

Econ 312: Midterm II

Sample Exam

Part I. Multiple Choice (15 questions worth 3 points each).

1. Finding a small value of the p-value (e.g. less than 5%)

- a. indicates evidence in favor of the null hypothesis.
- b. implies that the t-statistic is less than 1.96.
- c. indicates evidence in against the null hypothesis.
- d. will only happen roughly one in twenty samples.

Answer: c

2. A binary variable is often called a

- a. dummy variable
- b. dependent variable
- c. residual
- d. power of a test

Answer: a

3. Under imperfect multicollinearity

- a. the OLS estimator cannot be computed.
- b. two or more of the regressors are highly correlated.
- c. the OLS estimator is biased even in samples of $n > 100$.
- d. the error terms are highly, but not perfectly, correlated.

Answer: b

4. When there are omitted variables in the regression, which are determinants of the dependent variable, then

- a. you cannot measure the effect of the omitted variable, but the estimator of your included variable(s) is (are) unaffected.
- b. this has no effect on the estimator of your included variable because the other variable is not included.
- c. this will always bias the OLS estimator of the included variable.
- d. the OLS estimator is biased if the omitted variable is correlated with the included variable.

Answer: d

5. Imagine you regressed earnings of individuals on a constant, a binary variable ("Male") which takes on the value 1 for males and is 0 otherwise, and another binary variable ("Female") which takes on the value 1 for females and is 0 otherwise. Because females typically earn less than males, you would expect

- a. the coefficient for Male to have a positive sign, and for Female a negative sign.
- b. both coefficients to be the same distance from the constant, one above and the other below.
- c. none of the OLS estimators to exist because there is perfect multicollinearity.
- d. this to yield a difference in means statistic.

Answer: c

6. Omitted variable bias

- a. will always be present as long as the regression
- b. is always there but is negligible in almost all economic examples
- c. exists if the omitted variable is correlated with the included regressor but is not a determinant of the dependent variable
- d. exists if the omitted variable is correlated with the included regressor and is a determinant of the dependent variable

Answer: d

7. When testing joint hypothesis, you should

- a. use t-statistics for each hypothesis and reject the null hypothesis if all of the restrictions fail.
- b. use the F-statistic and reject all the hypothesis if the statistic exceeds the critical value.
- c. use t-statistics for each hypothesis and reject the null hypothesis once the statistic exceeds the critical value for a single hypothesis.
- d. use the F-statistics and reject at least one of the hypothesis if the statistic exceeds the critical value.

Answer: b

8. In the multiple regression model, the t-statistic for testing that the slope is significantly different from zero is calculated

- a. by dividing the estimate by its standard error.
- b. from the square root of the F-statistic.
- c. by multiplying the p-value by 1.96.
- d. using the adjusted R^2 and the confidence interval.

Answer: a

9. To test joint linear hypotheses in the multiple regression model, you need to

- a. compare the sums of squared residuals from the restricted and unrestricted model.
- b. use the F-statistic.
- c. use several t-statistics and perform tests using the standard normal distribution.
- d. compare the adjusted R^2 for the model which imposes the restrictions, and the unrestricted model.

Answer: b

10. At a mathematical level, if the two conditions for omitted variable bias are satisfied, then

- a. $E(\epsilon_i | X_{1i}, X_{2i}, \dots, X_{1ki}) \neq 0$.
- b. there is perfect multicollinearity.
- c. large outliers are likely.
- d. $(X_{1i}, X_{2i}, \dots, X_{1ki}, Y_i)$, $i = 1, \dots, n$ are not i.i.d.

Answer: a

11. All of the following are true, with the exception of one condition:

- a. a high R^2 or \bar{R}^2 does not mean that the regressors are a true cause of the dependent variable.
- b. a high R^2 or \bar{R}^2 does not mean that there is no omitted variable bias.
- c. a high R^2 or \bar{R}^2 always means that an added variable is statistically significant.
- d. a high R^2 or \bar{R}^2 does not necessarily mean that you have the most appropriate set of regressors.

Answer: c

12. If the estimates of the coefficients of interest change substantially across specifications,

- a. then this can be expected from sample variation.
- b. then you should change the scale of the variables to make the changes appear to be smaller.
- c. then this often provides evidence that the original specification had omitted variable bias.
- d. then choose the specification for which your coefficient of interest is most significant.

Answer: c

13. The interpretation of the slope coefficient in the model $\ln(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i$ is as follows:

- a. a 1% change in X is associated with a $\beta_1\%$ change in Y.
- b. a change in X by one unit is associated with a $100\beta_1\%$ change in Y.
- c. a 1% change in X is associated with a change in Y of $0.01\beta_1$.
- d. a change in X by one unit is associated with a β_1 change in Y.

Answer: b

14. To test whether or not the population regression function is linear rather than a polynomial of order r ,

- a. check whether the regression for the polynomial regression is higher than that of the linear regression.
- b. compare the TSS from both regressions.
- c. look at the pattern of the coefficients: if they change from positive to negative to positive, etc., then the polynomial regression should be used.
- d. use the test of $(r - 1)$ restrictions using the F-statistic.

Answer: d

15. In the equation $\widehat{TestScore} = 607.3 + 3.85Income - 0.0423Income^2$, the following income level results in the maximum test score

- a. 607.3.
- b. 91.02.
- c. 45.50.
- d. cannot be determined without a plot of the data.

Answer: c

Part II. Written Questions.

Question 1. The probability limit of the OLS estimator in the case of omitted variables is given in your lecture note by the following formula: $\hat{\beta}_1 \rightarrow^p \beta_1 + \rho_{X\epsilon} \frac{\sigma_\epsilon}{\sigma_X}$

Give an intuitive explanation for two conditions under which the bias will be small.

Answer: The bias will be small if there is little correlation between the included variable and the error term. The error term contains the omitted variable. If the omitted variable is correlated with the included variable, then the error term is correlated with the included variable. Now consider the case where the correlation between the included and omitted variable is low, resulting in a low correlation between the error term and the included variable. In that case, changes in the omitted variable will not result in changes in the included variable, which, in return, changes Y, and making it appear as if the included variable had changed Y.

The second condition is the size of the ratio of the two standard deviations. The formula suggests that if the included variable varies substantially more than the error term, which contains the omitted variable, then the inconsistency will be small. In that case, the relationship between the included variable and the dependent variable does not get disturbed much by variations in the omitted variable.

Question.2 One of your peers wants to analyze whether or not participating in varsity sports lowers or increases the GPA of students. She decides to collect data from 110 male and female students on their GPA and the number of hours they spend participating in varsity sports. The coefficient in the simple regression function turns out to be significantly negative, using the t-statistic and carrying out the appropriate hypothesis test. Upon reflection, she is concerned that she did not ask the students in her sample whether or not they were female or male. You point out to her that you are more concerned about the effect of omitted variables in her regression, such as the incoming SAT score of the students, and whether or not they are in a major from a high/low grading department. Elaborate on your argument.

Answer: The presence of omitted variables will result in an inconsistent estimator for the included variable (number of hours spent in varsity sports) if at least one of the following two conditions holds: the omitted variable is relevant in affecting the GPA and/or the omitted variable is correlated with the included variable. Incoming SAT scores are clearly relevant in predicting GPAs, at least in the earlier years. Hence it is relevant. Departmental differences in the general level of grading will even more obviously have an effect on the GPA. The relationship therefore suffers from omitted variable bias.

Question 3. Females, it is said, make 70 cents to the dollar in the United States. To investigate this phenomenon, you collect data on weekly earnings from 1,744 individuals, 850 females and 894 males. Next, you calculate their average weekly earnings and find that the females in your sample earned \$346.98, while the males made \$517.70.

a) Calculate the female earnings in percent of the male earnings. How would you test whether or not this difference is statistically significant?

Answer: Female earnings are at 67 percent of male earnings. Run a regression of earnings on a constant

and a binary variable, which takes on the value of one for females and is zero otherwise and then use a t-test on the slope of the binary variable for statistical significance.

b) A peer suggests that this is consistent with the idea that there is discrimination against females in the labor market. What is your response?

Answer: Differences in attributes of the individuals, such as education, ability, and tenure with an employer, have not been taken into account. Hence, in itself, this is weak evidence, at best, for discrimination.

c) You recall that additional years of experience are supposed to result in higher earnings. You reason that this is because experience is related to “on the job training.” One frequently used measure for (potential) experience is “Age-Education-6.” Explain the underlying rationale. Assuming that education is constant across the 1,744 individuals, you consider regressing earnings on age and a binary variable for gender. You estimate two specifications initially:

$$\begin{aligned}\widehat{Earn} &= 323.70 + 5.15Age - 169.78Female, & R^2 &= 0.13, SER = 274.75 \\ &(21.18) \quad (0.55) \quad (13.06) \\ \ln(\widehat{Earn}) &= 5.44 + 0.015Age - 0.421Female, & R^2 &= 0.17, SER = 0.75 \\ &(0.08) \quad (0.002) \quad (0.036)\end{aligned}$$

where *Earn* are weekly earnings in dollars, *Age* is measured in years, and *Female* is a binary variable, which takes on the value of one if the individual is a female and is zero otherwise. Interpret each regression carefully. For a given age, how much less do females earn on average? Should you choose the second specification on grounds of the higher regression R^2 ?

Answer: The potential experience variable is a reasonable proxy for “on the job training” if the individual started to work after completing her or his education, and stayed employed thereafter. Hence this is a better proxy for some than for others.

The linear specification suggests that for every additional year the individual receives \$5.15 of additional weekly earnings on average. Females make \$167.78 less than males at a given age. There is no data close to the origin, so the intercept should not be interpreted. The regression explains 13 percent of the variation in earnings.

The log-linear specification says that earnings increase by 1.5 percent for every additional year in an individual’s life. Females earn approximately 42.1 percent less than males at a given age. Again, the intercept should not be interpreted. The regression explains 17 percent of the variation in the log of earnings. You should not prefer this specification over the linear one on grounds of the higher regression R^2 since these cannot be compared as a result of the difference in the units of measurement of the dependent variable.

d) Your peer points out to you that age-earning profiles typically take on an inverted U-shape. To test this idea, you add the square of age to your log-linear regression.

$$\widehat{\ln(\text{Earn})} = 3.04 + 0.147\text{Age} - 0.421\text{Female} - 0.0016\text{Age}^2, \quad R^2 = 0.28, \text{SER} = 0.68$$

(0.18) (0.009) (0.033) (0.0001)

Interpret the results again. Are there strong reasons to assume that this specification is superior to the previous one? Why is the increase of the Age coefficient so large relative to its value in (c)?

Answer: The coefficient on the added variable is statistically significant and has resulted in a substantial increase in the regression R^2 . The increase in the Age coefficient is due to the fact that earnings increase more initially than later in life or, mathematically speaking, it compensates for the negative coefficient on Age^2 , which lowers earnings as individuals become older.

e) What other factors may play a role in earnings determination?

Answer: Students' answers will differ, but education, ability, regional differences, race, and professional choice are often mentioned.

Question.4 Use the data set CPS04.dat to answer following questions.

a. Run a regression of average hourly earning (AHE) on age (Age), gender (Female), and education (Bachelor). If Age increase from 25 to 26, how are earnings expected to change? If Age increase from 33 to 34, how are earnings expected to change?

```
. regress ahe age female bachelor
```

| Source | SS | df | MS | Number of obs = 7986 | | |
|----------|------------|------|------------|----------------------|---|--------|
| Model | 116386.54 | 3 | 38795.5133 | F(3, 7982) | = | 624.10 |
| Residual | 496180.729 | 7982 | 62.1624566 | Prob > F | = | 0.0000 |
| ----- | | | | R-squared | = | 0.1900 |
| Total | 612567.269 | 7985 | 76.7147487 | Adj R-squared | = | 0.1897 |
| ----- | | | | Root MSE | = | 7.8843 |

| ahe | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|----------|-----------|-----------|--------|-------|----------------------|-----------|
| age | .4392042 | .0305286 | 14.39 | 0.000 | .3793601 | .4990482 |
| female | -3.157864 | .1803647 | -17.51 | 0.000 | -3.511426 | -2.804302 |
| bachelor | 6.86515 | .1783686 | 38.49 | 0.000 | 6.515501 | 7.214799 |
| _cons | 1.883798 | .9202918 | 2.05 | 0.041 | .0797852 | 3.68781 |

Answer: If Age increases from 25 to 26, earnings are predicted to increase by \$0.439 per hour. If Age increases from 33 to 34, earnings are predicted to increase by \$0.439 per hour. These values are the same because the regression is a linear function relating AHE and Age.

b. Run a regression of the logarithm average hourly earnings, $\ln(AHE)$, on Age, Female, and Bachelor. If Age increase from 25 to 26, how are earnings expected to change? If Age increase from 33 to 34, how are earnings expected to change? (Stata: generate $\ln ahe = \ln(ahe)$)

```
generate lnahe=ln(ahe)
regress lnahe age female bachelor
```

| Source | SS | df | MS | | | |
|----------|------------|------|------------|-----------------|--------|--|
| Model | 397.245741 | 3 | 132.415247 | Number of obs = | 7986 | |
| Residual | 1667.74691 | 7982 | .208938476 | F(3, 7982) = | 633.75 | |
| | | | | Prob > F = | 0.0000 | |
| | | | | R-squared = | 0.1924 | |
| | | | | Adj R-squared = | 0.1921 | |
| | | | | Root MSE = | .4571 | |
| Total | 2064.99265 | 7985 | .258608974 | | | |

| lnahe | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|----------|-----------|-----------|--------|-------|----------------------|-----------|
| age | .0244429 | .0017699 | 13.81 | 0.000 | .0209735 | .0279124 |
| female | -.1804636 | .0104567 | -17.26 | 0.000 | -.2009616 | -.1599657 |
| bachelor | .4052749 | .010341 | 39.19 | 0.000 | .3850038 | .425546 |
| _cons | 1.856457 | .0533545 | 34.79 | 0.000 | 1.751868 | 1.961046 |

Answer: If Age increases from 25 to 26, $\ln(AHE)$ is predicted to increase by 0.024. This means that earnings are predicted to increase by 2.4%. If Age increases from 34 to 35, $\ln(AHE)$ is predicted to increase by 0.024. This means that earnings are predicted to increase by 2.4%. These values, in percentage terms, are the same because the regression is a linear function relating $\ln(AHE)$ and Age.

c. Run a regression of the logarithm average hourly earnings, $\ln(AHE)$, on $\ln(Age)$, Female, and Bachelor. If Age increase from 25 to 26, how are earnings expected to change? If Age increase from 33 to 34, how are earnings expected to change?

```
. generate lnage=ln(age)
. regress lnahe lnage female bachelor
```

| Source | SS | df | MS | | | |
|----------|------------|------|------------|-----------------|--------|--|
| Model | 397.892296 | 3 | 132.630765 | Number of obs = | 7986 | |
| Residual | 1667.10036 | 7982 | .208857474 | F(3, 7982) = | 635.03 | |
| | | | | Prob > F = | 0.0000 | |
| | | | | R-squared = | 0.1927 | |
| | | | | Adj R-squared = | 0.1924 | |
| | | | | Root MSE = | .45701 | |
| Total | 2064.99265 | 7985 | .258608974 | | | |

| lnahe | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|----------|-----------|-----------|--------|-------|----------------------|-----------|
| lnage | .0244429 | .0017699 | 13.81 | 0.000 | .0209735 | .0279124 |
| female | -.1804636 | .0104567 | -17.26 | 0.000 | -.2009616 | -.1599657 |
| bachelor | .4052749 | .010341 | 39.19 | 0.000 | .3850038 | .425546 |
| _cons | 1.856457 | .0533545 | 34.79 | 0.000 | 1.751868 | 1.961046 |

| lnage | | .7246973 | .0520447 | 13.92 | 0.000 | .6226762 .8267184 |
|----------|--|-----------|----------|--------|-------|---------------------|
| female | | -.1802958 | .010455 | -17.24 | 0.000 | -.2007903 -.1598013 |
| bachelor | | .4052329 | .010339 | 39.19 | 0.000 | .3849657 .4255 |
| _cons | | .1282838 | .17662 | 0.73 | 0.468 | -.2179376 .4745051 |

Answer: If Age increases from 25 to 26, then $\ln(\text{Age})$ has increased by $\ln(26) - \ln(25) = 0.0392$ (or 3.92%). The predicted increase in $\ln(\text{AHE})$ is $0.725 \times (.0392) = 0.0284$. This means that earnings are predicted to increase by 2.8%. If Age increases from 34 to 35, then $\ln(\text{Age})$ has increased by $\ln(35) - \ln(34) = .0290$ (or 2.90%). The predicted increase in $\ln(\text{AHE})$ is $0.725 \times (0.0290) = 0.0210$. This means that earnings are predicted to increase by 2.10%.

d. Run a regression of the logarithm average hourly earning, $\ln(\text{AHE})$ on, Age, Age^2 , Female, and Bachelor. If Age increase from 25 to 26, how are earnings expected to change? If Age increase from 33 to 34, how are earnings expected to change?

```
. generate agesqr=age*age
. regress lnawe age agesqr female bachelor
```

| Source | SS | df | MS | Number of obs = | 7986 |
|----------|------------|------|------------|-----------------|--------|
| Model | 399.069759 | 4 | 99.7674398 | F(4, 7981) = | 477.96 |
| Residual | 1665.9229 | 7981 | .20873611 | Prob > F = | 0.0000 |
| Total | 2064.99265 | 7985 | .258608974 | R-squared = | 0.1933 |
| | | | | Adj R-squared = | 0.1929 |
| | | | | Root MSE = | .45688 |

| lnawe | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|----------|-----------|-----------|--------|-------|----------------------|
| age | .1470452 | .0415124 | 3.54 | 0.000 | .0656701 .2284203 |
| agesqr | -.0020706 | .0007004 | -2.96 | 0.003 | -.0034436 -.0006975 |
| female | -.1797868 | .0104542 | -17.20 | 0.000 | -.2002797 -.1592938 |
| bachelor | .4050769 | .0103362 | 39.19 | 0.000 | .3848152 .4253386 |
| _cons | .0587333 | .6104789 | 0.10 | 0.923 | -1.137965 1.255431 |

Answer: When Age increases from 25 to 26, the predicted change in $\ln(\text{AHE})$ is $(0.147 \times 26 - 0.0021 \times 26^2) - (0.147 \times 25 - 0.0021 \times 25^2) = 0.0399$. This means that earnings are predicted to increase by 3.99%. When Age increases from 34 to 35, the predicted change in $\ln(\text{AHE})$ is $(0.147 \times 35 - 0.0021 \times 35^2) - (0.147 \times 34 - 0.0021 \times 34^2) = 0.0063$. This means that earnings are predicted to increase by 0.63%.

e. Do you prefer the regression in (c) to the regression in (b)? Explain.

Answer: The regressions differ in their choice of one of the regressors. They can be compared on the basis of the \bar{R}^2 . The regression in (3) has a (marginally) higher \bar{R}^2 so it is preferred.

f. Do you prefer the regression in (d) to the regression in (b)? Explain.

Answer: The regression in (4) adds the variable Age^2 to regression (2). The coefficient on Age^2 is statistically significant ($t = -2.96$), and this suggests that the addition of Age^2 is important. Thus, (4) is preferred to (2).

g. Do you prefer the regression in (d) to the regression in (c)? Explain.

The regressions differ in their choice of one of the regressors. They can be compared on the basis of the \bar{R}^2 . The regression in (4) has a (marginally) higher \bar{R}^2 so it is preferred.