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

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
```

There are two errors here: first "set seed 12345" is inside the for loop, so that it renders the for loop meaningless. Second, in the line "quietly generate y = 3 + 5*x + u" the code calls for x and u before they are defined. The correct code is:

```
set seed 12345
postfile buffer beta_0_hat beta_1_hat using mcs, replace

forvalues i=1/100 {
        quietly drop _all
        quietly set obs 1000
        quietly generate x = rchi2(1)
        quietly generate u = rnormal()
        quietly generate y = 3 + 5*x + u
        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
```

2. 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) How are the fitted values, $\hat{y}_i$, and the residuals, $\hat{u}_i$ constructed?

Given this new estimator (OLS no constant estimator), what is the corresponding fitted values (or estimator predicted values) and the corresponding residuals?

2. 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) How are the fitted values, $\hat{y}_i$, and the residuals, $\hat{u}_i$ constructed?

Solution: The fitted values are computed as $\hat{y}_i = \hat{\beta}_1 x_i$. The residuals are computed as $\hat{u}_i = y_i - \hat{y}_i$.

Note there is no $\hat{\beta}_0$ because this is a regression without an intercept/constant.

2. When you regress y on x without a constant, the OLS estimator is

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

(b) Find the variance of $\hat{\beta}_1$.

Variance of $\widehat{\beta_1} = var(\widehat{\beta_1}) = var\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2}\right)$

Some tricks of solving econometrics problem:

1) Sub y_i using the model formula, which is $y_i = \beta_0 + \beta_1 x_i + u$.
2) Add or Minus $\bar{x}, \bar{y}$
3) Sub $\widehat{y_i}$ using the estimating formula, which is $\widehat{y_i} = \widehat{\beta_1} x_i$ in this question's case but it is usually $\widehat{y_i} = \widehat{\beta_1} x_i + \widehat{\beta_0}$

Try your luck with these three tricks on any problem that seems way too hard!

2. When you regress y on x without a constant, the OLS estimator is

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

(b) Find the variance of $\hat{\beta}_1$.

Solution: We proceed as in the case we studied in class. Note that we can write:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \beta_1 + \frac{\sum_{i=1}^n x_i u_i}{\sum_{i=1}^n x_i^2}.$$

Let

$$\omega_i = \frac{x_i}{\sum_{i=1}^n x_i^2}$$

so that $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n \omega_i u_i$. Then, using properties of variance:

$$Var\left[\hat{\beta}_1\right] = \sum_{i=1}^n \omega_i^2 Var[u_i].$$

If we assume homoskedasticity, then $Var[u_i] = \sigma^2$ for every i, then

$$Var\left[\hat{\beta}_1\right] = \sigma^2 \sum_{i=1}^n \omega_i^2 = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

2. When you regress y on x without a constant, the OLS estimator is

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

(c) Show that, in this case, the mean of the residuals $\hat{u}_i$ is not always 0.

We excluded this question when grading because it is out of the scope, but the right answer for this question will be as follows:

For OLS estimator with a constant, we have two unknowns, β_0 and β_1, and we therefore need two conditions, which we force to be $\sum_i \hat{u}_i = 0$, and $\sum_i x_i \hat{u}_i = 0$
When we discard the constant, we only have one unknown β_1 so we only impose $\sum_i x_i \hat{u}_i = 0$, which is $\sum_i \hat{u}_i = 0$ cannot be ensured.

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: $\widehat{\beta_1} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$ 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. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(a) What sign do you expect the **unknown** parameter β_1 to have? What is the interpretation of β_1?

Take whatever sign you like to but find the point that can defend your side

4. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(a) What sign do you expect the **unknown** parameter β_1 to have? What is the interpretation of β_1?

Solution: I would expect β_1 to be positive because, all else equal, unlimited access to ChatGPT should increase productivity and learning outcomes of the student. The interpretation is: all else equal, a paid-subscription to ChatGPT increases GPA on average by β_1.

4. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(b) Name two possible factors in u: one positively correlated to the regressor, and one negatively correlated to the regressor.

This question is asking you to find the u that that could potentially violate our assumption $E[u|x] = 0$ which means $cov(X, u) \neq 0$.

Note: it is asking the relationship between u and x NOT u and y. So do not describe characteristics that raise students' GPA. We need to identify characteristics that make students more likely to use ChatGPT

So can there be any characteristic/factor such that, students who are having that characteristic will be more likely to use ChatGPT (ChatGPT=1) or less likely to use ChatGPT (ChatGPT=0)?

4. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(b) Name two possible factors in u: one positively correlated to the regressor, and one negatively correlated to the regressor.

Solution: A positively correlated factor could be the student motivation to learn. A negatively correlated factor could be how much a student cares about the environment, and thus refrains from using AI.

4. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(c) The mean independence assumption states that $E[u|\text{ChatGPT} = 0] = E[u|\text{ChatGPT} = 1] = 0$. Interpret this. Do you think that this assumption is reasonable?

Given that a student is or is not using ChatGPT, can you infer anything about the characteristics that help them achieve a higher or lower GPA?

The independence assumption says we can't.

Thinking of an unobserved factor that can influence GPA, passion of becoming an astronomer.

$E[u|ChatGPT = 0] = 0$ then will be
$E[becomeAstronomer|ChatGPT = 0] = 0$ means that, among the group of students who are not using ChatGPT, in average, their drive of becoming the astronomers = 0

4. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(c) The mean independence assumption states that $E[u|\text{ChatGPT} = 0] = E[u|\text{ChatGPT} = 1] = 0$. Interpret this. Do you think that this assumption is reasonable?

Given that a student is or is not using ChatGPT, can you infer anything about the characteristics that help them achieve a higher or lower GPA?

The independence assumption says we can't.

Thinking of an unobserved factor that can influence GPA, passion of becoming an astronomer.

$E[u|ChatGPT = 0] = 0$ then will be
$E[becomeAstronomer|ChatGPT = 0] = 0$ means that, among the group of students who are not using ChatGPT, in average, their drive of becoming the astronomers = 0

4. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(c) The mean independence assumption states that $E[u|\text{ChatGPT} = 0] = E[u|\text{ChatGPT} = 1] = 0$. Interpret this. Do you think that this assumption is reasonable?

Solution: This assumption is unlikely to hold, since we expect that u is correlated with ChatGPT (see the previous answer).

4. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(d) Suppose you collect observational data on college students' GPA and ChatGPT. The regression yields a negative $\hat{\beta}_1$. Would you trust the OLS estimator?

We will only trust OLS estimator if $E[u|x]$ is true!

4. The model relating GPA and whether a student has a paid-subscription to ChatGPT is

$$\text{GPA} = \beta_0 + \beta_1 \text{ChatGPT} + u$$

where the regressor "ChatGPT" is equal to 1 is the students has a paid-subscription, and 0 otherwise.

(d) Suppose you collect observational data on college students' GPA and ChatGPT. The regression yields a negative $\hat{\beta}_1$. Would you trust the OLS estimator?

Solution: It is likely that the OLS estimator is biased so I would not trust the estimate.