ECON6560 : Practice Midterm Solutions

Do all the questions. Good Luck

I Multiple Choice Questions (40 points, choose the best answer)

1. An estimator $\hat{\theta}$ of a parameter θ is consistent if

   (a) It has low variance.

   (b) It is unbiased and its variance approaches zero as sample size tends to infinity.

   (c) It takes on the same value in every sample.

   (d) It has low mean squared errors.

2. An estimator $\hat{\theta}$ of a parameter θ is unbiased if

   (a) it gives a correct estimate of the population parameter θ for any sample size.

   (b) it gives a correct estimate of the population parameter θ on average.

   (c) it converges to the population parameter of interest as the sample size goes to infinity.

   (d) It has low mean squared errors.

3. Suppose we set up the following hypothesis test framework about the mean of a population that is normally distributed: $H_0 : \mu = 0$ vs $H_1 : \mu < 0$.

   A random sample is collected and the sample mean $\bar{x} = 1.2$. Which of the following statement is true?

   (a) We will not reject the null regardless of the size of the standard error

   (b) We will reject the null in favour of the alternative if the standard error is sufficiently small.

   (c) Not enough information to make a decision.

   (d) None of the above

4. Suppose you construct a 95% (two sided) confidence interval around a sample mean, Which of the following statement is **false**?

   (a) Before you draw your random sample and calculate the coefficient estimate, the probability that the true coefficient value will lie in the confidence interval is 0.95.

   (b) Once you've calculated your confidence interval, the probability that the true value of the coefficient lies outside the interval is 0.05.

   (c) If you recalculate the confidence interval for the same sample at the 90% level, the interval will be narrower.

   (d) If you draw a new sample of data and recalculate the 95% confidence interval, its width will probably change with the new data.

5. Suppose you obtain the following fitted model:

$$\widehat{Grade}_i = \hat{\beta}_1 + \hat{\beta}_2 IQ_i + \hat{\beta}_3 ST_i,$$

   where *Grade* is the numerical scores (grade) of students in ECON4570/6560, IQ is the IQ score, ST is the time (measured in hours) student spent on ECON4570/6560.

   $\hat{\beta}_1$ is an estimate of

(a) the average grade of student in ECON4570/6560, when $IQ = 0$ and $ST = 0$.

(b) the average IQ of student in ECON4570/6560.

(c) the average time of student in ECON4570/6560. spent on this course.

(d) a weight average of IQ and study time of student in ECON4570/6560.

6. Using the information in Q5, $\hat{\beta}_2$ is an estimate of

(a) the average effect on grade for extra unit of IQ score when $ST = 0$.

(b) the average grade of students in ECON4570/6560 when $IQ = 1$ and $ST = 0$..

(c) the average effect of extra unit of IQ score on grade while holding ST constant.

(d) None of above

7. Suppose that you estimate the model $y = \beta_1 + \beta_2 x + \varepsilon$. You calculate residuals and find that the sum of squares due to regression is 400 and the total sum of squares is 1200. The R^2 is

(a) 1/4

(b) 1/3

(c) 1/2

(d) 2/3

8. Suppose that you estimate the model $y = \beta_1 + \beta_2 x + \varepsilon$. You calculate residuals and find that the sum of squares due to regression (SSR) is 400 and the total sum of squares (SST) is 1200 with sample size $n = 50$. The **adjusted** R^2 is

(a) 0.667

(b) 0.333

(c) 0.319

(d) none of above

9. Suppose the true population model of y is given by $y = \beta_1 + \beta_2 x + \varepsilon$. Which of the following will lead to a **higher** variance of the OLS estimator $\hat{\beta}_2$

(a) A smaller sample size.

(b) Less variation in x_2.

(c) larger variation in ε.

(d) All of the above

10. 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 highly unstable.

(d) (b) and (c)

11. Suppose the true population model of y is given by $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where x_2 and x_3 are correlated. But you actually estimate the following model: $y = \gamma_1 + \gamma_2 x_2 + e$ instead. Then

(a) γ_1 is an unbiased estimator for β_1

(b) γ_1 is a consistent estimator for β_1

(c) $\hat{\gamma}_2$ is a biased and inconsistent estimator for β_2

(d) $\hat{\gamma}_2$ is still an unbiased but inefficient estimator for β_2

12. A set of data whose histogram is extremely skewed yields a mean and standard deviation of 70 and 12, respectively. What is the minimum proportion of observations that lies between 46 and 94?

    (a) 25%.

    (b) 50%.

    (c) 75%.

    (d) Need more information concerning the distribution of data before making a decision.

13. Suppose the classical assumptions hold and the true population model of y_i is given by $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ with the sample mean of $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0$ and $\beta_1 \neq 0$. But by accident you estimate it $\hat{y}_i = \hat{\beta}_2 x_i$. with $\hat{\varepsilon}_i = y_i - \hat{y}_i$. Which of the following statement is **false**?

    (a) $\hat{\beta}_2$ is unbiased for β_2

    (b) $\sum_{i=1}^{n} \hat{\varepsilon}_i x_i = 0$

    (c) $\sum_{i=1}^{n} \hat{\varepsilon}_i \hat{y}_i = 0$

    (d) $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

14. The statistical significance of a parameter in a regression model refers to:

    (a) The conclusion of testing the null hypothesis that the parameter is equal to zero, against the alternative that it is non-zero.

    (b) The OLS estimate of this parameter is equal to zero.

    (c) The probability that the OLS estimate of this parameter equal to the true parameter is nonzero.

    (d) None of the above

15. The level of significance is the

    (a) maximum allowable probability of Type II error .

    (b) maximum allowable probability of Type I error

    (c) same as the confidence level

    (d) same as the p-value

16. The interpretation of the slope coefficient in the model : $\log(y) = \beta_1 + \beta_2 \log(x) + \varepsilon$ is as follows: a

    (a) change in x by one unit is associated with a $100\beta_2$% change in y.

    (b) change in x by one unit is associated with a β_2% change in y.

    (c) 1% change in x is associated with a β_2% change in y.

    (d) 1% change in x is associated with a change in y of $0.01\beta_2$ (unit)

17. The following linear hypothesis can be tested using the F-test with the exception of

    (a) $\beta_2 = 0$

    (b) $\beta_2 = \beta_3 = 0$

    (c) $\beta_2 + 2\beta_3 = 1$

    (d) $\beta_2 = \frac{1}{\beta_3}$

18. Suppose the classical assumptions hold and you run the OLS regressions for $y = \beta_1 + \beta_2 x + \varepsilon$ and $x = \alpha_1 + \alpha_2 y + \xi$ using the same set of data. You obtain the OLS estimates $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2$ with $\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, and $\hat{\sigma}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$. Then which of the following is true?

(a) $\hat{\alpha}_2\hat{\beta}_2 = 1$

(b) $\hat{\alpha}_1 = \hat{\beta}_1$

(c) $\boxed{\hat{\alpha}_2\hat{\sigma}_y^2 = \hat{\beta}_2\hat{\sigma}_x^2}$

(d) None of the above

19. Suppose the classical assumptions hold and you run the simple linear regression for the model: $y_i = \alpha + \beta x_i + \varepsilon$ using the sample data with sample size $n = 100$. You obtain estimates as $\hat{\alpha} = 2$, $\hat{\beta} = 0.8$, $\hat{\sigma}^2 = 4$ and $s.e(\hat{\beta}) = 0.108$. Now you want to test the the null hypothesis $H_0 : \beta = 0$ against the alternative hypothesis that $H_1 : \beta \neq 0$ using F test, you conclude that at 5% significance level

(a) H_0 is not rejected

(b) $\boxed{H_0 \text{ is rejected}}$

(c) the F test is not applicable in this case

(d) the information given is not sufficient to make a decision regrading H_0 and H_1

20. If you reject a joint null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ using the F-test in a multiple hypothesis setting $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, then

(a) the F-statistic must be negative.

(b) R^2 is close to one.

(c) all of the hypotheses are always simultaneously rejected.

(d) $\boxed{\text{a series of individual t-tests may or may not give you the same conclusion.}}$

II Questions (60 points)

1. (20 points) There are 100 children in a small town. Someone wants to know what fraction of the children have obesity. They take a random sample of 50 children. 30 out of 50 children had obesity.

(a) Suppose that, in fact, 70 children in the town have obesity. Find the probability that more than 40 children have obesity in a random sample of 50 children.

Answer: let

$$X_i = \begin{cases} 1 & \text{if obesity} \\ 0 & \text{no obesity} \end{cases},$$

Then $\sum_{i=1}^{n} X_i$ denote the total number of kids have obesity and $\sum_{i=1}^{n} X_i \sim B(n,p)$ with $p = \frac{70}{100} = 0.7$ and $n = 50$. That is, $\sum_{i=1}^{50} X_i \sim B(50, p(1-p))$. because of the large numbers involved, we should approximate the binomial to a normal distribution. We know that a binomial can be approximated to $N(35, 10.5)$ because sample size is 50 large enough to apply central limit theorem. Thus

$$\begin{aligned} P(40 \leq \sum_{i=1}^{50} X_i \leq 50) &= P\left(\frac{40-35}{\sqrt{10.5}} \leq \frac{\sum_{i=1}^{50} X_i - 35}{\sqrt{10.5}} \leq \frac{50-35}{\sqrt{10.5}}\right) \\ &= P(1.53 \leq Z \leq 4.63) \\ &\approx 0.063 \end{aligned}$$

(b) Obtain a 90% confidence interval for the proportion of children in the town with obesity.

Answer: 30 out of 50 had obesity. So, $\hat{p} = \frac{3}{5} = 0.6$. We know that for a large $n = 50$, 90% confidence interval for a proportion is

$$\begin{aligned} \hat{p} \pm z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.6 \pm 1.645\sqrt{\frac{0.6 \times 0.4}{50}} \\ &= (0.486, 0.714) \end{aligned}$$

2. (20 points) Sir Francis Galton, a cousin of James Darwin, examined the relationship between the height of children and their parents towards the end of the 19th century. It is from this study that the name "regression" originated. His data contains 896 children. Let *height* denote the height of a child and $Midheight$ the average height of his parents. (Following Galton's methodology, both variables were adjusted so that the average female height was equal to the average male height.) The estimated relationship is given by

$$\widehat{height}_i = \hat{\beta}_1 + \hat{\beta}_2 Midheight_i$$

with $\hat{\beta}_1 = 22.26$, $\hat{\beta}_2 = 0.668$, $R^2 = 0.104$, $s = \sqrt{SSE/(n-2)} = 3.39$ and $s.e(\hat{\beta}_2) = 4.369$. the mean of Midheight is $\overline{Midheight} = 66.75$.

(a) Interpret $\hat{\beta}_1$ and $\hat{\beta}_2$.

Answer: $\hat{\beta}_1 = 22.26$ means that when the $Midheight = 0$, the average height for children is 22.26, which is meaningless as we have have any sample with Midheight being close to zero.
$\hat{\beta}_2 = 0.668$ means that when $Midheight_i$ increases by one unit, the height for children will increase by 0.668 unit on average.

(b) Interpret R^2 and compute adjusted R^2.

Answer: $R^2 = 0.104$ means about 10.4% variation in children's height is explained by the variable Midheight in the simple linear regression model. The R^2 is a bit low in this model, indicating that other (missing) variables may be needed and/or the functional form may not be linear.
$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} = 1 - 0.896 \times \frac{895}{894} = 0.103$.

(c) Compute the confidence interval for $E(height_i)$ when $Midheight_i = 70$.

Answer: when $Midheight_i = 70$, $\widehat{height}_i = 22.26 + 0.668 \times 70 = 69.02$. Now since

$$s.e(\hat{\beta}_2) = \frac{s}{\sqrt{\sum_{i=1}(Midheight_i - \overline{Midheight})^2}}$$

so

$$\sum_{i=1}(Midheight_i - \overline{Midheight})^2 = \left(\frac{s}{s.e(\hat{\beta}_2)}\right)^2 = \left(\frac{3.39}{4.369}\right)^2 = 0.602$$

In addition, $(Midheight_i - \overline{Midheight})^2 = (70 - 66.75)^2 = 10.563$ Thus

$$\begin{aligned} s.e.(\widehat{height}_i) &= s\sqrt{\frac{1}{n} + \frac{(Midheight_i - \overline{Midheight})^2}{\sum_{j=1}(Midheight_j - \overline{Midheight})^2}} \\ &= 3.39\sqrt{\frac{1}{896} + \frac{10.563}{0.602}} = 14.201 \end{aligned}$$

So the 95% confidence interval for $E(height_i)$ when $Midheight_i = 70$ is then given by

$$\begin{aligned} & [69.02 - t_{894,0.025} \times s.e.(\widehat{height}_i),\ 69.02 + t_{894,0.025} \times s.e.(\widehat{height}_i)] \\ = \ & [69.02 - z_{0.025} \times s.e.(\widehat{height}_i),\ 69.02 + z_{0.025} \times s.e.(\widehat{height}_i)] \\ = \ & [69.02 - 1.96 \times 14.201,\ 69.02 + 1.96 \times 14.201] \\ = \ & [41.186,\ 96.854] \end{aligned}$$

(d) Compute the prediction interval for $height_i$ when $Midheight_i = 70$.

Answer: when $Midheight_i = 70$, $\widehat{height}_i = 22.26 + 0.668 \times 70 = 69.02$.

$$\begin{aligned} s.e.(forecast\ error) &= s\sqrt{1 + \frac{1}{n} + \frac{(Midheight_i - \overline{Midheight})^2}{\sum_{j=1}(Midheight_j - \overline{Midheight})^2}} \\ &= 3.39\sqrt{1 + \frac{1}{896} + \frac{10.563}{0.602}} = 14.6 \end{aligned}$$

So So the 95% prediction interval for $height_i$ when $Midheight_i = 70$ is then given by

$$\begin{aligned} & [69.02 - t_{894,0.025} \times s.e.(forecast\ error),\ 69.02 + t_{894,0.025} \times s.e.(forecast\ error)] \\ = \ & [69.02 - z_{0.025} \times s.e.(forecast\ error),\ 69.02 + z_{0.025} \times s.e.(forecast\ error)] \\ = \ & [69.02 - 1.96 \times 14.6,\ 69.02 + 1.96 \times 14.6] \\ = \ & [40.404,\ 97.636] \end{aligned}$$

3. (20 points) The table below has four models relating the college grade point average (COLGPA) of 427 students to their high school grade point average (HSGPA) and to their verbal and math scores in the Scholastic Aptitude Test (VSAT and MSAT). A number of dummy variables are also included: DCAM=1 if the student lived on campus, DPUB=1 if the student graduated from a public high school, and several dummy variables for major categories: science (DSCI), social science (DSOC), humanities (DHUM), and arts (DARTS). (The values in parentheses are standard errors.)

Estimated Models for the Grade Point Average Data

Variable	Model A	Model B	Model C	Model D
CONSTANT	0.357	0.363	0.423	0.422
	(0.224)	(0.224)	(0.220)	(0.221)
HSGPA	0.406	0.414	0.398	0.389
	(0.063)	(0.062)	(0.061)	(0.062)
VSAT	0.00073	0.00068	0.00074	0.00079
	(0.00029)	(0.00029)	(0.00028)	(0.00029)
MSAT	0.0011	0.0011	0.0010	0.0010
	(0.0003)	(0.0003)	(0.0003)	(0.0003)
DSCI	−0.027	−0.026		
	(0.057)	(0.057)		
DSOC	0.056	0.054		
	(0.073)	(0.073)		
DHUM	−0.0041	−0.0068		
	(0.142)	(0.141)		
DARTS	0.229	0.243		
	(0.189)	(0.188)		
DCAM	−0.041			−0.040
	(0.052)			(0.052)
DPUB	0.029			0.033
	(0.063)			(0.063)
SSE	96.204	96.421	97.164	96.932
$\bar{R}^2$	0.211	0.213	0.215	0.213

Useful facts: $t_{417}(0.025) \simeq 1.96$, $t_{421}(0.025) \simeq 1.96$, $P(|t| > 0.788) \simeq 0.44$, $P(|t| > 0.769) \simeq 0.44$, $P(|t| > 0.46) \simeq 0.64$. $P(|t| > 0.52) \simeq 0.60$, $F_{4,417}(0.10) = 1.94$, $F_{6,417}(0.10) = 1.77$.

(a) In Model A, using a t test to check whether there is any difference due to the students living on campus or off campus at 5% significance level. State the null and alternative hypotheses, the test statistic and its distribution, and the critical value.

Answer: $H_0 : \beta_{DCAM} = 0$ vs $H_1 : \beta_{DCAM} \neq 0$

For Model A, under H_0

$$t = \frac{\hat{\beta}_{DCAM}}{s.e.(\hat{\beta}_{DCAM})} \sim t_{417}$$

At $\alpha = 5\%$ level, the critical value is $t_{417}(0.025) \simeq 1.96$. Since $t = \frac{\hat{\beta}_{DCAM}}{s.e.(\hat{\beta}_{DCAM})} = \frac{-0.041}{0.052} = -0.788$, and hence $|t| = 0.788 < t_{417}(0.025)$, we do not reject the H_0 at 5% level.

(b) In Model A, test whether there is any difference due to the students having graduated from a public or other type of school at 5% significance level. State the null and alternative hypotheses, the test statistic and its distribution.

Answer: $H_0 : \beta_{DPUB} = 0$ vs $H_1 : \beta_{DPUB} \neq 0$

For Model A, under H_0

$$t = \frac{\hat{\beta}_{DPUB}}{s.e.(\hat{\beta}_{DPUB})} \sim t_{417}$$

At $\alpha = 5\%$ level, the critical value is $t_{417}(0.025) \simeq 1.96$. Since $t = \frac{\hat{\beta}_{DPUB}}{s.e.(\hat{\beta}_{DPUB})} = \frac{0.029}{0.063} = 0.46$, and hence $|t| = 0.46 < t_{417}(0.025)$, we do not reject the H_0 at 5% level.

(c) In Model A, test the hypothesis that all of the dummy variables have zero regression coefficients at 10% significance level.

Answer: $H_0 : \beta_{DSCI} = \beta_{DSOC} = \beta_{DHUM} = \beta_{DARTS} = \beta_{DCAMP} = \beta_{DPUB} = 0$ vs $H_1 : H_0$ is not true.

Restricted model: Model C

Unrestricted Model: Model A

Then Under H_0

$$\begin{aligned} F &= \frac{(SSE_C - SSE_A)/6}{SSE_A/(n-10)} \\ &= \frac{(SSE_C - SSE_A)/6}{SSE_A/(427-10)} \\ &\backsim F_{6,417} \end{aligned}$$

We reject H_0 if the F statistic $F > F_{6,417}(\alpha)$. Now $\alpha = 0.1$ and $F = \frac{(96.421-96.204)/6}{96.204/417} = 0.157$. Since $F_{6,417}(0.10) = 1.77$, $F = 0.157 < F_{6,417}(0.10)$, we do not reject H_0 at 10% significance level based on the statistical evidence!

Note that we can also use the following formula to do the test.

$$\begin{aligned} F &= \frac{(R_A^2 - R_C^2)/6}{(1-R_A^2)/417} \\ &\sim F_{6,417} \text{ under } H_0. \end{aligned}$$

The use $R^2 = 1 - \frac{n-k}{n-1}(1-\bar{R}^2)$ to compute R_A^2 and R_C^2 and hence the F statistic.

(d) In Model A, test whether the dummy variables for majors are jointly significant at the 10 percent significance level. Be sure to state the null and alternative hypotheses, the test statistic and its distribution, and the critical value.

Answer: $H_0 : \beta_{DSCI} = \beta_{DSOC} = \beta_{DHUM} = \beta_{DARTS} = 0$ vs $H_1 : H_0$ is not true.

Restricted model: Model D

Unrestricted Model: Model A

$$\begin{aligned} F &= \frac{(SSE_D - SSE_A)/4}{SSE_A/(n-10)} \\ &= \frac{(SSE_D - SSE_A)/4}{SSE_A/(427-10)} \\ &\sim F_{4,417} \text{ under } H_0. \end{aligned}$$

We reject H_0 if the F statistic $F > F_{4,417}(\alpha)$. Now $\alpha = 0.1$ and

$$F = \frac{(96.932 - 96.204)/4}{96.204/417} = 0.789 < F_{4,417}(0.10) = 1.94$$

we do not reject H_0 at 10% level based on the statistical evidence.
Again, Note that we can also use the following formula to do the test.

$$\begin{aligned} F &= \frac{(R_A^2 - R_D^2)/4}{(1 - R_A^2)/417} \\ &\sim F_{4,417} \text{ under } H_0. \end{aligned}$$

The use $R^2 = 1 - \frac{n-k}{n-1}(1 - \bar{R}^2)$ to compute R_A^2 and R_D^2 and hence the F statistic.