Introductory Econometrics I – Spring 2023
Midterm Suggested Solution

Notes:

- Please write your name and student ID clearly on the first page of the answer book and exam 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 your answer book, cheat sheet, and exam question book at the end of the exam.

1. (**Hypothesis Testing**) We are interested in exploring factors associated with individuals' sleeping length. We have a data set of the following variables:

- $sleep$ = mins sleep at night, per week
- age = age in years
- $educ$ = years of schooling
- $totwrk$ = mins worked per week
- $exper$ = working experience, calculated as age - $educ$ - 6 in the data
- $marr$ = 1 if married, 0 otherwise
- $male$ = 1 if male, 0 otherwise

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) Consider the regression model:

$$sleep = \beta_0 + \beta_1 totwrk + u. \tag{1}$$

We estimate the model using the data and obtain the regression results in Figure 1. Fill in the three blanks in the regression table. (6 points)

```
. regress sleep totwrk

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(1, 704)       =       (a)
       Model |  14381717.2         1  14381717.2   Prob > F        =    0.0000
    Residual |   124858119       704  177355.282   R-squared       =    0.1033
-------------+----------------------------------   Adj R-squared   =    0.1020
       Total |   139239836       705  197503.313   Root MSE        =    421.14

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1507458   .0167403    -9.00   0.000    -.1836126         (b)
       _cons |   3586.377   38.91243      (c)   0.000     3509.979    3662.775
------------------------------------------------------------------------------
```

Figure 1: Regression Results, (a)

- **Solution**: (a): $= |-9|^2$ or $14381717.2/(124858119/704)$ (b) $= -0.1507458*2-(-0.1836126)$, or $-0.1507458 + 1.96 * 0.0167403$ $(c) = 3586.377/38.91243$

(b) For the following questions, consider the regression model:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 age + \beta_3 educ + \beta_4 marr + \beta_5 male + u. \tag{2}$$

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

	sleep	sleep	sleep	sleep
totwrk	**-0.165**	**-0.167**	**-0.166**	**-0.166**
	(0.018)	**(0.018)**	**(0.018)**	**(0.018)**
age	**1.969**			**1.964**
	(1.443)			**(1.443)**
educ	**-11.597**	**-13.885**	**-13.732**	**-11.756**
	(5.872)	**(5.658)**	**(5.664)**	**(5.866)**
marr	**30.360**		**30.124**	
	(41.880)		**(41.905)**	
male	**83.137**	**90.969**	**86.157**	**87.993**
	(34.982)	**(34.274)**	**(34.934)**	**(34.323)**
_cons	**3615.422**	**3747.517**	**3720.912**	**3642.467**
	(117.938)	**(81.006)**	**(89.086)**	**(111.844)**
N	**706**	**706**	**706**	**706**
r2	**0.122**	**0.119**	**0.120**	**0.122**

Standard errors in parentheses

Figure 2: Regression results, (b)

Think about the model from a causal perspective. How to understand that $\hat{\beta}_1 = -0.165$? (4 points)

- **Solution**: Holding fixed *educ*, *age*, *marr*, and *male*, working one more minute per week causes the sleeping minutes to reduce by 0.165 per week.

(c) Is it okay to add *exper* to the regression model (2)? Explain. (5 points)

- **Solution**: No. In the sample, $exper = age - educ - 6$, which is a linear function of *age* and *educ*. Adding *exper* along with *age* and *educ* would results in perfect colinearity.

(d) We are interested in testing in the model (2), whether after holding fixed *totwrk*, *educ*, and *male*, the variables *age* and *marr* have no effect on sleep length, against the alternative that this is not true. Write out the null and alternative hypotheses. Please explain how to decide whether to reject H_0 at the 5% level. [Hint: use the regression results in other columns of Figure 2. 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.](8 points)

- **Solution**: $H_0 : \beta_2 = 0$ 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:

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(N - k - 1)} = \frac{(0.122 - 0.119)/2}{(1 - 0.122)/(706 - 5 - 1)}.$$

We can compare F with the critical value c, and reject H_0 if $F > c$. c is the 95th percentile of an F distribution with (2, 706-5-1) 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, 706-5-1) degrees of freedom.

(e) Suppose we want to test whether the partial impact of *male* on *sleep* is twice as large as the partial impact of *marr* on *sleep* in model (2). $H_0 : \beta_5 = 2\beta_4$. and $H_1 : \beta_5 \neq 2\beta_4$. Briefly explain how you can test this using a t-test. (7 points)

- **Solution**: Define $\theta = \beta_5 - 2\beta_4$. Then $\beta_5 = \theta + 2\beta_4$. We can rewrite the regression model as:

$$\begin{aligned} sleep &= \beta_0 + \beta_1 totwrk + \beta_2 age + \beta_3 educ + \beta_4 marr + \beta_5 male + u \\ &= \beta_0 + \beta_1 totwrk + \beta_2 age + \beta_3 educ + \beta_4 (marr + 2male) + \theta male + u. \end{aligned}$$

So we can regress *sleep* on 1, *totwrk*, *age*, *eudc*, $marr + 2male$, and *male*. We then test whether the slope coefficient of *male* in this regression is 0, by calculating the t-statistic and compare with the critical value.

2. (**Estimating Methods**) We obtain a random sample of $\{(y_i, x_{i1}, x_{i2}) : i = 1, ..., N\}$. We are interested in forecasting y using the following equation:

$$y = \frac{\beta_0 + \beta_1 x_1}{1 + \beta_2 x_2} + u,$$

where u is the error term. Explain how you want to estimate β_0, β_1, and β_2 using the least squares idea. [Hint: You only need to explain your method without worrying about deriving the explicit expressions for your estimators.](20 points)

- **Solution**: Define $\hat{u}_i$ to be

$$\hat{u}_i = y_i - \frac{\hat{\beta}_0 + \hat{\beta}_1 x_{i1}}{1 + \hat{\beta}_2 x_{i2}}.$$

We choose $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ to minimize

$$R = \sum_{i=1}^{N} \hat{u}_i^2.$$

The minimum value of R occurs when the first-order derivative with regard to $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are zero. Often these equations do not have a closed-form solution, and we could use numeric approximation to solve them.

3. (**Properties of the OLS Estimators**) Consider the population regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 z + u,$$

with assumptions that $E(u|x) = 0$, $E(z|x) = E(z) = 0$. Suppose we have a random sample $\{(y_i, x_i, z_i) : i = 1, ..., N\}$, where x_i and z_i are not all the same.

(a) Regress y on x only (with intercept). Let $\tilde{\beta}_1$ denote the OLS estimator for β_1. Is it consistent? Is it unbiased? Explain. (10 points)

- **Solution**: It is consistent and unbiased. Define $v = \beta_2 z + u$. Then $E(v|x) = E(\beta_2 z + u|x) = \beta_2 E(z|x) + E(u|x) = 0$. Given that the model satisfies SLR.1-4, we can apply the result for simple regression to get the unbiasedness and consistency of $\tilde{\beta}_1$.

(b) Regress z on x (with intercept). Let $\tilde{\delta}$ denote the slope coefficient. In large samples (when the sample size N gets large), what value do you think $\tilde{\delta}$ gets close to? You only need to explain your intuition. [Hint: show $cov(x, z) = 0$.] (5 points)

- **Solution**: Note that $E(z|x) = E(z) = 0$ implies that $E(xz) = E(E(xz|x)) = E(xE(z|x)) = 0$. Given this, we can show that $cov(x, z) = E(xz) - E(x)E(z) = 0$. As a result, x and z are uncorrelated. In the regression model $z = \delta_0 + \delta_1 x + v$, the (population) parameter $\delta_1 = \frac{cov(x,z)}{var(x)} = 0$. In large samples, $\tilde{\delta}_1$ converges to 0.

(c) Regress y on x and z (with intercept). Let $\hat{\beta}_1$ denote the OLS estimator for β_1. Is it consistent? Explain. [Hint: consider the omitted variable bias formula, where $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}$. Besides, use the following results: if $plim(T_n) = \alpha$ and $plim(U_N) = \gamma$, then $plim(T_N + U_N) = \alpha + \gamma, plim(T_N U_N) = \alpha\gamma$.] (10 points)

- **Solution**: From (b), we know that $plim(\tilde{\delta}) = 0$. The consistency of $\tilde{\beta}_1$ means that $plim(\tilde{\beta}_1) = \beta_1$. As a result, $plim(\hat{\beta}_1) = plim(\tilde{\beta}_1 - \hat{\beta}_2 \tilde{\delta}) = plim(\tilde{\beta}_1) - plim(\hat{\beta}_2) plim(\tilde{\delta}) = \beta_1$.

4. (**Financial Incentives and Teacher Performance**) Development economists notice that public school teachers are often absent from school. For example, a survey found that 24 percent of teachers in India were absent during school hours.[1] Researchers propose that low financial compensation might be one reason leading to low teacher attendance, and want to quantify the causal effects: if the government increases rural teachers' salaries, how much more would they work? Let *work* denote rural public school teachers' working days per year, and *salary* denote their annual salaries (measured in 1000 yuan). Researchers estimate the following model:

$$work = \beta_0 + \beta_1 salary + u.$$

and find that $\hat{\beta}_1 = 5, se(\hat{\beta}_1) = 0.5$.

[1] Kremer, Michael, Nazmul Chaudhury, F. Halsey Rogers, Karthik Muralidharan, and Jeffrey Hammer. 2005. "Teacher Absence in India: A Snapshot." *Journal of the European Economic Association* 3 (2–3): 658–67.

(a) Suppose they estimate the model by drawing a random sample of 500 rural public school teachers in China and measure *work* and *salary*. To get accurate measures, *work* is collected via anonymous visits by researchers. Think about the model from a descriptive perspective. How should we interpret the slope coefficient? (5 points)

- **Solution**: On average, rural teachers who earn 1000 yuan more work five days more per year.

(b) In this case, can we conclude that higher salaries cause teachers to work more? Explain. (5 points)

- **Solution**: Probably not. The positive correlation between salary and work could also be explained by 1) teachers' compensation is linked to their performance, so working more leads them to earn a higher salary, or 2) more responsible teachers are assigned to higher-salary positions. Their personal traits cause them to work more and earn a higher salary.

(c) Suppose instead, researchers conduct an experiment: they randomly assign teachers to different salary levels, and compare their working days. In this case, can we interpret the model causally? Explain. (5 points)

- **Solution**: Yes. Salaries are randomly assigned to teachers, so they are not correlated to u, i.e. other factors affecting working days. We have $E(u|salary) = 0$ and we can interpret the model causally: increasing salaries by 1000 yuan leads to teachers increasing their working days by 5 days.

(d) Suppose the data are collected in the experiment, are the effects statistically significant at a 5% significance level? Is it economically significant? [Hint: Use the two-sided t-test. The 97.5th percentile of the standard normal distribution is 1.96.] (5 points)

- **Solution**: It is statistically significant at 5% level, because $\hat{\beta}_1/se(\hat{\beta}_1) = 5/0.5 > 1.96$. The economic importance depends on the magnitude of $\hat{\beta}_1$, and is an open-ended question up to your interpretation.

(e) In the data collected from the experiment, researchers also notice that younger teachers tend to work more days than older workers: $E(work|age)$ decreases with age. To estimate β_1, researchers can either regress *work* on *salary*, or regress *work* on both *age* and *salary*. Are these two estimators consistent? Explain. [Hint: consider your answer to Question 3.] (5 points)

- **Solution**: Both estimators are consistent. The fact that salary is randomly assigned suggests that it is not correlated to age. We can apply the results from Question 3, where *age* is z and *salary* is x. As shown above, both estimators are consistent.

- THE END -