Introductory Econometrics I – Spring 2024
Midterm Exam Suggested Solution

Notes:

- Please write your name and student ID clearly on the first page of the answer book.
- Please do not open the exam question book until the proctors ask you to do so.
- Use the last pages of this exam question book as the scratch paper. You may take them off.
- 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 at the end of the exam. You may keep the exam question book (this one), your cheat sheet, and the scratch papers.

1. (**Regression Model Interpretations**) We collect a dataset of two variables, *index* and *cloud*. The variable *index* indicates the Shanghai Composite Index (上证综合指数), which measures the stock market performance. The variable *cloud* measures the cloud cover (云层覆盖率) in Shanghai and ranges between 0 and 1; 0 means no visible cloud, and 1 means the sky is completely cloudy. In the dataset, each observation is a trading day randomly drawn from 2000 to 2023. The sample size is 50. Figure 1 plots the two variables, with *cloud* on the x-axis and *index* on the y-axis.

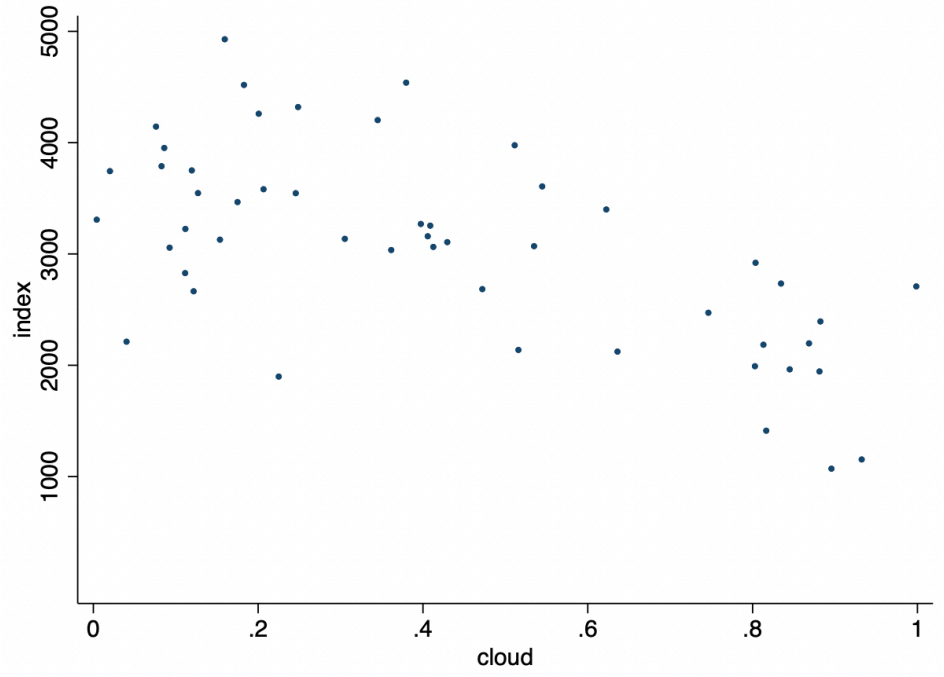

Figure 1: *index* and *cloud*

(a) Construct a simple linear regression model to describe the pattern in the data. Then explain with words the meaning of the model. [Hint: you do not need to estimate the model.] (4 points)

- **Solution**: The model is:

$$index = \beta_0 + \beta_1 cloud + u.$$

or, equivalently,

$$E(index|cloud) = \beta_0 + \beta_1 cloud.$$

The model shows how the expected value of *index* varies with *cloud*. In general, we find that stock prices are higher on less cloudy days; in other words, the two variables are negatively correlated. We give credit to all reasonable answers.

(b) Behavioral economists claim that weather like cloud cover impacts stock market prices because the weather condition changes the investor's mood. Use a simple linear regression model to capture this causal relationship. Explain the meaning of your model elements, for example, the parameters, the variables, etc. (5 points)

- **Solution**: We may construct the following model:

$$index = \beta_0 + \beta_1 cloud + u.$$

In the model, *cloud* and *index* are defined as in the question. u indicates factors affecting *index* other than *cloud*, for example, the monetary policy, the macroeconomic conditions, etc. β_1 measures the causal impact of *cloud* on *index*: if *cloud* increases by one unit, the change in *index* because of that. Note: we give credit to all other reasonable answers. For example, we give full points if you specify the causal relationship between cloud and mood.

(c) Under what conditions can we interpret the model in (b) causally? Explain whether you think it is satisfied in this example. (6 points)

- **Solution**: To interpret the model causally, we require the zero conditional mean assumption, i.e., $E(u|cloud) = 0$. The condition means that other factors affecting the stock price index have similar expected values on cloudy days and sunny days. Whether it is satisfied in this example is an open-ended question. We give credit to all reasonable answers. Again, if you think about other causal relationships in (b), as long as you argue whether the zero conditional mean assumption is satisfied, we give full credit.

2. (**Estimating Methods**) Consider the population regression model:

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

with assumptions that $E(u|x) = 0$. Suppose we have a random sample $\{(y_i, x_i) : i = 1, ..., N\}$.

(a) Show that $E(x^3 u) = 0$ and $E(xu) = 0$. [Hint: use the law of iterated expectations.] (4 points)

- **Solution**: Using the law of iterated expectations and the fact that $E(u|x) = 0$:

$$E(x^3 u) = E(E(x^3 u|x)) = E(x^3 E(u|x)) = 0,$$
$$E(xu) = E(E(xu|x)) = E(xE(u|x)) = 0.$$

(b) Write the sample analogues of $E(x^3 u) = 0$ and $E(xu) = 0$. (5 points)

- **Solution**: Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote our estimators, and define $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.
The sample analogues are:

$$\frac{1}{N}\sum_{i=1}^{N} x_i^3 \hat{u}_i = 0,$$
$$\frac{1}{N}\sum_{i=1}^{N} x_i \hat{u}_i = 0.$$

(c) Propose an estimator of β_1 and β_0 based on the sample analogues you derived in (b). Note: you do not need to explicitly solve the estimators. (5 points)

- **Solution**: Given the two sample analogues, we have two equations:

$$\frac{1}{N}\sum_{i=1}^{N} x_i^3 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$
$$\frac{1}{N}\sum_{i=1}^{N} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Using the two equations we can solve for $\hat{\beta}_0$ and $\hat{\beta}_1$.

(d) Explain what it means that the estimators in (c) are unbiased and what it means that the estimators in (c) are consistent. Note: you do not need to show whether they are unbiased or consistent. (6 points)

- **Solution**: The unbiasedness means that expected values of the estimators are the same as the true values: $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$. Consistency means that with sufficiently large sample sizes, the estimators are arbitrarily close to the true parameter values: $plim(\hat{\beta}_0) = \beta_0$ and $plim(\hat{\beta}_1) = \beta_1$. We give credits if you either state the mathematical equations or argue using words.

3. (**Hypothesis Testing**) Consider a Cobb-Douglas production function:

$$Q = AK^{\alpha}L^{\beta}e^{u}, \tag{1}$$

where Q is the total output, K is capital, L is labor, A is technology (which is unobserved and a constant with $A > 0$), e is the base of the natural logarithm, and u is an error term. To estimate the model, we collect a random sample of size N. In the sample, we observe Q, K, and L.

(a) Transform equation (1) into a linear regression model. [Hint: $log(x^y) = ylog(x), log(e^x) = x$, where log is the natural logarithm.] (4 points)

- **Solution**: The transformed equation is:

$$log(Q) = log(A) + \alpha log(K) + \beta log(L) + u.$$

(b) Interpret the meaning of α causally. (5 points)

- **Solution**: α means, holding fixed labor inputs, if the capital inputs increase by 1%, then the total output increases by α%.

(c) We are interested in testing whether there are constant returns to scale in the production function. Constant returns to scale mean that a proportional increase in the inputs produces the same proportional increase in output. In other words, if $K^* = \gamma K$ and $L^* = \gamma L$, then constant returns to scale means $Q(K^*, L^*) = \gamma Q(K, L), \forall \gamma > 0$, where $Q(K, L) = AK^{\alpha}L^{\beta}$. Write out the null hypothesis that there are constant returns to scale as an expression of the coefficients in the model you derived in (a). [Hint: $e^{x+y} = e^x e^y$.] (5 points)

- **Solution**: $Q(K^*, L^*) = A(\gamma K)^{\alpha}(\gamma L)^{\beta} = A\gamma^{(\alpha+\beta)}K^{\alpha}L^{\beta} = \gamma^{(\alpha+\beta)}Q(K, L)$. Constant returns to scale means $Q(K^*, L^*) = \gamma Q(K, L), \forall \gamma > 0$. For this to hold, we require $\alpha + \beta = 1$. The null hypothesis is $\alpha + \beta = 1$.

(d) Explain how we can test the null hypothesis in (c) against the alternative that it is not true. Your answer should specify: 1) what are the test statistics? 2) what are the rejection rules? (6 points)

- **Solution**: We may use an F test. If H_0 is true, the restricted is:

$$\begin{aligned} log(Q) &= log(A) + (1-\beta)log(K) + \beta log(L) + u, \\ log(Q) - log(K) &= log(A) + \beta(log(L) - log(K)) + u, \\ log(Q/K) &= log(A) + \beta log(L/K) + u \end{aligned}$$

We estimate the model and obtain the SSR_r. We then estimate the original model in (a) and obtain SSR_{ur}. We then calculate the F statistic as

$$F = \frac{(SSR_r - SSR_{ur})/1}{SSR_{ur}/(N-3)}.$$

Pick a significance level of α. Find the critical value as the $1-\alpha^{th}$ percentile of the F distribution with numerator degree of freedom of 1 and denominator degree of freedom of $N-3$. We reject the null hypothesis if F is larger than the critical value. Note: you may also use a t-test. We give credit to all reasonable answers.

4. (**Empirical Exercises**) Baby's birth weight is an important indicator of health. We are interested in exploring factors affecting the birth weight. We collect a data set of 1664 babies with the following variables:

   - $bwght$ = birth weight, grams. The World Health Organization defines low birth weight as the birth of < 2500 grams. These kids have a higher probability of having bad health after birth. The average birth weight in the data is 3500.
   - $pareage$ = mother age plus father age, in years
   - $meduc$ = mother's education, in years
   - $feduc$ = father's education, in years
   - $cigs$ = average cigarettes per day smoked by mother
   - $drinks$ = average drinks per week by mother

   Using the data, we estimate the following multiple linear regression model:

$$bwght = \beta_0 + \beta_1 pareage + \beta_2 meduc + \beta_3 feduc + \beta_4 cigs + \beta_5 drinks + u. \tag{2}$$

   The regression results are displayed in Figure 2.

Source	SS	df	MS	Number of obs	=	1,664
				F(5, 1658)	=	3.28
Model	5269653.51	5	1053930.7	Prob > F	=	0.0059
Residual	532129170	(d)	320946.423	R-squared	=	0.0098
				Adj R-squared	=	0.0068
Total	537398823	1,663	323150.225	Root MSE	=	(a)

bwght	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pareage	2.749829	1.537402	1.79	0.074	-.2656251	5.765284
meduc	-2.175729	8.419011	(b)	0.796	-18.68874	14.33728
feduc	9.149812	7.647633	1.20	0.232	(c)	24.14985
cigs	-9.669166	3.509162	-2.76	0.006	-16.55202	-2.78631
drink	-12.47638	48.18475	-0.26	0.796	-106.9858	82.03299
_cons	3153.301	119.1094	26.47	0.000	2919.68	3386.921

Figure 2: Regression Results

   (a) Fill in the blanks in the table. Note: you can give answers that may involve sums, products, quotients, or square roots 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. (12 points)

- **Solution**: Note: there might be multiple ways to solve the numbers. We give credit to all reasonable answers. (For the last answer, we also give credit if you use $N = 1644$.)
  i. $\sqrt{320946.423}$
  ii. $\frac{-2.175729}{8.419011}$
  iii. $9.149812 \times 2 - 24.14985$
  iv. $1664 - 5 - 1$

(b) We estimate that the coefficient of *pareage* is about 2.75. Interpret the meaning of this value causally. (5 points)

- **Solution**: Holding fixed all other variables, if parent age increases by 1 year, then the kid birth weight increases by 2.75 grams.

(c) Is the coefficient of *pareage* statistically significant at the 10% significance level? Do you think it is economically significant? Explain your answer. (5 points)

- **Solution**: Yes, we can compare the p-value with 10%. We find that $0.074 < 10\%$, so *pareage* is significant at the 10% level. To determine the economic significance, we consider whether the magnitude of 2.75 is large in this context. Whether it is economically significant is an open-ended question. We give credit for all reasonable answers.

(d) Suppose in the dataset, we also observe mother's age *mage* and father's age $fage$. Is it okay if we add these two variables to equation (2)? Explain. (5 points)

- **Solution**: No. $pareage = mage + fage$. Including them would cause perfect collinearity.

5. (**Properties of the OLS Estimators**) Suppose that we have estimated the parameters of the multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

using the ordinary least squares (OLS) and a random sample of $\{(y_i, x_{i1}, x_{i2}), i = 1, 2, ..., N\}$. Denote the OLS estimates as $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, the estimated residuals as $\hat{u}_i$ and the fitted values as $\hat{y}_i$. Denote the sample average of y, x_1, and x_2 as $\bar{y}$, $\bar{x}_1$ and $\bar{x}_2$.

(a) Suppose $\hat{\beta}_1 = \hat{\beta}_2 = 0$. What is the R^2 of the model? Prove it. [Hint: You may use the fact that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2$.] (6 points)

- **Solution**: The R^2 is 0. There might be different ways to show this. We give credit to all reasonable answers. Below is a suggested proof.
  When $\hat{\beta}_1 = \hat{\beta}_2 = 0$, $\bar{y} = \hat{\beta}_0$. As a result, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} = \bar{y} + 0 + 0 = \bar{y}, \forall i$.
  Plug in the formula for R^2:

$$R^2 = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = \frac{\sum_{i=1}^N (\bar{y} - \bar{y})^2}{\sum_{i]1}^N (y_i - \bar{y})^2} = 0.$$

(b) Now suppose $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ may take on any values. Using the sample, we regress y_i on $\hat{y}_i$. Show that the R^2 of this regression is the same as the R^2 of the original regression model in which we regress y on x_1 and x_2. [Hint: you may use the fact that $\bar{\hat{y}} \equiv \frac{1}{N} \sum_{i=1}^N \hat{y}_i = \bar{y}$.] (6 points)

- **Solution**: There might be different ways to show this. We give credit to all reasonable answers. Below is a proof.

If we regress y_i on $\hat{y}_i$, the slope coefficient is:

$$\frac{\sum_{i=1}^N(\hat{y}_i-\bar{y})(y_i-\bar{y})}{\sum_{i=1}^N(\hat{y}_i-\bar{y})^2}=\frac{\sum_{i=1}^N(\hat{y}_i-\bar{y})(\hat{y}_i-\bar{y}+\hat{u}_i)}{\sum_{i=1}^N(\hat{y}_i-\bar{y})^2}=\frac{\sum_{i=1}^N(\hat{y}_i-\bar{y})^2+\sum_{i=1}^N(\hat{y}_i-\bar{y})\hat{u}_i}{\sum_{i=1}^N(\hat{y}_i-\bar{y})^2}=1.$$

We use the fact that $y_i=\hat{y}_i+\hat{u}_i$ and $\sum_{i=1}^N\hat{u}_i=0$, $\sum_{i=1}^N\hat{y}_i\hat{u}_i=0$. The intercept is $\bar{y}-1\times\bar{\hat{y}}=0$. Thus the fitted value if we regress y_i on $\hat{y}_i$ will also be $\hat{y}_i$. The R^2 is

$$\frac{SSE}{SST}=\frac{\sum_{i=1}^N(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^N(y_i-\bar{y})^2},$$

which is the same as the R^2 of the original regression model.

(c) Now suppose $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ may take on any values. Using the sample, if we regress $\hat{u}_i$ on x_{i1} and x_{i2}, what is the R^2? Prove it. (6 points)

- **Solution**: The R^2 is 0. There might be different ways to show this. We give credit to all reasonable answers. Below is a proof.
  First, let $\hat{\delta}_1$ denote the slope coefficient of x_1 when we regress $\hat{u}_i$ on x_1 and x_2. Let $\hat{r}_{i1}$ denote the residual obtained from regressing x_1 on x_2, and $\hat{\alpha}_0$ and $\hat{\alpha}_1$ denote the intercept and the slope coefficients. Using the FWL theorem:

$$\hat{\delta}_1=\frac{\sum_{i=1}^N\hat{u}_i\hat{r}_{i1}}{\sum_{i=1}^N\hat{r}_{i1}^2}=\frac{\sum_{i=1}^N\hat{u}_i(x_{i1}-\hat{\alpha}_0-\hat{\alpha}_1x_{i2})}{\sum_{i=1}^N\hat{r}_{i1}^2}=0.$$

  Note that the numerator is zero because $\sum_{i=1}^N\hat{u}_i=0,\sum_{i=1}^N\hat{u}_ix_{i1}=0$, and $\sum_{i=1}^N\hat{u}_ix_{i2}=0$. Using the same logic, we could show that the slope coefficient of x_2 when we regress $\hat{u}_i$ on x_1 and x_2 is also 0. Finally, using the conclusion from part (a), we show that the R^2 is zero.

- THE END -