The Chinese University of Hong Kong, Shenzhen
School of Management and Economics

Midterm Exam ECO3121 Fall 2021

Name:_____

Student ID:_____

Signature:_____

Read the Instructions Carefully

- November 1st, 2021. 7.30-9 pm.
- Write your answers in the spaces below each question. The spaces are ENOUGH, and the answers should be SHORT. Writing overly long answers with unrelated information does NOT help.
- Page 1-6 are exam questions. Page 7-8 are the "official cheatsheet" which may (or may not) contain some formulas or equations that you find useful. Everything need to be turned in at the end of the exam. DO NOT write your answers on page 7-8.
- Total points: 100 points.
- Note that you do not need to know all institutional details in order to correctly answer the questions. You can ask the TAs for translation if you feel certain words prevent you from understanding the questions. But you CANNOT ask the TAs for translation of the terminologies that are covered in class.

To examine the effects of incumbency (running for re-election to a house seat you currently hold) in the 2018 US House of Representatives election, we run the following regression on our dataset of election winners:

$$log(vshare_i) = \beta_0 + \beta_1 Inc_i + \beta_2 D_i + \beta_3 log(share16_i) + U_i, \tag{1}$$

where the dependent variable, $log(vshare)$, is the log of the vote-share the winner of a given seat received in the 2018 election; Inc is a dummy variable with $Inc = 1$ if the winner was an incumbent and $Inc = 0$ if the winner was not the incumbent; D is a dummy variable indicating if the winner was a democrat or republican with $D = 1$ for Democrat and $D = 0$ for Republican (assume everyone in the sample is either a Democrat or Republican); $log(share16)$ is the share of the vote the winner's party received in the district in 2016. Assume $\mathbb{E}[U_i|Inc_i, D_i, log(share16_i)] = 0$.

Part I

1. (6') How should we interpret β_3?

Solution:

- β_3 is the average (or expected) percentage change of current (2018) vote-share from one percent increase of lagged (2016) vote share, given the same incumbency status and political party (or say given all else equal.)

Grading Guidelines

+ 3 for use "percentage" interpretation correctly

+ 3 clearly state the other variables in the multiple regression are held constant.

2. (10') Suppose one wants to test if incumbency has any effect on the candidate's vote-share, given all else equal. Write down (i) the null and alternative hypotheses, (ii) the name of the test statistic, (iii) its distribution for large sample, (iv) how to get the critical value if the significance level is 5%. (Providing the exact critical value is not necessary.)

Solution:

i. Null (H_0) and Alternative (H_1) Hypotheses are:

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

ii. Use the t-statistic

iii. Under H_0 and large n, the t-statistic has a standard normal distribution.

iv. For a two-sided 5% significance test choose critical value, c, such that $P(|t| > c) = .05$, or say the critical value is the 97.5 % quantile of the standard normal distribution.

Alternatively we would have accepted an F-statistic.

Grading Guidelines

+ 2 for each of (i), (ii), (iii), and 4 for (iv). Fine if they say F-test for (ii) but then (iii) and (iv) must follow for F-statistic.

3. (8') Suppose one wants to test that the incumbency status and being a Democrat have no effect on vote-share (after accounting for 2016 vote-share). Write down (i) the null and alternative hypotheses, (ii) the name of the test statistic, (iii) how to get the critical value if the significance level is 5%. (Providing the exact critical value is not necessary.)

Solution:

i. Null (H_0) and Alternative (H_1) Hypotheses are:

$$H_0 : \beta_1 = \beta_2 = 0 \quad H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

ii. Use the F-statistic

iii. When conducting an F-test, we reject H_0 when F is large so we pick our critical value, c, such that $P(F > c) = .05$, or say the critical value is the 95% quantile of this distribution $\frac{\chi^2_2}{2}$. We also give credits if the distribution is written as $F_{2,\infty}$ as in the book.

Grading Guidelines

+ 2 for each of (i), (ii), and 4 for (iv).

4. (8') Suppose that being a Democrat was an advantage in the 2018 elections and incumbents were more likely to be Republicans.

   Suppose we estimate the model without the D variable. Will the OLS estimate of β_1, the coefficient of Inc, suffer from positive or negative omitted variable bias?

   Your argument should be based on the omitted variable bias formula. No hypothesis testing is involved. You can ignore the existence of the regressor $log(share16)$ when analyzing the omitted variable bias.

Solution:

If we omit D, the regression we run is:

$$log(vshare_i) = \beta_0 + \beta_1 Inc_i + \beta_3 log(share16_i) + \epsilon_i,$$

where the error $\epsilon_i = \beta_2 D_i + U_i$. Since the question states being a Democrat was an advantage in 2018, we know $\beta_2 > 0$. The question also stated that incumbents were more likely to be republicans, so D_i (and hence ϵ_i) is negatively correlated with Incumbency. Since $E[U_i|Inc_i, D_i, log(share16_i)] = 0$, $corr(Inc_i, U_i) = 0$ so $\rho = corr(Inc_i, \epsilon_i) = corr(Inc_i, \beta_2 D_i) < 0$. Plugging into the bias function we get

$$\hat{\beta}_1 \to \beta_1 + \beta_2 \delta < \beta_1$$

Hence omitting D will lead to negative omitted variable bias.

Grading Guidelines

+ 4 if correlations are right or if bias equation was used correctly conditional on the wrong correlations.

+ 8 (full credit) for correct answer

5. (6') Think of another variable that is not included in the regression equation (1) and argue that omitting that variable causes a bias in the estimation of β_2.

Solution:

For an omitted variable to cause a bias in our estimate of β_2, it must be that the omitted variable is correlated with D and the omitted variable must have a non-zero effect on the outcome variable, logged vote-share, i.e., it is relevant in the regression equation.

The simplest such example is whether a candidate supports a policy that is typically associated with one party more than the other, where this policy affects vote-share. For example let W_i be a dummy that indicates whether a candidate supports the border wall policy. W_i is negatively correlated with D_i since republicans are more likely to support the policy. It is also reasonable to assume support for the border wall policy will affect vote-share, so W_i has a non-zero affect on $log(vshare_i)$. Hence omitting W_i should cause an bias in estimating β_2.

Grading Guidelines

+ 3 if justifies how the omitted variable is correlated with D.
+ 3 if justifies how the omitted variable affects the outcome.

  The sign of the bias is not required to get full points.

Part II

6. (10') Given the OLS estimator $\hat{\beta}_0$,$\hat{\beta}_1$,$\hat{\beta}_2$,$\hat{\beta}_3$, write down the sum of square residuals. Use the first order condition of the OLS minimization problem to show that the residuals sum to 0.

Solution:

The predicted residuals $\hat{U}_i$ are

$$\hat{U}_i = log(vshare_i) - \hat{\beta}_0 - \hat{\beta}_1 Inc_i - \hat{\beta}_2 D_i - \hat{\beta}_3 log(share16_i).$$

The sum of squared residuals (SSR) are $\sum_{i=1}^{n} \hat{U}_i^2$. You can also write the SSR with the estimators plug in directly.

We know that $\hat{\beta}_0, ..., \hat{\beta}_3$ were selected to minimize $\sum_{i=1}^{n} \hat{U}_i^2$. Writing out that min-

imization problem we get

$$\hat{\beta}_0, ..., \hat{\beta}_3 = \text{argmin}_{b_0,...,b_3} \sum_{i=1}^{n} (log(vshare_i) - b_0 - b_1 Inc_i - b_2 D_i - b_3 log(share16_i))^2$$

Take the derivative with respect to b_0 and evaluate it at the OLS estimator. Since the OLS estimator must satisfy the first order condition for this minimization problem, we get

$$-2\sum_{i=1}^{n} \Big(log(vshare_i) - \hat{\beta}_0 - \hat{\beta}_1 Inc_i - \hat{\beta}_2 D_i - \hat{\beta}_3 log(share16_i)\Big) = -2\sum_{i=1}^{n} \hat{U}_i = 0$$

This shows that the sum of squared residuals $\sum_{i=1}^{n} \hat{U}_i = 0$.

Grading Guidelines

+ 3 write the SSR correctly

+ 7 if knew to take FOC of the intercept, not other variables.

- Simply writing down FOC for all variables only get 5 instead of 7. It is important to make it clear which FOC gives the result. To get full point, $\hat{U}_i = log(vshare_i) - ...$ should be clearly stated.

- 1 for small math errors

+ 10 (full points) if everything is right

7. (6') Compare two regression models with and without D as a regressor. You found that the one with D as a regressor has a higher sample R^2.

   Does this mean that the true value of β_2 must be different from 0?

   Does the model with D has a smaller sum of squared residual from the OLS estimation?

   Explain to get credits.

Solution:

- We know that R^2 increases whenever you add a regressor to the model, regardless of whether or not the coefficient of that regressor equals zero. Since

R^2 goes up from adding D both when $\beta_2 = 0$ and when $\beta_2 \neq 0$, we cannot say if $\boldsymbol{\beta_2}$ is non-zero.

- The model with D has a smaller SSR. There are two ways to explain this: 1. A higher R^2 means that a higher proportion of total sum of squares (TSS), i.e., which is the sum of the variance of the dependent variable, is explained by the regressors. The SSR is the unexplained part. 2. Alternatively, you can say that the model without D can be viewed as a "restricted regression" where the coefficient of D is set to 0. The OLS estimator minimizes the SSR unrestrictedly and reduce SSR when compared to setting the coefficient to 0.

Grading Guidelines

+ 3 for right explanation of R^2
+ 3 for smaller SSR.

8. (6') Suppose the researchers found out that in the sample all incumbents are Democrats. Do you expect a multicollinearity issue? What if you also find Democrats are all incumbents?

Solution:

This is not an issue unless all incumbents are Democrats AND all Democrats are incumbents, in which case you run into an issue with multi-collinearity because that would imply $D = Inc$.

Grading Guidelines

+ 3 for No Issue.
+ 3 explain the multi-collinearity when $D = Inc$.
+ Give 2 points for the first question if states there is an issue with reasonable explanations.

9. (6') Explain the "partialing-out procedure" to estimate $\hat{\beta}_1$. (No need to lay out all the maths)

Solution:

Step 1: regress Inc on D and $log(share16)$, and get the residual $\hat{r}$. This yields the part of Inc that are unexplained by D and $log(share16)$.
Step 2: regress $log(vshare)$ on $\hat{r}$ to get $\hat{\beta}_1$. This yields the partial effect of Inc on $log(vshare)$.

Grading Guidelines

+ 3 points for each step. Deduct 1 point for each step if no explation.

10. (6') Suppose you obtain data for another 100 individuals. Suppose all observations are independent and from the same distribution. How do you expect the standard error for $\hat{\beta}_1$ to change now you have 535 observations? Explain to get credits.

Solution:

Given an increased sample size, the standard error for $\hat{\beta}_1$ should decrease as now the estimator $\hat{\beta}_1$ can be pinned down more precisely. This can be directly seen from the formula for the variance.

Grading Guidelines

+ 3 for decrease

+ 3 for reasonable explanation

11. (6') A friend runs an alternative regression

$$log(vshare_i) = \alpha_0 + \alpha_1 Inc_i + \alpha_2 Rep_i + \alpha_3 log(share16_i) + U_i.$$

with a binary variable Rep_i (1 if the candidate is Republican, 0 otherwise). How will the R^2 change? Explain to get credits.

Solution:

R^2 is a unit-free measure as it captures the fraction of variance explained by the regression and doesn't change with units.

Grading Guidelines

+ 4 for no change

+ 2 for reasonable explanation

Part III

Now consider a nonlinear specification of the model

$$log(vshare_i) = \beta_0 + \beta_1 Inc_i + \beta_2 D_i + \beta_3 log(share16_i) + \beta_4 (Inc_i \times D_i) + U_i. \quad (2)$$

12. (5') What is the main advantage of this new regression compared with the one in part I? Explain how this model improves upon the model in equation (1).

Solution:

This new model allows expected change in the percentage of vote-share from incumbency (or incumbency advantage) to be different between Democrats and Republicans, holding the vote share in 2016 constant. Alternatively, the answer could be the other way: the expect difference in the the percentage of vote-share between Republicans and Democrats depends on the incumbency status.

Grading Guidelines

+ As long as the main idea is right, give full points even if they forget the expected/average change, percentage, or holding share16 constant. The key testing point here, compared to question 1, is the different change for two groups.

13. (6') Write out the incumbency advantage for a Republican and for a Democrat as functions of $\{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\}$ under the model given by (2).

Solution:

The incumbency advantage is the expected percentage change in vote share (change in logged vote share) between a candidate who is an incumbent and a candidate who is not.

- For Republicans that difference is $\boldsymbol{\beta_1}$.
- For Democrats that difference is $\boldsymbol{\beta_1 + \beta_4}$.

Grading Guidelines

+ 3 for republicans.

+ 3 for democrats

14. (5') In this new model, what is the predicted difference in logged vote-share for the 2018 election between a democratic incumbent with logged 2016 vote-share = -.65 (this corresponds to winning the 2016 election with 52% of the vote) and a democratic incumbent with logged 2016 vote-share = -.24 (this corresponds to winning the 2016 election with 79% of the vote)? Assume the OLS estimator $\hat{\beta}_0 = 1, \hat{\beta}_1 = 1, \hat{\beta}_2 = 1, \hat{\beta}_3 = 1$.

Solution:

The only difference between the two candidates is the difference in logged voteshare for their district in 2016, so the predicted difference in logged voteshare between the candidate one and candidate two is

$$\hat{\beta}_3 log(share16_1) - \hat{\beta}_3 log(share16_2) = \hat{\beta}_3(-.65) - \hat{\beta}_3(-.24) = -0.41\hat{\beta}_3 = -0.41.$$

Grading Guidelines

+ 3 if they use the vote share numbers instead of the logged vote share numbers.

15. (6') Suppose you want to know whether the effect of 2016 logged vote share depends on incumbency. What additional regressor should be added to equation (2)? Write down the new model. What are the null and alternative hypotheses in the new model you propose?

Solution:

We should add in an interaction term between $log(share16)$ and Inc. The new model becomes:

$$log(vshare_i) = \beta_0 + \beta_1 Inc_i + \beta_2 D_i + \beta_3 log(share16_i) + \beta_4 (Inc_i \times D_i) + \beta_5 log(share16_i) \times Inc_i + U_i$$

The Null (H_0) and Alternative (H_1) Hypotheses are:

$$H_0 : \beta_5 = 0, \quad H_1 : \beta_5 \neq 0$$

Grading Guidelines

+ 4 for adding in the interaction term

+ 2 for the right hypothesis.