Introductory Econometrics I – Spring 2022
Midterm Solution

Notes:

- Please write your name and student ID clearly on the first page of the answer book.
- Use the last page of this exam question book as the scratch paper.
- Please do not open the exam question book until the proctors ask you to do so.
- No credit will be given unless you show your work.
- Feel free to use either English or Chinese to answer the questions.
- Return the exam question book, your answer book, and the cheat sheet at the end of the exam.

1. (**Vaccine and Death Rate**) Researchers are interested in examining whether COVID-19 vaccines are effective in reducing death rates among those who got the disease. To do so, researchers collected information from N cities. For each city, researchers observe the death rates because of COVID-19, and the vaccination rates, and estimated the following regression model:

$$death_i = \beta_0 + \beta_1 vaccine_i + u_i.$$

$death_i$ is the death rate among those who got COVID-19 (ranges between 0 and 1), and $vaccine_i$ is the fraction of individuals in a city who gets at least one COVID-19 vaccine shot (ranges between 0 and 1).

(a) Think the model from a descriptive perspective. How should we interpret β_1? (5 points)

- **Solution**: β_1 represents the expected difference in death rates between cities where everyone gets at least one vaccine shot, and the cities where nobody gets any vaccine shot.

(b) Suppose we want to interpret the model causally. Explain the meaning of β_1. (5 points)

- **Solution**: Increasing the fraction of people getting at least one vaccine shot from 0 to 1 causes the death rate to change by β_1.

(c) What assumption do we need to justify the causal interpretation? Can you think of one potential omitted variable that might bias the estimate? Explain why. (10 points)

- **Solution**: $E(u|x) = 0$. In other words, factors other than *vaccine* that could affect the death rate are the same (on average) for cities with different vaccine rates. The zero conditional mean assumption may not be true. Potential omitted variables could include: the age distribution of a city, the medical resources etc.

(d) Take the one omitted variable you come up with in c). Do you think the likely bias is positive or negative due to this omitted variable? Please briefly explain. (5 points)

- **Solution**: We need to consider how the omitted variable might affect the death rate, and its correlation with *vaccine*. For example, cities with better medical resources should have lower death rate, and they may also have a higher vaccine rate. Both facts will result in a negative bias.

(e) Suppose researchers get more detailed measure on *vaccine*. Now they know the fraction of people with no vaccine shot (*novac*), one vaccine (*onevac*), two vaccine shots (*twovac*), and three or more vaccine shots (*threevac*). Is it okay to add all four variables into a single regression model? Why? (5 points)

- **Solution**: No. $novac + onevac + twovac + threevac = 1$, so including all of them into a regression equation with an intercept would violate MLR3, no perfect colinearity assumption.

2. (**Sample Analogue**) Consider a regression model:

$$y_i = \beta_0^{\beta_1} + \beta_2 x_{i1} + (\beta_2 - \beta_0) x_{i2} + x_{i1} u_i.$$

Assume that $E(u_i | x_{i1}, x_{i2}) = 0$.

Explain how you want to estimate β_0, β_1 and β_2 using the sample analogue idea. [Hint: come up with some population equations, then translate them into sample analogues. You only need to explain your method. There is no need to derive the explicit expressions for your estimators.](20 points)

- **Solution**: There are many different potential answers. We have three parameters, so we need three equations to solve the model. Two common correct answers are:

  Answer 1: the idea is to directly replicate from what we have for a typical regression model. $E(u_i|x_{i1}, x_{i2}) = 0$ implies that $E(u_i) = E[E(u_i|x_{i1}, x_{i2})] = 0$ (we use the law of iterated expectations). Similarly, we could show that $E(u_i x_{i1}) = 0$, and $E(u_i x_{i2}) = 0$.

Population expectations	Sample analogue
$E(u_i) = 0$	$\frac{1}{N}\sum_{i=1}^{N} \hat{u}_i = 0$
$E(u_i x_{i1}) = 0$	$\frac{1}{N}\sum_{i=1}^{N} \hat{u}_i x_{i1} = 0$
$E(u_i x_{i2}) = 0$	$\frac{1}{N}\sum_{i=1}^{N} \hat{u}_i x_{i2} = 0$

  where $\hat{u_i} = \frac{1}{x_{i1}}(y_i - \hat{\beta}_0^{\hat{\beta}_1} - \hat{\beta}_2 x_{i1} + (\hat{\beta}_2 - \hat{\beta}_0) x_{i2})$.

  Answer 2: view $x_{i1}u_i$ as a new error term, and view x_{i1} and x_{i2} as the two independent variables. Using law of iterated expectations, we can show that $E(u_i x_{i1}) = 0$, $E(u_i x_{i1}^2) = 0$, and $E(u_i x_{i1} x_{i2}) = 0$. The advantage of this method is that we do not need to worry if $x_{i1} = 0$.

Population expectations	Sample analogue
$E(u_i x_{i1}) = 0$	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}_0^{\hat{\beta}_1} - \hat{\beta}_2 x_{i1} + (\hat{\beta}_2 - \hat{\beta}_0) x_{i2}) = 0$
$E(u_i x_{i1}^2) = 0$	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}_0^{\hat{\beta}_1} - \hat{\beta}_2 x_{i1} + (\hat{\beta}_2 - \hat{\beta}_0) x_{i2}) x_{i1} = 0$
$E(u_i x_{i1} x_{i2}) = 0$	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}_0^{\hat{\beta}_1} - \hat{\beta}_2 x_{i1} + (\hat{\beta}_2 - \hat{\beta}_0) x_{i2}) x_{i2} = 0$

3. (**College GPA**) We are interested in exploring factors affecting college GPA and estimate the following regression model:

$$colGPA_i = \beta_0 + \beta_1 hsGPA_i + \beta_2 ACT_i + \beta_3 bgfriend_i + \beta_4 voluntr_i + \beta_5 skipped_i + u_i,$$

where $colGPA$ is college GPA, $hsGPA$ is high school GPA, ACT is college entrance exam score, $bgfriend$ is one if the individual has a boyfriend or girlfrend (and 0 otherwise), $voluntr$ is one if the individual participates in voluntary works, and $skpped$ denotes the total number of lectures missed during the semester.

Figure 1 column (1) shows the regression results, where standard errors are shown in parenthesis and put next to the coefficients. `N` is the number of observations, and `r2` is the R^2. We also estimate

	(1) colGPA		(2) colGPA		(3) colGPA		(4) colGPA	
hsGPA	0.421	(0.0934)	0.412	(0.0937)	0.413	(0.0937)	0.419	(0.0935)
ACT	0.0145	(0.0105)	0.0147	(0.0106)	0.0151	(0.0106)	0.0140	(0.0105)
bgfriend	0.0828	(0.0556)					0.0780	(0.0555)
voluntr	-0.0748	(0.0678)			-0.0669	(0.0679)		
skipped	-0.0870	(0.0262)	-0.0831	(0.0260)	-0.0871	(0.0263)	-0.0826	(0.0259)
_cons	1.347	(0.332)	1.390	(0.332)	1.394	(0.332)	1.345	(0.332)
N	141		141		141		141	
r2	0.251		0.234		0.239		0.245	

Standard errors in parentheses

some other models using the same sample. The results are shown in Figure 1 columns (2) to (4).

Throughout this part, you can give answers that may involve sums, products, quotients or square root of known values; and you do not have to actually calculate a value. For example, feel free to write $(2+3)/4$ instead of 1.25.

(a) Think the model from a causal perspective. How to understand that $\hat{\beta}_5 = -0.0870$? (5 points)

- **Solution**: Holding fixed $colGPA$, $hsGPA$, ACT, $bgfriend$, and $voluntr$, skipping one more lecture causes the GPA to reduce by 0.0870.

(b) Construct a 90% confidence interval for β_1. [Hint: the critical value is 1.656].(5 points)

- **Solution**: $[0.421 - 1.656 \times 0.0934, 0.421 + 1.656 \times 0.0934]$

(c) We are interested in testing whether after holding fixed $hsGPA$, ACT and $skipped$, the variable $bgfriend$ and $voluntr$ have no effect on college GPA, against the alternative that this is not true. Write out the null and alternative hypothesis. Please explain how to decide whether to reject H_0 at the 5% level. [Hint: use the regression results in other columns of Figure 1. Derive the necessary statistics (you do not need to calculate the specific value), and then state how to use that statistic to decide whether to reject H_0.](10 points)

- **Solution**: $H_0 : \beta_3 = 0$ and $\beta_4 = 0$, $H_1 : H_0$ is not true. We use the F test to test the hypothesis. The restricted model is in column (2) and the R^2 information is given. So we use the R^2 version of the F statistic:

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(N - k - 1)} = \frac{(0.251 - 0.234)/2}{(1 - 0.251)/(141 - 5 - 1)}.$$

We can compare F with the critical value c, and reject H_0 if $F > c$. c is the 95-th percentile of an F distribution with (2, 141-5-1) degrees of freedom. Equivalently, we could compare the p-value with 5% and reject H_0 if the p-value is smaller than 5%. The p-value is $P(\mathcal{F} > F)$, where F is calculated above, and $\mathcal{F}$ is an F random variable with (2, 141-5-1) degrees of freedom.

(d) We are interested in testing whether $\beta_2 = 0.2$, against the alternative hypothesis that $\beta_2 \neq 0.2$. Derive the t-statistic for this test. State how to use the statistic to decide whether to reject H_0 at the 1% level. (5 points)

- **Solution**:

$$t = \frac{0.0145 - 0.2}{0.0105}.$$

We can compare $|t|$ with the critical value c, and reject H_0 if $|t| > c$. c is the 99.5th percentile of a t random variable with (141-5-1) degrees of freedom (or, simply use the 99.5th percentile of the standard normal distribution). Equivalently, we could compare the p-value with 1% and reject H_0 if the p-value is smaller than 1%. The p-value is $P(|\mathcal{T}| > |t|)$, where t is calculated above, and $\mathcal{T}$ is a T random variable with (141-5-1) degrees of freedom.

(e) Suppose we want to test $H_0 : \beta_1 = \beta_2$, against the alternative that H_0 is not true. Briefly explain how you can test this using a t-test. [Hint: transform the regression model into a new one involving $\theta = \beta_1 - \beta_2$.] (5 points)

- **Solution**:

$$\begin{aligned} colGPA &= \beta_0 + \beta_1 hsGPA + \beta_2 ACT + \beta_3 bgfriend + \beta_4 voluntr + \beta_5 skipped + u \\ &= \beta_0 + (\theta_1 + \beta_2) hsGPA + \beta_2 ACT + \beta_3 bgfriend + \beta_4 voluntr + \beta_5 skipped + u \\ &= \beta_0 + \theta_1 hsGPA + \beta_2 (hsGPA + ACT) + \beta_3 bgfriend + \beta_4 voluntr + \beta_5 skipped + u. \end{aligned}$$

So we can regress $colGPA$ on 1, $hsGPA$, $hsGPA + ACT$, and other variables. We then test whether the slope coefficient of $hsGPA$ in this regression is 0, by calculating the t-statistic and compare with the critical value.

4. (**Properties of OLS**) Suppose we are interested in estimating the model:

$$y = \beta_0 + \beta_1 x_1 + u.$$

Suppose $E(u|x_1) = 0$ and $Var(u|x) = \sigma^2$. We collected a random sample following the above model with size N.

(a) Suppose we use only the first half observations to estimate the model using OLS. Assume that $\{x_i : i = 1, ..., N/2\}$ are not all the same. The estimators are noted as $\tilde{\beta}_0, \tilde{\beta}_1$. Are these estimators unbiased? Are they consistent? Briefly explain your reason. (10 points)

- **Solution**: They are consistent and unbiased, because the estimators satisfy SLR1-SLR4.

(b) Now suppose we use the full sample to estimate the model using OLS. Assume that $\{x_i : i = 1, ..., N\}$ are not all the same. The estimators are noted as $\hat{\beta}_0, \hat{\beta}_1$. Compared to $\tilde{\beta}_0, \tilde{\beta}_1$, which of these estimators do you prefer? Explain why. [Hint: consider the variance of these estimators.] (10 points)

- **Solution**: We prefer $\hat{\beta}_0, \hat{\beta}_1$. Note that $\hat{\beta}_0, \hat{\beta}_1$ are also unbiased and consistent, because they also satisfy SLR.1-SLR.4. The difference between the two estimators are the variance. Consider the variance of the slope coefficient. Note that the only difference in the variances of these estimators are SST_x: because $\hat{\beta}_1$ uses the full sample, its $SST = \sum_{i=1}^{N}(x_i - \bar{x})^2$ will be larger than if we use half of the sample. (Note that because $\{x_i : i = 1, ..., N/2\}$ are not all the same, $\sum_{i=1}^{N/2}(x_i - \bar{x})^2$ will be positive.) So the variance of $\hat{\beta}_1$ is smaller, and for unbiased estimators, we prefer the one with a smaller variance.