be:

$$\begin{aligned}
\tilde{\beta_1} &= \frac{\sum_i (y_i - \bar{y})(x_i^* - \bar{x^*})}{\sum_i (x_i^* - \bar{x^*})^2} \\
&= \frac{\sum_i (y_i - \bar{y})\left(\frac{x}{a} - \frac{\bar{x}}{a}\right)}{\sum_i (\frac{x}{a} - \frac{\bar{x}}{a})^2} \\
&= \frac{\frac{1}{a}\sum_i (y_i - \bar{y})\left(x - \frac{\bar{x}}{a}\right)}{\frac{1}{a^2}\sum_i (x - \bar{x})^2} \\
&= a\frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\
&= a\hat{\beta_1}
\end{aligned}$$

Thus, $\hat{\beta}_0 = \bar{y} - \tilde{\beta}_1 \overline{x^*} = \bar{y} - a\hat{\beta}_1 \frac{\bar{x}}{a} = \hat{\beta}_0$.

The R^2 has to stay the same, since there cannot be any gain in fit by just changing the scale of the observations. To be clear, let's use an example - a regression of GPA on Instagram use:

4. Suppose the model for the outcome is $y = \beta_1 x + u$. When you regress y on x without a constant, the OLS estimator is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(a) Show that $\hat{\beta}_1$ is an unbiased estimator of β_1, that is, $E[\hat{\beta}_1] = \beta_1$.

(b) Explain what does it mean that $\hat{\beta}_1$ is an unbiased estimator of β_1?

Whenever you see y_i or, dependent variable, try to sub the model:$y = \beta_0 + \beta_1 x + u$, which is $y_i = \beta_0 + \beta_1 x_i + u_i$ for observation-wise

4. Suppose the model for the outcome is $y = \beta_1 x + u$. When you regress y on x without a constant, the OLS estimator is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(a) Show that $\hat{\beta}_1$ is an unbiased estimator of β_1, that is, $E[\hat{\beta}_1] = \beta_1$.

(b) Explain what does it mean that $\hat{\beta}_1$ is an unbiased estimator of β_1?

Answer:

a)

$$\begin{aligned}
E[\hat{\beta}_1] &= E[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}] \\
&= E[\frac{\sum_{i=1}^{n} x_i(\beta_1 x_i + u_i)}{\sum_{i=1}^{n} x_i^2}] \\
&= E[\frac{\sum_{i=1}^{n} x_i \beta_1 x_i + \sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}] \\
&= E[\frac{\beta_1 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}] \\
&= E[\beta_1 + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}]
\end{aligned}$$

Because $\sum_{i=1}^{n} x_i u_i = 0$, one can have $E[\hat{\beta}_1] = E[\beta_1] = \beta_1$

b) If we draw different samples from the population infinitely many times and use the OLS estimator to estimate $\hat{\beta}_1$ for each sample, then take the average of all the estimated $\hat{\beta}_1$ values from different samples, the result will be the true β_1.

5. You gather data for a sample of different countries about average weekly working hours, denoted Y and GDP, denoted X. The following model is estimated using OLS:

$$y = \beta_0 + \beta_1 \log x + u$$

(a) Argue whether you expect the **unknown parameter** β_1 to be positive or negative. What is the economic interpretation of β_1?

(b) Provide one example of a factor contained in u that might be **correlated with GDP**.

(c) Suppose the regression yields $\hat{\beta}_1 = -22$. Interpret this estimate carefully. Would you trust the OLS estimator in this setting?

Answer:

(a) Given that people value their leisure, an increase in GDP (X) is likely to reduce working hours (**Y**), so β_1 is expected to be negative.

For example, if $\beta_1 = -3$, then in a linear-log model, a 1% increase in GDP is associated with a decrease of 0.03 hours in average weekly working hours $\beta_1/100 = -3/100 = -0.03$.

If people instead disliked leisure, β_1 could be positive. The key point is that, in this example, leisure preferences are assumed constant across countries.

5. You gather data for a sample of different countries about average weekly working hours, denoted Y and GDP, denoted X. The following model is estimated using OLS:

$$y = \beta_0 + \beta_1 \log x + u$$

(a) Argue whether you expect the **unknown parameter** β_1 to be positive or negative. What is the economic interpretation of β_1?

(b) Provide one example of a factor contained in u that might be **correlated with GDP**.

(c) Suppose the regression yields $\hat{\beta}_1 = -22$. Interpret this estimate carefully. Would you trust the OLS estimator in this setting?

(b) Regional differences, which are unobserved and therefore included in the error term (u), can affect leisure preferences across countries. Unlike part (a), people's preference for leisure may vary by region — for example, countries with higher GDP may include regions like San Jose where people are more eager to work, while other regions, like Santa Cruz, may prefer to be more relaxed. Since our dataset does not include regional information, these unobservable differences are absorbed into u. Because regions with stronger work preferences tend to be in countries with higher GDP, u is likely correlated with X.

5. You gather data for a sample of different countries about average weekly working hours, denoted Y and GDP, denoted X. The following model is estimated using OLS:

$$y = \beta_0 + \beta_1 \log x + u$$

(a) Argue whether you expect the **unknown parameter** β_1 to be positive or negative. What is the economic interpretation of β_1?

(b) Provide one example of a factor contained in u that might be **correlated with GDP**.

(c) Suppose the regression yields $\hat{\beta}_1 = -22$. Interpret this estimate carefully. Would you trust the OLS estimator in this setting?

(b) Regional differences, which are unobserved and therefore included in the error term (u), can affect leisure preferences across countries. Unlike part (a), people's preference for leisure may vary by region — for example, countries with higher GDP may include regions like San Jose where people are more eager to work, while other regions, like Santa Cruz, may prefer to be more relaxed. Since our dataset does not include regional information, these unobservable differences are absorbed into u. Because regions with stronger work preferences tend to be in countries with higher GDP, u is likely correlated with X.

5. You gather data for a sample of different countries about average weekly working hours, denoted Y and GDP, denoted X. The following model is estimated using OLS:

$$y = \beta_0 + \beta_1 \log x + u$$

(a) Argue whether you expect the **unknown parameter** β_1 to be positive or negative. What is the economic interpretation of β_1?

(b) Provide one example of a factor contained in u that might be **correlated with GDP**.

(c) Suppose the regression yields $\hat{\beta}_1 = -22$. Interpret this estimate carefully. Would you trust the OLS estimator in this setting?

(c) Based on part (a), we can interpret $\hat{\beta}_1 = -22$ as meaning that a 1% increase in GDP is associated with a decrease of 0.22 hours in average weekly working hours ($-22 \times 0.01 = -0.22$ hours). However, the OLS estimator is not fully trustworthy in this setting. As mentioned in part (b), unobserved differences in leisure preferences by region are included in the error term u and may be systematically related to GDP, which means

$$E[u \mid X] \neq 0.$$

This violates the OLS assumption $E[u \mid X] = 0$ and leads to biased estimates.