University of Illinois at Urbana-Champaign

ECON471

Introduction to Applied Econometrics

Exam 1

October 6, 2020

H. Arvin-Rad

Instructions.

This is a closed book exam. Show your work. Unsupported or unreadable answers receive no credit. Answer all questions in parts I and II. In part I, each multiple-choice question has 3 points. In part II, the weight of each part of a question is the number in parenthesis next to the part.

Please hand-write your answers in the space provided for each part. If you don't have access to a printer to get a printout of the exam, make sure you clearly label each question and the parts in that question.

Please join the zoom meeting while taking the test. The password for the meeting is: EXAM1.

Upload your completed exam in a single file by 10:15pm, CST. You can only submit your exam once.

Please take a picture of your student ID and attach it to your exam, and sign the code of honor below. **Exams uploaded without the attached photo ID or no signing of the code of honor will be penalized.** Violations of academic integrity as given in the Code on Campus Affairs will be taken extremely seriously. Students found cheating will be penalized according to the Code's guidelines.

Please sign the honor code:

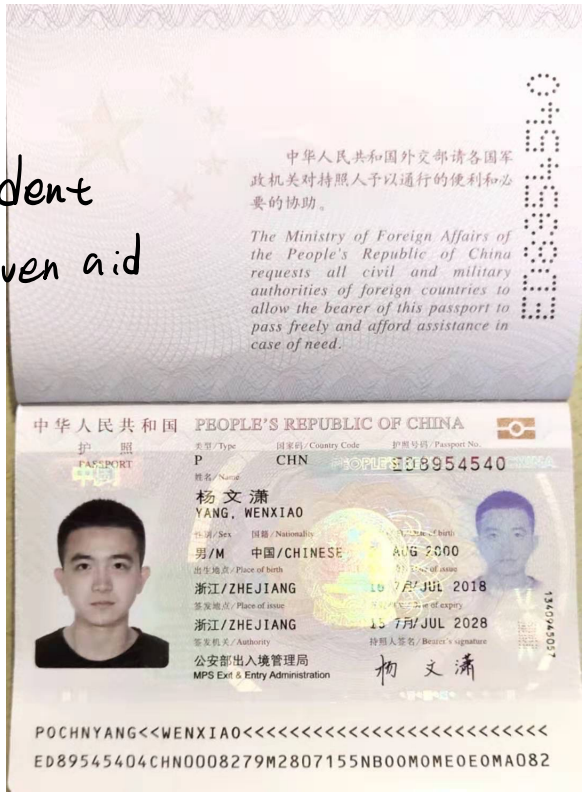
On my honor as a student, I have neither received nor given aid

Signature: On my honor as a student I have neither received nor given aid on this exam.

Name: Wenxiao Yang

NetID: wenxiao 5

Section: B3



I am a new transfer student

I haven't got my ID card

This is the picture of my passport.

Part I (30 points). Multiple Choice Questions. Please enter your answers for this part in the multiple-choice answer sheet (page 4) by circling the letter of your choice.

1. Consider the multiple regression model with two regressors X_1 and X_2, where both variables are determinants of the dependent variable. First regress Y on X_1 only and obtain the estimate of the slope $\tilde{\beta}_1$. Then regress Y on X_1 and X_2 and obtain the slope estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. What is the relationship between $\tilde{\beta}_1$ and $\hat{\beta}_1$?

   a. $\tilde{\beta}_1 = \hat{\beta}_1$
   b. $\tilde{\beta}_1$ will be greater than $\hat{\beta}_1$ if X_1 and X_2 are positively correlated and the partial effect of X_2 on Y is positive
   c. $\tilde{\beta}_1$ will be greater than $\hat{\beta}_1$ if X_1 and X_2 are positively correlated and the partial effect of X_2 on Y is negative
   d. $\tilde{\beta}_1$ will always be greater than $\hat{\beta}_1$ since it shows the effect of both X_1 and X_2 on Y.

2. In the bivariate regression model, the slope estimator, $\hat{\beta}_1$, has a smaller standard error, other things equal, if

   a. there is more variation in the explanatory variable, X.
   b. there is a large variance of the error term, u.
   c. the sample size is smaller.
   d. the intercept, β_0, is small.

3. Consider the following least squares specification between tests cores and income: $\widehat{TestScore} = 557.8 + 36.42\ ln\ (Income)$. According to this equation, a 1% increase income is associated with an increase in test scores of

   a. 36.42 points
   b. 557.8 points
   c. 0.36 points
   d. cannot be determined from the information given here

4. Which of the following is NOT a good reason for including a disturbance term in a regression equation?

   a. It captures omitted determinants of the dependent variable
   b. It allows for the non-zero mean of the dependent variable
   c. It allows for errors in the measurement of the dependent variable
   d. It allows for using proxies for some hard-to-measure independent variables

5. Which of the following statements is correct concerning the conditions required for OLS to be a usable estimation technique?

   a. The model must be linear in the parameters
   b. The model must be linear in the variables
   c. The model must be linear in the variables and the parameters
   d. The model must be linear in the residuals.

$\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$

6. When the estimated slope coefficient, $\hat{\beta}_1$, in the linear regression model $y = \beta_0 + \beta_1 x + u$, is zero, then

$\hat{y}_i = \beta_0 \qquad SSE = 0$

a. $0 < R^2 < 1$.
b. $R^2 > SSE/SST$
c. R^2 may be negative.
d. sample correlation between *x* and *y* is zero.

7. A random sample of *n* observations is used to estimate the linear regression model $y = \beta_0 + \beta_1 x + u$. Let $\hat{u}$ be the OLS residual. Which of the following statements is correct?

a. $\sum_{i=1}^{n} u_i x_i = 0$
b. $\sum_{i=1}^{n} \hat{u}_i y_i = 0$
c. $\sum_{i=1}^{n} \hat{u}_i x_i^2 > 0$
d. $\sum_{i=1}^{n} \hat{u}_i x_i = 0$

8. Which of the following statements is true?

a. A variable has a causal effect on another variable if both variables increase or decrease simultaneously.
b. Difficulty in inferring causality disappears when studying data at fairly high levels of aggregation.
c. The notion of 'ceteris paribus' plays an important role in causal analysis.
d. The problem of inferring causality arises if experimental data is used for analysis.

9. Which of the following is a difference between panel and pooled cross-sectional data?

a. A panel data set consists of data on different cross-sectional units over a given period of time while a pooled cross-sectional data set consists of data on the same cross-sectional units over a given period of time.
b. A panel data set consists of data on the same cross-sectional units over a given period of time while a pooled cross-sectional data set consists of data on different cross-sectional units over a given period of time.
c. A panel data consists of data on a single variable measured at a given point in time while a pooled cross-sectional data set consists of data on the same cross-sectional units over a given period of time.
d. A panel data set consists of data on a single variable measured at a given point in time while a pooled cross-sectional data set consists of data on more than one variable at a given point in time.

10. Which of the following is true of R^2?

a. R^2 is also called the standard error of regression.
b. A low R^2 indicates that the Ordinary Least Squares line fits the data well.
c. R^2 usually decreases with an increase in the number of independent variables in a regression.
d. R^2 shows what percentage of the total variation in the dependent variable, *Y*, is explained by the explanatory variables.

Multiple-Choice Answer Sheet

Please circle the letter of your choice for each multiple-choice question.

1.	a	**b**	c	d	e
2.	**a**	b	c	d	e
3.	a	b	**c**	d	e
4.	a	**b**	c	d	e
5.	a	b	c	**d**	e
6.	a	b	c	**d**	e
7.	a	b	c	**d**	e
8.	a	b	**c**	d	e
9.	a	**b**	c	d	e
10.	a	b	c	**d**	e

Part II (70 points).

1. (8 points) In the linear regression model $y = \beta_0 + \beta_1 x + u$, let y be the selling price of a house in dollars and x be its living area in square feet. Define a new variable $x^* = x - 1000$ (that is, $x^*$is the square feet in excess of 1000), and estimate the model $y = \beta_0^* + \beta_1^* x^* + v$.

a. (4) Show the relationship between the OLS estimators $\widehat{\beta_1^*}$ and $\widehat{\beta_1}$.

$$X_i^* - \bar{X}^* = (X_i - 1000) - (\bar{X} - 1000) = X_i - \bar{X}$$

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n} (X_i^* - \bar{X}^*)(y_i - \bar{y})}{\sum_{i=1}^{n} (X_i^* - \bar{X}^*)^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \hat{\beta}_1$$

b. (4) Show the relationship between the OLS estimators $\widehat{\beta_0^*}$ and $\widehat{\beta_0}$.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_0^* = \bar{y} - \hat{\beta}_1^* \bar{X}^* = \bar{y} - \hat{\beta}_1 (\bar{X} - 1000) = \bar{y} - \hat{\beta}_1 \bar{X} + 1000 \hat{\beta}_1$$

$$= \hat{\beta}_0 + 1000 \hat{\beta}_1$$

Hence $\hat{\beta}_0^* = \hat{\beta}_0 + 1000 \hat{\beta}_1$

$$= \hat{\beta}_0 + \frac{1000 \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

2. (30 points) A sample of 274 male employees was used to investigate the relationship between hourly wage rates y_i (measured in dollars per hour) and firm tenure x_i (measured in years). Preliminary analysis of the sample data produces the following sample information:

$n = 274$ $\quad \sum_{i=1}^{n} y_i = 1945.26$ $\quad \sum_{i=1}^{n} x_i = 1774.00$ $\quad \sum_{i=1}^{n} {y_i}^2 = 18536.73$
$\sum_{i=1}^{n} {x_i}^2 = 30608.00$ $\quad \sum_{i=1}^{n} x_i y_i = 16040.72$ $\quad \sum_{i=1}^{n} {\hat{u}_i}^2 = 4105.297$

Use the above sample information to answer all the following questions. Show explicitly all formulas and calculations.

a. (6) Prove algebraically that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$.

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}(-2\bar{x}\,x_i) + n(\bar{x})^2$$
$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$
$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

b. (5) Prove algebraically that $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$.

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\bar{y}x_i - \sum_{i=1}^{n}\bar{x}y_i + n\bar{x}\bar{y}$$
$$= \sum_{i=1}^{n} x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y}$$
$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

c. (6) Compute OLS estimates of the intercept coefficient β_0 and the slope coefficient β_1.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{16040.72 - 274\times\frac{1774}{274}\times\frac{1945.26}{274}}{30608 - 274\left(\frac{1774}{274}\right)^2}$$

$$= 0.180220089$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = \frac{1945.26}{274} - 0.180220089\times\frac{1774}{274}$$

$$= 5.932662635$$

d. (4) Calculate an estimate of σ^2, the error variance.

$$\hat{\sigma}^2 = \frac{SSR}{n-2} = \frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2} = \frac{4105.297}{272} = 15.09300368$$

e. (3) Calculate an estimate of $Var(\hat{\beta}_1)$.

$$\widehat{Var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{15.09300368}{30608 - 274\left(\frac{1774}{274}\right)^2} = 0.000189287$$

f. (6) Compute the value of R^2, the coefficient of determination. Briefly explain what the calculated value of R^2 means.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i-\bar{Y}_i)^2} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}_i^2} = 1 - \frac{4105.297}{18536.73 - 274\left(\frac{1945.26}{274}\right)^2} = 0.13141$$

the pecentage of the total variation in dependent variable Y is explained by the explanatory variable X

3. (12 points) Suppose you are interested in estimating the effect of hours spent in an SAT preparation course (*hours*) on total SAT score (*sat*). The population is all college-bound high school seniors for a particular year.

a. (4)Suppose you are given a grant to run a controlled experiment. Explain how you would structure the experiment in order to estimate the causal effect of *hours* on *sat*.

divide students into different groups by their grades and we controll the number of hours in each group

ceteris paribus

For example, in group one we let $\frac{1}{5}$ students study { 2 hours per day, 4 hours per day, 6 hours per day, 8 hours per day, 10 hours per day }

b. (4) Consider the more realistic case where students choose how much time to spend in a preparation course, and you can only randomly sample *sat* and *hours* from the population. Write the population model as

$$sat = \beta_0 + \beta_1 hours + u$$

where, as usual in a model with an intercept, we can assume $E(u) = 0$. Consider innate ability, family income, and general health on the day of the exam, as factors that might have an effect on *sat*. Are these factors likely to have positive or negative correlation with *hours*?

innate ability may have negative correlation with hours (smart students study really fast)

family income may have positive correlation with hours. (high income family can support their children take more courses)

General health on the day of the exam may have negative correlation with hours.

(too much study may cause bad health, or bad health can't support too much study).

c. (2) In the equation from part (b), what should be the sign of β_1 if the preparation course is effective?

positive

d. (2) In the equation from part (b), what is the interpretation of β_0?

the average SAT score students can get without preparation.

4. (20 points) The median starting salary for new law school graduates is determined by

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u,$$

where *LSAT* is the median LSAT score for the graduating class, *GPA* is the median college GPA for the class, *libvol* is the number of volumes in the law school library, *cost* is the annual cost of attending law school, and *rank* is a law school ranking (with *rank* = 1 being the best).

a. (3)Explain why we expect $\beta_5 \leq 0$.

the better the law school the smaller the rank

So when the rank increases the law school may be worse

the salary may be lower.

b. (5)What signs do you expect for the other slope parameters? Justify your answers.

β_1 is positive } better grades show the graduates have more knowledge and have more chance to get higher salary.

β_2 is positive

β_3 is positive: better law schools often have more volumn in library graduates have more chance to read more books.

β_4 is positive: good law schools' expenditure is huge. cost may also show the quality of the schol.

c. (4)Using the data in LAWSCH85, the estimated equation is

$$\log(\widehat{salary}) = 8.34 + 0.0047LSAT + 0.248GPA + 0.095\log(libvol)$$

$$+0.038\log(cost) - 0.038rank$$

$$n = 136, \quad R^2 = 0.842.$$

What is the predicted ceteris paribus difference in salary for schools with a median GPA different by one point? (report your answer as a percentage.)

$\%\widehat{\Delta salary} = 100\,\Delta s \approx 100 \times 0.248 \times \Delta GPA = 24.8\%$

d. (4)Interpret the coefficient on the variable $\log(libvol)$.

every 1% change in libvol will cause 0.095% change in salary.

e. (4)Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary?

Yes.

$\%\widehat{\Delta salary} = 100\,\Delta s \approx 100 \times (-0.038) \times 20 = 76\%$

A difference in ranking of 20 will cause 76% difference in predicted starting salary

Formula Sheet

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$SST = \sum (Y_i - \bar{Y})^2 \qquad SSE = \sum (\hat{Y}_i - \bar{Y})^2 \qquad SSR = \sum (Y_i - \hat{Y}_i)^2$$

$$R^2 = \frac{SSE}{SST}$$

$$r_{x,y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}.$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

$$Var(\hat{\beta}_0) = \frac{\sigma^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2}$$

$$\hat{\sigma}^2 = \frac{SSR}{n-k-1}, \text{ where k is the number of regressors.}$$

$$\widehat{Var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}$$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}$$