《计量经济学导论》期中、期末试题及答案 - 2021 年春季学期 - 丁一

THE CHINESE UNIVERSITY OF HONG KONG, SHENZHEN
School of Management and Economics

ECONOMICS 3121
Introductory Econometrics

Spring term 2021

MIDTERM EXAM

DATE: **Saturday March 27, 2021**

TIME: **150 minutes; 2:00 p.m. – 4:30 p.m.**

INSTRUCTIONS: The exam consists of **FIVE (5)** questions. Students are required to answer **ALL FIVE (5)** questions.

Answer all questions in the exam booklets provided. Be sure your ***name*** and ***student number*** are printed clearly on the front of all exam booklets used.

Do not write answers to questions on the front page of the first exam booklet.

Please label clearly each of your answers in the exam booklets with the appropriate number and letter.

Please write legibly.

This exam is **CONFIDENTIAL**. This question paper must be submitted in its entirety with your answer booklet(s); otherwise your exam will not be marked.

A formula sheet and a table of percentage points of the t-distribution is given on the last pages of the exam.

MARKING: The marks for each question are indicated in parentheses immediately above each question. **Total marks** for the exam **equal 150**.

GOOD LUCK!

Question 1: Multiple Choice (40 points, 2 points each)
Select the **BEST** response to each of the following questions.

1. The OLS estimator is derived by
   a) connecting the y_i corresponding to the lowest x_i observation with the y_i corresponding to the highest x_i observation.
   b) making sure that the standard error of the regression equals the standard error of the slope estimator.
   c) maximizing the R^2.
   d) minimizing the sum of absolute values of residuals.

Answer: c

2. In the multiple linear regression model, bias in the OLS estimator is caused by:
   a) $E(u|x) \neq 0$
   b) $Var(u|x) \neq 0$
   c) $Var(u|x) \neq \sigma^2$
   d) $E(u|x) = E(u) = 0$

Answer: a

3. Which of the following set of random variables are perfectly collinear?
   a) x_1, x_1^2.
   b) $x_1, x_1^2, ln(x_1^2)$.
   c) $x_1,\ ln(x_2 + x_1)$.
   d) y, x_1, x_2

Answer: no answer

4. The sample regression line estimated by OLS (including a constant)
   a) will always have a positive intercept.
   b) is the same as the population regression line.
   c) can run above all data points in the scatter plots of (x_i, y_i).
   d) will always run though the interior of the scatter plots of (x_i, y_i).

Answer: d

5. The normality assumption in the OLS regression is important because:
   a) many explanatory variables in real life are normally distributed.
   b) it allows econometricians to develop methods for statistical inference.
   c) it is necessary for the causal interpretation of the coefficients.
   d) it is necessary for the BLUE property of OLS.

Answer: b

6. A data set that consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken repeatedly over time, is called a _____.
   a) cross-sectional data set
   b) longitudinal data set
   c) time series data set

d) experimental data set

Answer: b

7. Adding the dependent variable by 100 leaves the
   a) OLS estimate of the slope the same.
   b) OLS estimate of the intercept the same.
   c) regression R^2 changed.
   d) variance of the OLS estimators changed.

Answer: a

8. In the classical linear model, the assumption of homoscedasticity is needed for
   (i) unbiasedness
   (ii) simple calculation of variance and standard errors of coefficient estimates
   (iii) the claim that OLS estimator is BLUE
   (iv) hypothesis testing

   a) (i) and (iii) only
   b) (ii) and (iii) only
   c) (ii), (iii), and (iv) only
   d) (i), (ii), (iii), and (iv)

Answer: c

9. Using 151 observations, assume that you had estimated a simple regression function and that your estimate for the slope was 0.04, with a standard error of 0.01. You want to test whether or not the estimate is statistically significant. Which of the following possible decisions is the only correct one:
   a) you decide that the coefficient is small and hence most likely is zero in the population.
   b) the slope is statistically significant since it is four standard errors away from zero.
   c) the response of y given a change in x must be economically important since it is statistically significant.
   d) since the slope is very small, so must be the regression R^2.

Answer: b

10. Suppose the true population model of y is given by $= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$. Which of the following will lead to a higher variance of the OLS estimator, $\hat{\beta}_3$.
   (i) A smaller sample size
   (ii) Lower variation in x_3
   (iii) Greater variation in u
   (iv) Higher correlation between x_1 and x_2
   (v) Higher correlation between x_1 and x_3

   a) (ii) and (v) only
   b) (i), (iii) and (v) only
   c) (i), (ii), (iii), and (v) only
   d) (i), (ii), (iii), (iv), and (v)

Answer: c

11. The R^2 of a regression can be used to test the null hypothesis that
   a) all slope coefficients are zero.
   b) the sample regression line explains 50% of the total variation in y.
   c) the intercept in the regression and at least one, but not all, of the slope coefficients is zero.
   d) the slope coefficient of the variable of interest is zero, but that the other slope coefficients are not.

Answer: a

12. The OLS estimators of the coefficients in multiple regression will have omitted variable bias
   a) only if an omitted determinant of y_i is a continuous variable.
   b) if an omitted variable is correlated with at least one of the regressors, even though it is not a determinant of the dependent variable ($\beta = 0$ for the omitted variable in the population model).
   c) only if the omitted variable is not normally distributed.
   d) if an omitted determinant of y_i is correlated with at least one of the regressors.

Answer: d

13. In a simple log-log regression of $\log(y)$ on $\log(x)$, the estimated coefficient on $\log(x)$ measures
   a) the elasticity of y with respect to x.
   b) the change in x which the model predicts for a unit change in y.
   c) the ratio y/x.
   d) the growth rate of y given x.

Answer: a

14. Suppose the variable x_2 has been omitted from the following regression equation, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. $\widetilde{\beta_1}$ is the estimator obtained when x_2 is omitted from the equation. The bias in $\widetilde{\beta_1}$ is negative if _____.
   a) $\beta_2 > 0$ and x_1 and x_2 are positively correlated
   b) $\beta_2 < 0$ and x_1 and x_2 are positively correlated
   c) $\beta_2 < 0$ and x_1 and x_2 are negatively correlated
   d) $\beta_2 = 0$ and x_1 and x_2 are negatively correlated

Answer: b

15. In a regression model where $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$, to test the hypothesis that $\beta_1 = -\beta_2$ we should:
   a) Regress y on $x_1, (x_2 + x_1), x_3$ and examine the significance of the coefficient for x_1
   b) Regress y on $x_1, (x_2 + x_1), x_3$ and examine the significance of the coefficient for $(x_2 + x_1)$
   c) Regress y on $x_1, (x_2 - x_1), x_3$ and examine the significance of the coefficient for x_1
   d) Regress y on $x_1, (x_2 - x_1), x_3$ and examine the significance of the coefficient for x_3

Answer: c

16. What are the degrees of freedom in Ordinary Least Square (OLS) residuals of multiple linear model $y = \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3 + u$?
    a) 4
    b) n-2
    c) n-3
    d) n-4

Answer: c

17. Which of the following statements is true?
    a) If the p-value is smaller than the significance level, we fail to reject the null hypothesis at the chosen significance level.
    b) The F statistic is always nonnegative as SSRr is never smaller than SSRur.
    c) Degrees of freedom of a restricted model is always less than the degrees of freedom of an unrestricted model.
    d) The F statistic is more flexible than the t statistic to test a hypothesis with a single restriction.

Answer: b

The following regression for Savings (in thousands of dollars) is used in the next three questions.

OLS Dependent Variable: Savings (1000$)
Included observations: 534

Variable	Coefficient	Std. Error	t	P>|t|
$\log(Income)$	0.821491	0.321702	2.553576	0.0109
$Married$	0.837728	0.460136	1.820609	0.0692
Sex	-0.286274	0.135118	-2.118699	0.0346
Age	2.474931	0.586957	4.216548	0.0000
$Intercept$	7.000713	1.071595	6.532983	0.0000

18. In the above regression, which of the variables are significant at the 1% level?
    a) $\log(Income)$, Sex, Age
    b) $\log(Income)$, Sex,
    c) $Married$, sex
    d) Age

Answer: d

19. In the above regression, which of the variables show a significant **positive** effect at the 1% level?
    a) $\log(income)$, Age
    b) Age
    c) $Married$
    d) None of the variables are positively significant at the 1% level

Answer: a

20. According to the above regression results what is the expected change in Savings when Income changes, all else remaining constant?

a) Savings will increase by $821.491 when Income increases by $1000.
b) Savings will increase by $8.21491 when Income increases by 1%.
c) Savings will increase by 82.1491% when Income increases by $1000.
d) Savings will increase by $821.491 when Income increases by 1%

Answer: b

Question 2 (16 marks)

Suppose that you are interested in estimating the ceteris paribus relationship between y and x_1. For this purpose, you can collect data on two control variables, x_2 and x_3. Let $\tilde{\beta}_1$ be the simple regression estimate from y on x_1 and let $\hat{\beta}_1$ be the multiple regression estimate from y on x_1, x_2, x_3.

(4 marks)

(a) If x_1 is highly correlated with x_2 and x_3 in the sample, and x_2 and x_3 have large partial effects on y, would you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar or very different? Explain.

Very different (2 marks). Because $\hat{\beta}_1$ measures the sample relationship between y and x_1 after x_2 and x_3 have been partialled out. If x_1 is highly correlated with x_2 and x_3, and these latter variables have large partial effects on y, then partialling out x_2 and x_3 would largely change the sample relationship between y and x_1. Therefore, the simple and multiple regression coefficients on x_1 can differ by large amounts. **(2 marks)**

(**Or,** one can derive that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1 + \hat{\beta}_3\tilde{\gamma}_1$, where $\tilde{\delta}_1$ is the slope estimator when regress x_2 on x_1 and $\tilde{\gamma}_1$ is the slope estimator when regress x_3 on x_1. If one *writes* $E[\tilde{\beta}_1] = \beta_1 + \beta_2 E[\tilde{\delta}_1] + \beta_3 E[\tilde{\gamma}_1]$, **no marks**)

(4 marks)

(b) If x_1 is uncorrelated with x_2 and x_3 in the sample, but x_2 and x_3 are highly correlated, will $\tilde{\beta}_1$ and $\hat{\beta}_1$ tend to be similar or very different? Explain.

Similar (2 marks). Omitting x_2 and $x_3$won't cause any bias in the estimated coefficient on x_1 if x_1 is uncorrelated with x_2 and x_3. The amount of correlation between x_2 and x_3 does not directly affect the multiple estimated coefficient on x_1. **(2 marks)**

(4 marks)

(c) If x_1 is highly correlated with x_2 and x_3, and x_2 and x_3 have no partial effects on y, which one would you expect to be smaller, $se(\tilde{\beta}_1)$ or $se(\hat{\beta}_1)$? Explain.

$se(\tilde{\beta}_1)$(2marks). In this case, we are (unnecessarily) introducing multicollinearity into the regression. We know that $Var(\tilde{\beta}_1) = \sigma^2/SST_1$ and $Var(\hat{\beta}_1) = \sigma^2/SST_1(1 - R_1^2)$. Adding x_2 and x_3 won't change the residual variance (the numerator) but will largely decrease the denominator by multiplying it with $(1 - R_1^2)$. Therefore, it increases the standard error of the coefficient on x_1, so $se(\hat{\beta}_1)$ is likely to be larger than $se(\tilde{\beta}_1)$. **(2 marks)**

(4 marks)

(d) If x_1 is uncorrelated with x_2 and x_3, x_2 and x_3 have large partial effects on y, and x_2 and x_3 are highly correlated, which one would you expect to be smaller, $se(\tilde{\beta}_1)$ or $se(\hat{\beta}_1)$? Explain.

$se(\hat{\beta}_1)$(2marks). In this case, adding x_2 and x_3 will decrease the residual variance (the numerator) without changing the denominator (because $R_1^2 = 0$ when x_1 is uncorrelated with x_2 and x_3), so we should see $se(\hat{\beta}_1)$ smaller than$se(\tilde{\beta}_1)$.

Question 3 (18 marks)

Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1 through SLR.5. The usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the estimator of β_0 and β_1 obtained by connecting the first two observations (x_1, y_1) and (x_2, y_2). $[\tilde{\beta}_0 = y_1 - \frac{y_2 - y_1}{x_2 - x_1} x_1$ and $\tilde{\beta}_1 = \frac{y_2 - y_1}{x_2 - x_1}]$

(9 marks)

(a) Prove that $\tilde{\beta}_0$and $\tilde{\beta}_1$ are unbiased for β_0 and β_1.

We have the formula of $\tilde{\beta}_1$: $\tilde{\beta}_1 = \frac{y_2 - y_1}{x_2 - x_1}$.

Plugging in $y_i = \beta_0 + \beta_1 x_i + u_i$ gives

$$\tilde{\beta}_1 = \frac{\beta_1 (x_2 - x_1) + (u_2 - u_1)}{x_2 - x_1}$$

(2 marks)

Simplifying,

$$\tilde{\beta}_1 = \beta_1 + \frac{u_2 - u_1}{x_2 - x_1}$$

(1 marks)

Conditional on the x_i , we have

$$\begin{aligned} E(\tilde{\beta}_1 | x) &= \beta_1 + E\left[\frac{u_2}{x_2 - x_1} \middle| x\right] - E\left[\frac{u_1}{x_2 - x_1} \middle| x\right] \\ &= \beta_1 \end{aligned}$$

(2 marks)

For $\tilde{\beta}_0$,

$$\begin{aligned} \tilde{\beta}_0 &= y_1 - \frac{y_2 - y_1}{x_2 - x_1} x_1 \\ &= \beta_0 + u_1 - \frac{u_2 - u_1}{x_2 - x_1} x_1 \end{aligned}$$

(2 marks)

Conditional on x, the second and third terms have zero conditional expectations, so the law of iterated expectations implies $E[\tilde{\beta}_0] = \beta_0$.

(2 marks)

(9 marks)

(b) Find $Var(\tilde{\beta}_1 | x)$, and prove that OLS is more efficient. (Hint: prove that $Var(\tilde{\beta}_1 | x) \geq Var(\hat{\beta}_1 | x)$)

From the expression for $\tilde{\beta}_1$ in part (b) we have, $\tilde{\beta}_1 = \beta_0 \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n {x_i}^2} + \beta_1 + \frac{\sum_{i=1}^n x_i u_i}{\sum_{i=1}^n {x_i}^2}$, conditional on the x_i, we have

$$\begin{aligned} Var(\tilde{\beta}_1 | x) &= Var\left(\frac{u_2 - u_1}{x_2 - x_1} \middle| x\right) \\ &= \frac{2\sigma^2}{(x_2 - x_1)^2} \end{aligned}$$

(4 marks)

We know that

$$Var(\hat{\beta}_1|x) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

(3 marks)

So we need to prove $\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \frac{(x_2 - x_1)^2}{2}$. This can be done by

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 \geq \frac{[(x_2 - \bar{x}) - (x_1 - \bar{x})]^2}{2} = \frac{(x_2 - x_1)^2}{2}.$$

(2 marks)

Question 4 (38 marks)

A researcher is using data for a sample of 274 male employees 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:

$n = 274$ $\quad \sum_{i=1}^{n} y_i = 1945.26$ $\quad \sum_{i=1}^{n} x_i = 1774.00$ $\quad \sum_{i=1}^{n} y_i^2 = 18536.73$

$\sum_{i=1}^{n} x_i^2 = 30608.00$ $\quad \sum_{i=1}^{n} x_i y_i = 16040.72$ $\quad \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 3446.226$

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 4726.377$ $\quad \sum_{i=1}^{n}(x_i - \bar{x})^2 = 19122.32$ $\quad \sum_{i=1}^{n} \hat{u}_i^2 = 4105.297$

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

(8 marks)

(a) Use the above information to compute OLS estimates of the intercept coefficient β_0 and the slope coefficient β_1

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{3446.226}{19122.32} = 0.1802201 = \mathbf{0.18022}$$ **(4 marks)**

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1945.26}{274} = 7.09949 \text{ and } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1774.00}{274} = 6.47445$$

Therefore **(4 marks)**

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 7.09949 - 0.18022 * 6.47445 = 7.09949 - 1.166825 = \mathbf{5.93266}$$

(2 marks)

(b) Interpret the slope coefficient estimate you calculated in part (a) -- i.e., explain what the numeric value you calculated for $\widehat{\beta_1}$ means.

Note: $\widehat{\beta_1} = \mathbf{0.18022}$. y_i is measured in **dollars per hour**, and x_i is measured in **years**.

The estimate **0. 18022** of $\widehat{\beta_1}$ means that an **increase (decrease) in firm tenure of 1 year** is associated on average with **an increase (decrease) in male employees' hourly wage rate equal to 0.18 of dollars per hour,** or **18 cents per hour.**

(4 marks)
(c) Calculate the unbiased estimator of σ^2, the error variance.

$$\hat{\sigma}^2 = \frac{SSR}{n-2} = \frac{\sum_{i=1}^{n} \hat{u}_i^{\,2}}{n-2} = \frac{4105.297}{274-2} = \mathbf{15.0930}$$

(4 marks)
(d) Calculate an estimate of $Var(\hat{\beta}_1)$, the variance of $\hat{\beta}_1$

$$\hat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{15.093}{19122.32} = 0.00078929$$

(6 marks)
(e) Compute the value of R^2, the coefficient of determination for the estimated OLS sample regression equation. Briefly explain what the calculated value of R^2 means.

$$SSE = SST - SSR = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} \hat{u}_i^{\,2} = 4726.377 - 4105.297 = 621.08$$

$$R^2 = \frac{SSE}{SST} = \frac{621.08}{4726.377} = \mathbf{0.1314} \qquad \textbf{(4 marks)}$$

Interpretation of $R^2 = 0.1314$: The value of 0.1314 indicates that 13.14 percent of the total sample (or observed) variation in employees' hourly wage rates is attributable to, or explained by, firm tenure. **(2 marks)**

(8 marks)
(f) Perform a test of the null hypothesis H_0: $\beta_1 = 0$ against the alternative hypothesis H_1: $\beta_1 > 0$ at the 1% significance level (i.e., for significance level $\alpha = 0.01$). Show how you calculated the test statistic. State the decision rule you use, and the inference you would draw from the test. Briefly explain what the test outcome means.

A one-sided alternative hypothesis ⇒ a one-tailed test

Test statistic is $t(\widehat{\beta_1}) = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim t_{n-2}$
Calculate the estimated standard error of $\widehat{\beta_1}$:

$$se(\widehat{\beta_1}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_1 - \bar{x})^2}} = \sqrt{15.093/19122.32} = \mathbf{0.0280943} \qquad \textbf{(1 marks)}$$

Calculate the sample value of the t-statistic (1) under H_0: set $\beta_1 = 0$ in (1).

$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} = \frac{0.18022 - 0}{0.0280943} = \mathbf{6.415}$ **(3 marks)**

Degree of freedom $= n - 2 = 274 - 2 = 272$
Decision Rule: This is a ***one-tail* test**. Compare the sample t statistic with the α-level critical value of the t_{272} distribution.
1. If $|t| \leq c_\alpha$, do not reject H_0 at the significance level α.
2. If $|t| > c_\alpha$, reject H_0 at the significance level α.

Critical value of t_{272} at 1% significance level ($\alpha = 0.01$) is $\boldsymbol{c_{0.01} \approx 2.33}$

Inference: **(3 mark)**
Since $\boldsymbol{t = 6.415 > 2.33 = c_{0.01}}$, **reject H_0** at the **1%** significance level.

Conclusion: **(1 mark)**
The sample evidence **favors the alternative hypothesis that $H_1: \beta_1 > 0$ at the 1% significance level**. It thus provides strong evidence that firm tenures have a positive impact on the hourly wage.

(6 marks)
(g) Compute the two-sided 95% confidence interval for the slope coefficient β_1.

The **two-sided $(1-\alpha)$ -level**, or **$100(1-\alpha)$ percent, confidence interval for** β_1 is computed as

$\widehat{\beta_1} - c_\alpha se(\widehat{\beta_1}) \leq \beta_1 \leq \widehat{\beta_1} + c_\alpha se(\widehat{\beta_1})$ **(2 marks)**

$\widehat{\beta_1} = \mathbf{0.18022}$ $se(\widehat{\beta_1}) = \mathbf{0.0280943}$

Degree of freedom is $n - 2 = 272$
$\alpha = 0.05 \Rightarrow c_\alpha = \mathbf{1.96}$ in a two-tailed test

$c_\alpha se(\widehat{\beta_1}) = 1.960 * 0.0280943 = 0.05506$

Lower 95% confidence limit for β_1 is: **(2 marks)**

$\widehat{\beta_1} - c_\alpha se(\widehat{\beta_1}) = 0.18022 - 0.05506 = \mathbf{0.12516}$

Upper 95% confidence limit for β_1 is: **(2 marks)**

$\widehat{\beta_1} + c_\alpha se(\widehat{\beta_1}) = 0.18022 + 0.05506 = \mathbf{0.23528}$

Result: The two-sided 95% confidence interval for β_1 is: [0.12516, 0.23528]

Question 5 (38 marks)

You are conducting an econometric investigation into the effect on house prices of proximity to a power plant, which presumably generates negative externalities for homeowners and others located close to it. The sample data consist of observations for 321 houses that were sold in a single metropolitan area in the year 2012 on the following variables:

$price_i$: selling price of house i, in dollars.
$hsize_i$: house size of house i, in square meters.
age_i : age of house i, in years.
$dist_i$: distance of house i from power plant, in meters.

Using the given sample data on 321 houses, your trusty research assistant has estimated the following regression equation (1) and obtained the estimation results (with estimated standard errors given in parentheses below the coefficient estimates):

$$price_i = \beta_0 + \beta_1 hsize_i + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 dist_i + \beta_5 dist_i^2 + u_i \quad (1)$$

$$\underset{(12,689)}{\hat{\beta}_0 = 17{,}222} \qquad \underset{(28.54)}{\hat{\beta}_1 = 361.1} \qquad \underset{(163.5)}{\hat{\beta}_2 = -915.7}$$

$$\underset{(1.059)}{\hat{\beta}_3 = 3.743} \qquad \underset{(3.456)}{\hat{\beta}_4 = 8.220} \qquad \underset{(0.000254)}{\hat{\beta}_5 = -0.000695}$$

$$SST = 597{,}850{,}000{,}000 \qquad SSR = 271{,}740{,}000{,}000 \qquad N = 321$$

(8 marks)
(a) Use the estimation results for regression equation (1) to test the joint significance of the slope coefficient estimates $\hat{\beta}_j$ $(j = 1, \ldots, 5)$ in regression equation (1). State the null and alternative hypotheses, and show how you calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level.

This is an overall significance test:

Test: $H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ versus H_1: H_0 is not true **(2 marks)**

Compute F statistic: **(4 marks)**

$$R_{ur}^2 = 1 - \frac{SSR}{SST} = 1 - \frac{271{,}740{,}000{,}000}{597{,}850{,}000{,}000} = \mathbf{0.5455}$$

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)} = \frac{(0.5455 - 0)/5}{(1 - 0.5455)/(321 - 5 - 1)} = \mathbf{75.61}$$

Decision Rule: Compare the sample F statistic with the α-level critical value of the $\mathbf{F_{5,315}}$ distribution.
1. If $F \leq c_\alpha$, do not reject H_0 at the significance level α.
2. If $F > c_\alpha$, reject H_0 at the significance level α.

Critical value of $\mathbf{F_{5,315}}$ at 5% significance level ($\alpha = 0.05$) is $\boldsymbol{c_{0.05} = 2.21}$

Inference: **(2 marks)**

Since $F = 75.61 > 2.21 = c_{0.05}$, **reject H_0** at **5%** significance level.

(10 marks)

(b) Compute a test of the proposition that the distance from the power plant had an increasing marginal effect on mean prices as the distance increased. State the coefficient restrictions on regression equation (1) implied by this proposition; that is, state the null hypothesis H_0 and the alternative hypothesis H_1. Find the p-value of the test (you may write down the expression of the probability without computing the number). Can you reject the null hypothesis at the 10% level?

The **marginal effect of $dist_i$** on house prices is:

$\frac{\partial price_i}{\partial dist_i} = \beta_4 + 2\beta_5 dist_i$

Test: $H_0: \beta_5 = 0$ versus $H_1: \beta_5 > 0$ **(4 marks)**

Compute t statistic:

$$t = \frac{\hat{\beta}_5}{se(\hat{\beta}_5)} = -0.000695/0.000254 = -2.74$$

(2 marks)

This is a right-tail test. The df is **321-6=315**. The p-value is

$$p - value = Pr(t_{315} \geq -2.74).$$

(2 marks)

The p-value is greater than 50%, so we fail to reject the null hypothesis at the 10% level.

(2 marks)

(11 marks)

(c) Compute a test of the proposition that the house size was the **only** factor that affects the house prices and, furthermore, if the house size increases by one square meter, then the predicted house prices increase by 350 dollars. State the coefficient restrictions on regression equation (1) implied by this proposition; that is, state the null hypothesis H_0 and the alternative hypothesis H_1. Write the restricted regression equation implied by the null hypothesis H_0. OLS estimation of this restricted regression equation yields a Residual Sum-of-Squares value of **SSR = 357,850,000,000**. Use this information, together with the results from OLS estimation of equation (1), to calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level.

Test: $H_0: \beta_1 = 350, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ versus H_1: H_0 is not true **(4 marks)**

Restricted regression implied by H_0:

$price_i - 350 hsize_i = \beta_0 + u_i$ **(1 marks)**

Compute F statistic: **(4 marks)**

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(357{,}850{,}000{,}000 - 271{,}740{,}000{,}000)/5}{271{,}740{,}000{,}000/(321-5-1)} = \mathbf{19.96}$$

Decision Rule: Compare the sample F statistic with the α-level critical value of the $\mathbf{F_{5,315}}$ distribution.
1. If $F \leq c_\alpha$, do not reject H_0 at the significance level α.
2. If $F > c_\alpha$, reject H_0 at the significance level α.

Critical value of $\mathbf{F_{5,315}}$ at 5% significance level ($\alpha = 0.05$) is $\boldsymbol{c_{0.05} = 2.21}$

Inference: **(2 marks)**
Since $\boldsymbol{F = 19.96 > 2.21 = c_{0.05}}$, **reject $\mathbf{H_0}$** at **5%** significance level.

(9 marks)
(d) Use the estimation results for regression equation (1) to perform a two-tail test of the null hypothesis $\beta_2 = -3\beta_1$ at the 5 percent significance level. The estimated covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ is $\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2) = 1418.9597$. State the null and alternative hypotheses, show how you calculate the required test statistic, and state its null distribution. State the decision rule you use, and the inference you would draw from the test at the 5 percent significance level.

Null hypothesis: $H_0: 3\beta_1 + \beta_2 = 0$
Alternative hypothesis: H_1: H_0 is not true **(3 marks)**

Compute t-statistic:
$3\hat{\beta}_1 + \hat{\beta}_2 = 3 * 361.1 - 915.7 = \mathbf{167.6}$

$$\hat{Var}(3\hat{\beta}_1 + \hat{\beta}_2) = 9\hat{Var}(\hat{\beta}_1) + \hat{Var}(\hat{\beta}_2) + 6cov(\hat{\beta}_1, \hat{\beta}_2)$$
$$\hat{Var}(3\hat{\beta}_1 + \hat{\beta}_2) = 9(28.54)^2 + (163.5)^2 + 6(1418.9597)$$
$$\hat{Var}(3\hat{\beta}_1 + \hat{\beta}_2) = \mathbf{42576.7926}$$
$$se(3\hat{\beta}_1 + \hat{\beta}_2) = \sqrt{\hat{Var}(3\hat{\beta}_1 + \hat{\beta}_2)} = \mathbf{206.3414}$$ **(2 marks)**
$$t = \frac{3\hat{\beta}_1 + \hat{\beta}_2}{se(3\hat{\beta}_1 + \hat{\beta}_2)} = \frac{167.6}{206.3414} = \mathbf{0.8122}$$

(2 marks)
Decision Rule: This is a ***two-tail* test**. Compare the sample t statistic with the α-level critical value of the t_{314} distribution.
1. If $|t| \leq c_\alpha$, do not reject H_0 at the significance level α.
2. If $|t| > c_\alpha$, reject H_0 at the significance level α.

Critical value of $\boldsymbol{t_{315}}$ at 5% significance level ($\alpha = 0.05$) is $\boldsymbol{c_{0.05} = 1.96}$

Inference: **(2 mark)**
Since $\boldsymbol{t = 0.8122 < 1.96 = c_{0.05}}$, **$\mathbf{H_0}$ cannot be rejected** at the **5%** significance level.

THE CHINESE UNIVERSITY OF HONG KONG, SHENZHEN
School of Management and Economics

ECONOMICS 3121
Introductory Econometrics

Spring term 2021

FINAL EXAM

DATE: **Thursday May 20, 2019**

TIME: **180 minutes; 4:00 p.m. – 7:00 p.m.**

INSTRUCTIONS: The exam consists of **FOUR (4)** questions. Students are required to answer **ALL FOUR (4)** questions.

Answer all questions in the exam booklets provided. Be sure your ***name*** and ***student number*** are printed clearly on the front of all exam booklets used.

Do not write answers to questions on the front page of the first exam booklet.

Please label clearly each of your answers in the exam booklets with the appropriate number and letter.

Please write legibly.

This exam is **CONFIDENTIAL**. This question paper must be submitted in its entirety with your answer booklet(s); otherwise your exam will not be marked.

A formula sheet and tables of critical values of the t distribution and F distribution are given on the last pages of the exam.

MARKING: The marks for each question are indicated in parentheses immediately above each question. **Total marks** for the exam **equal 170**.

GOOD LUCK!

Question 1: Multiple Choice (50 points, 2 points each)
Select the **BEST** response to each of the following questions.

1. Which of the following Gauss-Markov assumptions is violated by the linear probability model?
   a) The assumption of constant variance of the error term.
   b) The assumption of zero conditional mean of the error term.
   c) The assumption of no exact linear relationship among independent variables.
   d) The assumption that none of the independent variables are constants.

Answer: a

2. If $\widehat{\beta}_j$, an unbiased estimator of β_j, is consistent, then the:
   a) distribution of $\hat{\beta}_j$ becomes more and more loosely distributed around β_j as the sample size grows.
   b) distribution of $\hat{\beta}_j$becomes more and more tightly distributed around β_j as the sample size grows.
   c) distribution of $\hat{\beta}_j$ tends toward a standard normal distribution as the sample size grows.
   d) distribution of $\hat{\beta}_j$ remains unaffected as the sample size grows.

Answer: b

3. Which of the following assumptions is required to obtain a first-differenced estimator in a two-period panel data analysis?
   a) The independent variable does not change over time for any cross-sectional unit.
   b) The independent variable changes by the same amount in each time period.
   c) The variance of the error term in the regression model is not constant.
   d) The idiosyncratic error at each time period is uncorrelated with the independent variables in both time periods.

Answer: d

4. What will you conclude about a regression model if the Breusch-Pagan test results in a small p-value?
   a) The model contains homoskedasticty.
   b) The model contains heteroskedasticty.
   c) The model contains dummy variables.
   d) The model omits some important explanatory factors.

Answer: b

5. In a multiple linear regression model with heteroskedasticity, if you use Stata to obtain the heteroskedasticity-robust standard error, you will
   a) have estimates that are unbiased, but parameter standard errors that are too small as standard errors are biased under heteroskedasticity
   b) have parameter estimates that are unbiased, but standard errors with unknown bias
   c) have the best estimates you can come up with, because OLS is the Best Linear Unbiased Estimator (BLUE)
   d) have unbiased parameter estimates and unbiased standard errors, but OLS is not the

best estimator to use, as it is no longer the minimum variance estimator

Answer: d

6. In a first differenced equation, which of the following assumptions is needed for the usual standard errors to be valid when differencing with more than two time periods?
   a) The regression model exhibits heteroskedasticty.
   b) The differenced idiosyncratic error or Δu_{it} is uncorrelated over time.
   c) The unobserved factors affecting the dependent variable are time-constant.
   d) The regression model includes a lagged independent variable.

Answer: b

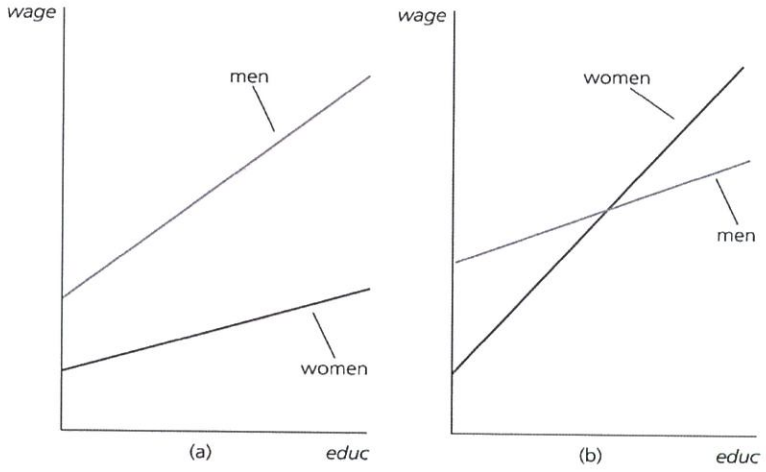
7. Which of the following types of variables cannot be included in a fixed effects model?
   a) Dummy variable
   b) Discrete dependent variable
   c) Time-varying independent variable
   d) Time-constant independent variable

Answer: d

8. A researcher has analyzed the differences in the effect of education on wages for men and women using the following regression model:

$$wage = \beta_0 + \beta_1 educ + \delta_0 female + \delta_1 female * educ + u$$

where $wage$ is log of hourly wages, $educ$ is the number of years of education, and $female$ is a dummy variable that is 1 for women. The following graphs show the estimated regression results from two different populations, (a) and (b).



What is the sign of δ_0 and δ_1 in the two populations?
   a) δ_0 is negative in (a) and negative in (b), δ_1 is positive in (a) and positive in (b)
   b) δ_0 is negative in (a) and negative in (b), δ_1 is negative in (a) but positive in (b)
   c) δ_0 is negative in (a) but positive in (b), δ_1 is negative in (a) and negative in (b)
   d) δ_0 is positive in (a) and positive in (b), δ_1 is negative in (a) but positive in (b)

Answer: b

9. Refer to the following finite distributed lag model.
$y_t = \alpha_0 + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + u_t$
$\beta_0 + \beta_1 + \beta_2 + \beta_3$ represents:
   a) the short-run change in y given a temporary increase in x.
   b) the short-run change in y given a permanent increase in x.
   c) the long-run change in y given a permanent increase in x.
   d) the long-run change in y given a temporary increase in x.

Answer: c

10. Which of the following is true of the OLS t statistics?
   a) The heteroskedasticity-robust t statistics are justified only if the sample size is large.
   b) The heteroskedasticty-robust t statistics are justified only if the sample size is small.
   c) The usual t statistics do not have exact t distributions if the sample size is large.
   d) In the presence of homoscedasticity, the usual t statistics do not have exact t distributions if the sample size is small.

Answer: a

11. In the following regression equation, y is a binary variable:
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$
In this case, the estimated slope coefficient, $\hat{\beta}_1$ measures _____.
   a) the predicted change in the value of y when x_1 increases by one unit, everything else remaining constant
   b) the predicted change in the value of y when x_1 increase by one percentage point, everything else remaining constant
   c) the predicted change in the probability of success when x_1 increase by one percentage point, everything else remaining constant
   d) the predicted change in the probability of success when x_1 increases by one unit, everything else remaining constant

Answer: d

12. Which of the following is a reason for using independently pooled cross sections?
   a) To obtain data on different cross sectional units
   b) To increase the sample size
   c) To select a sample based on the dependent variable
   d) To select a sample based on the independent variable

Answer: b

13. A Chow test _____.
   a) is used to test the presence of heteroskedasticty in a regression model.
   b) is used to determine how multiple regression differs across two groups.
   c) cannot detect changes in the slope coefficients of dependent variables over time.
   d) cannot be computed for more than two time periods.

Answer: b

14. Idiosyncratic error is the error that occurs due to _____.

a) incorrect measurement of an economic variable
b) unobserved factors that affect the dependent variable and change over time
c) unobserved factors that affect the dependent variable and do not change over time
d) correlation between the independent variables

Answer: b

15. Which of the following assumptions is required for obtaining unbiased fixed effect estimators?
a) The errors are heteroskedastic.
b) The errors are serially correlated.
c) The independent variables are strictly exogenous.
d) The unobserved effect is correlated with the independent variables.

Answer: c

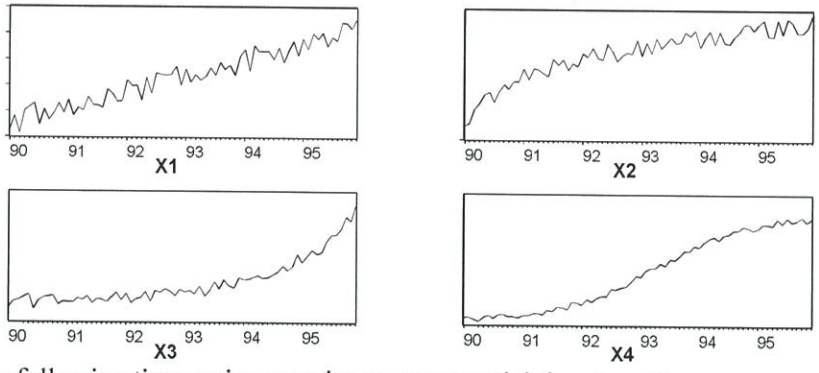
16. A pooled OLS estimator that is based on the time-demeaned variables is called the _____.
a) random effects estimator
b) fixed effects estimator
c) least absolute deviations estimator
d) instrumental variable estimator

Answer: b

17. What should be the degrees of freedom (df) for fixed effects estimation if the data set includes 'N' cross sectional units over 'T' time periods and the regression model has 'k' independent variables?
a) N-kT
b) NT-k
c) NT-N-k
d) N-T-k

Answer: c

18. The time series variables depicted below contain various types of time trend components.



which of the following time series contains an exponential time trend?
a) X1
b) X2
c) X3

d) X4

Answer: c

19. Which of the following is true of heteroskedasticity?
    a) Heteroskedasticty causes inconsistency in the Ordinary Least Squares estimators.
    b) Population R^2 is affected by the presence of heteroskedasticty.
    c) The Ordinary Least Square estimators are not the best linear unbiased estimators if heteroskedasticity is present.
    d) It is not possible to obtain F statistics that are robust to heteroskedasticity of an unknown form.

Answer: c

20. In an OLS multiple regression, if you divide the explained variable y by 10, which of the following will happen?
    a) The estimated intercept parameter will be reduced by 10 times
    b) the estimated slope parameter will be reduced by 10 units
    c) the Total Sum of Squares for the regression will be reduced by $\sqrt{10}$ times
    d) R-squared for the regression will be reduced by 10 times

Answer: a

21. Which of the following statements is true when the dependent variable, $y > 0$
    a) Taking log of a variable often expands its range.
    b) Models using $\log(y)$ as the dependent variable may satisfy CLM assumptions more closely than models using the level of y.
    c) Taking log of variables make OLS estimates more sensitive to extreme values.
    d) Taking logarithmic form of variables make the slope coefficients more responsive to rescaling.

Answer: b

22. Consider the following regression model: $y = \beta_0 + \beta_1 x + u$. Which of the following is a property of Ordinary Least Square (OLS) estimates of this model and their associated statistics?
    a) The sum, and therefore the sample average of the OLS residuals, is positive.
    b) The sum of the OLS residuals is negative.
    c) The sample covariance between the regressors and the OLS residuals is positive.
    d) The point $(\bar{x}, \bar{y})$ always lies on the OLS regression line.

Answer: d

23. First-differenced estimation in a panel data analysis is subject to serious biases if _____.
    a) key explanatory variables vary significantly over time
    b) the explanatory variables do not change by the same unit in each time period
    c) one or more of the explanatory variables are measured incorrectly
    d) the regression model exhibits homoscedasticity

Answer: c

24. Weighted least squares estimation is used only when _____.
    a) the dependent variable in a regression model is binary
    b) the independent variables in a regression model are correlated
    c) the error term in a regression model has a constant variance
    d) the form of the error variances is known

Answer: d

25. In a multiple linear regression model with heteroskedasticity, if you use Stata to obtain the heteroskedasticity-robust standard error, you will
    a) have estimates that are unbiased, but parameter standard errors that are too small as standard errors are biased under heteroskedasticity
    b) have parameter estimates that are unbiased, but standard errors with unknown bias
    c) have the best estimates you can come up with, because OLS is the Best Linear Unbiased Estimator (BLUE)
    d) have unbiased parameter estimates and unbiased standard errors, but OLS is not the best estimator to use, as it is no longer the minimum variance estimator

Answer: a

Question 2 (20 points)
Answer parts (a), (b), and (c) below.

(6 points)
(a) Briefly explain what strict exogeneity is in time series study.

$E(u_t | X) = 0$ (3)

The mean value of the error is uncorrelated to the values of the independent variables in all periods. (3)

(6 points)
(b) Suppose the unobserved effects panel data model is as follows:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \beta_3 x_{it3} + a_i + u_{it}, \quad t = 1, 2, \ldots, T$$

where a_i is the individual unobserved effect, u_{it} is idiosyncratic error.

Write down the first differenced equation and the time-demeaned equation.

FD: $\Delta y_{it} = \beta_1 \Delta x_{it1} + \beta_2 \Delta x_{it2} + \beta_3 \Delta x_{it3} + \Delta u_{it}, \quad t = 2, \ldots, T$ (3)

FE: $y_{it} - \bar{y}_i = \beta_1 (x_{it1} - \bar{x}_{i1}) + \beta_2 (x_{it2} - \bar{x}_{i2}) + \beta_3 (x_{it3} - \bar{x}_{i3}) + (u_{it} - \bar{u}_i)$ (3)
$, \quad t = 1, \ldots, T$

1变量. ≥2个数.

(8 points)
(c) Briefly explain how to estimate the Linear Probability model by Weighted Leaset Squares?

(1). Estimate the model by OLS and obtain the predicted values $\hat{y}_i$ (2)

(2) Determine whether all the predicted values are inside the unit interval. If not, some adjustment is needed to bring all predicted value in the unit interval (2)

(3) Calculate $\hat{h}_i$ (2)

(4) Find WLS estimation (2) run regression.

$$\left[\frac{y_i}{\sqrt{h_i}}\right] = \beta_0 \left[\frac{1}{\sqrt{h_i}}\right] + \beta_1 \left[\frac{x_{i1}}{\sqrt{h_i}}\right] + \cdots + \beta_k \left[\frac{x_{ik}}{\sqrt{h_i}}\right] + \left[\frac{u_i}{\sqrt{h_i}}\right]$$

Question 3 (46 marks)

You are investigating the relationship between the birth weights of newborn babies and four of their determinants: mother's average daily cigarette consumption during pregnancy; the number of prenatal visits made by the mother to a physician or medical facility during pregnancy; the mother's age; and the mother's race. You have sample data for 1656 babies born during a given year on the following variables

$bwght_i$ = birth weight of the baby born to the i-th mother, measured in *hundreds of grams*;
$cigs_i$ = average number of cigarettes per day smoked by the i-th mother during pregnancy, measured in *cigarettes per day*;
$npvis_i$ = number of prenatal visits to a doctor or medical facility made by the i-th mother during pregnancy;
age_i = age of the i-th mother, measured in *years*;
$white_i$ = an indicator variable defined such that $white_i = 1$ if the i-th mother is white, and $white_i = 0$ the i-th mother is non-white;

Using the given sample data on 1656 newborn babies, your trusty research assistant has estimated regression equation (1) and obtained the following estimation results (with estimated standard errors given in parentheses below the coefficient estimates):

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 npvis_i + \beta_3 npvis_i^2 + \beta_4 age_i + \beta_5 age_i^2 + \beta_6 white_i + \beta_7 white_i age_i + u_i \quad (1)$$

$\hat{\beta}_0 = 16.076$ (4.618) $\quad \hat{\beta}_1 = -0.1010$ (0.03338) $\quad \hat{\beta}_2 = 0.2971$ (0.1061) $\quad \hat{\beta}_3 = -0.006045$ (0.003429)

$\hat{\beta}_4 = 0.8083$ (0.2825) $\quad \hat{\beta}_5 = -0.0101$ (0.004534) $\quad \hat{\beta}_6 = 6.8850$ (2.587) $\quad \hat{\beta}_7 = -0.2039$ (0.08734)

$SSR = 53099.886 \quad SST = 54516.777 \quad N = 1656$

(8 marks)

(a) Use the estimation results for regression equation (1) to test the joint significance of the slope coefficient estimates $\hat{\beta}_j$ $(j = 1, \ldots, 7)$ in regression equation (1). Perform the test at the 1 percent significance level (i.e., for significance level $\alpha = 0.05$). State the null and alternative hypotheses, and show how you calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level.

H_0: $\beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0, \beta_6 = 0, \beta_7 = 0$, (2)

H_1: H_0 is not true

$$R^2_{ur} = 1 - \frac{SSR}{SST} = 1 - \frac{53099.886}{54516.777} = 0.026$$

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)} = \frac{0.026/7}{(1-0.026)/1648} = 6.2845 \quad (4)$$

$C_{0.05} = 2.01$ (1)

(or $F = \frac{(54516.777 - 53099.886)/7}{53099.886/1648} = 6.2821$)

Since $F = 6.2845 > 2.01 = C_{0.05}$

Reject the H_0 at 5% significance level (1)

(4 marks)

(b) Interpret the slope coefficient estimate $\hat{\beta}_1$ in sample regression equation (1). That is, explain in words what the numerical value of the slope coefficient estimate $\hat{\beta}_1$ means.

the estimate $\hat{\beta}_1 = -0.101$ means that 1 cigarettes per day increase in mother's smoking is associated with an decrease (1) in average (1) birth weight of her body equal to 0.101 hundreds of grams (1) (1)

(10 marks)

(c) Use the estimation results for regression equation (1) to compute the two-sided 95 percent confidence interval for the slope coefficient β_1. Use the computed confidence interval for β_1 to perform a test of the proposition that mothers' cigarette consumption during pregnancy ($cigs_i$) has no effect on the birth weight of their babies. Perform the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$). State the null hypothesis H_0 and the alternative hypothesis H_1. State the inference you would draw from the test at the 5 percent significance level.

two sided 95 percent CI of β_1 is:

$\hat{\beta}_1 - c_{0.05} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + c_{0.05} se(\hat{\beta}_1)$ (2)

$c_{0.05} = 1.96$

Lower bound: $-0.101 - 1.96 \cdot 0.03338 = -0.1664$ (2)

upper bound: $-0.101 + 1.96 \cdot 0.03338 = -0.03558$ (2)

Two sided 95% CI is $[-0.1664, -0.03558]$

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$ (1)

Since $0 \notin [-0.1664, -0.03558]$ (1)

reject H_0 at 5% significance level (1)

(8 marks)

(d) Compute a test of the proposition that the number of prenatal visits by the mother to a doctor or medical facility has no effect on the birth weight of newborn babies. Perform the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$). State the coefficient restrictions on regression equation (1) implied by this proposition; that is, state the null hypothesis H_0 and the alternative hypothesis H_1. Write the restricted regression equation implied by the null hypothesis H_0. OLS estimation of this restricted regression equation yields a Residual Sum-of-Squares value of **SSR = 53542.009**. Use this information, together with the results from OLS estimation of equation (1), to calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level.

$\frac{\partial\, bwght_i}{\partial\, npvis_i} = \beta_2 + 2\beta_3 npvis_i$

$H_0: \beta_2 = 0, \beta_3 = 0$ (2)

H_1: H_0 is not true

restricted model:

$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_4 age_i + \beta_5 age_i^2 + \beta_6 white_i + \beta_7 white_i age_i + u_i$ (1)

$F = \frac{(53542.009 - 53091.886)/2}{53091.886/(1656-7-1)}$ (2)

$= 6.8608$

$c_{0.05} = 3$ (1) Since $F = 6.8608 > 3 = c_{0.05}$

reject H_0 at 5% significance level (1)

(8 marks)

(e) Use the estimation results for regression equation (1) to test the proposition that the marginal effect of mother's age (age_i) on the birth weight of newborn babies is negatively related to age_i. Perform the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$). State the null hypothesis H_0 and the alternative hypothesis H_1 implied by this proposition. Show how you calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level.

$\frac{\partial\, bwght_i}{\partial\, age_i} = \beta_4 + 2\beta_5 age_i + \beta_7 white_i$

$H_0: \beta_5 = 0$ (or $\beta_5 \geq 0$)

$H_1: \beta_5 < 0$ left-tail test (2)

$t = \frac{-0.0101}{0.004534} = -2.23$ (2)

$C_{0.05} = 1.645$ (1)

Since $t = -2.23 < -1.645 = C_{0.05}$ (1)

reject H_0 at the 5% significance level (1)

(8 marks)

(f) Compute a test of the proposition that the marginal effect of mother's age (age_i) on birth weight is zero for babies born to white mothers. State the coefficient restrictions on regression equation (1) implied by this proposition; that is, state the null hypothesis H_0 and the alternative hypothesis H_1. Write the restricted regression equation implied by the null hypothesis H_0. OLS estimation of this restricted regression equation yields a R-squared value of $\boldsymbol{R^2}$**=0.0193**. Use this information, together with the results from OLS estimation of equation (1), to calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$)

$\frac{\partial E(bwght \mid white_i = 1)}{\partial age_i} = \beta_4 + 2\beta_5 age_i + \beta_7$

$= (\beta_4 + \beta_7) + 2\beta_5 age_i$

$H_0: \beta_4 + \beta_7 = 0,\ \beta_5 = 0$

$H_1:$ H_0 is not true (2)

Restricted model

$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 npvis_i + \beta_3 npvis_i^2$

$+ \beta_4 (1 - white_i)\, age_i + \beta_6 white_i + u_i$ (2)

$F = \frac{(0.026 - 0.0193)/2}{(1-0.026)/(1646-7-1)}$ (2)

$= 5.6682$

$C_{0.05} = 3$

Since $F = 5.6682 > 3 = C_{0.05}$

reject H_0 at 5% significance level (1)

Question 4 (54 marks)

You are conducting an econometric investigation into the selling prices of houses in a single urban area in the years 2009 and 2015. You have available for this purpose a random sample of houses that were sold in the years 2009 and 2015. The sample data consist of observations for these houses on the following variables:

$price_i$ = the selling price of house i, in *thousands of dollars*;
$hsize_i$ = the living area of house i, in *hundreds of square feet*;
lot_i = lot size of house i, in *hundreds of square feet*;
age_i = age of house i, in years;
$d15_i$ = an indicator variable defined to equal 1 if house i was sold in 2015, and 0 if house i was sold in 2009
$mall_i$ = an indicator variable defined to equal 1 if house i is near the shopping mall, and 0 otherwise

The regression model you are asked to estimate is given by the population regression equation

$$\begin{aligned} price_i = {} & \beta_0 + \beta_1 hsize_i + \beta_2 hsize_i^2 + \beta_3 lot_i + \beta_4 age_i + \beta_5 hsize_i age_i + \beta_6 mall_i \\ & + \beta_7 d15_i + \beta_8 d15_i hsize_i + \beta_9 d15_i hsize_i^2 + \beta_{10} d15_i lot_i \\ & + \beta_{11} d15_i age_i + \beta_{12} d15_i hsize_i age_i + \beta_{13} d15_i mall_i + u_i \end{aligned} \quad (1)$$

where the β_j $(j = 0, 1, 2, \ldots, 13)$ are regression coefficients and u_i is an error term

State the null hypothesis H_0 and alternative hypothesis H_1 of the statistical test that you would perform on regression equation (1) to assess the evidence for each of the following empirical propositions. In addition, for each hypothesis test, state which of the following tests you would use: (1) a two-tail t-test; (2) a left-tail t-test; (3) a right-tail t-test; or (4) an F-test.

(6 marks)

(a) The marginal effect of house size on price was zero in 2009.

$\frac{\partial E(price_i | d15=0)}{\partial hsize} = \beta_1 + 2\beta_2 hsize_i + \beta_5 age_i$ ④ H_0

$H_0: \beta_1=0, \beta_2=0, \beta_5=0$, H_1: H_0 is not true, F-test ②

(6 marks)

(b) The marginal effect of house size on price increases as house size increases in 2015.

$\frac{\partial E(price_i | d15=1)}{\partial hsize} = \beta_1 + 2\beta_2 hsize_i + \beta_5 age_i + \beta_8 + 2\beta_9 hsize_i + \beta_{13} age_i$ ②

$H_0: \beta_2 + \beta_9 = 0$ ($\beta_2 + \beta_9 \leq 0$), $H_1: \beta_2 + \beta_9 > 0$, ③ right-tail t-test ②

(6 marks)

(c) The marginal effect of house age on price was the same in 2009 as it was in 2015.

$\frac{\partial price_i}{\partial age_i} = \beta_4 + \beta_5 hsize_i + \beta_{11} d15_i + \beta_{12} d15_i hsize_i$

$H_0: \beta_{11}=0, \beta_{12}=0$ ④, H_1: H_0 is not true, F-test ②

(6 marks)

(d) House size and house age were substitutable for one another in determining house prices in 2015 (i.e., the effect of house age on price decreases as house size increases in 2015).

$H_0: \beta_5 + \beta_{12} = 0$ (or $\beta_5 + \beta_{12} \geq 0$) ② $H_1: \beta_5 + \beta_{15} < 0$, ② left-tail t-test. ②

(6 marks)

(e) The marginal effect of house size on price was constant in 2015.

$H_0: \beta_5 + \beta_{12} = 0$, $\beta_2 + \beta_9 = 0$ ④ H_1: H_0 is not true, F-test ②

H_0 才 · H_1 × 扣一半

(6 marks)

(f) The marginal effect of lot size on price was zero in both 2009 and 2015.

$\frac{\partial price_i}{\partial lot_i} = \beta_3 + \beta_{10} d15$

$H_0: \beta_3=0, \beta_{10}=0$ H_1: H_0 is not true ④ H_0, F-test ②

(6 marks)

(g) The slope coefficients do not change over time.

$H_0: \beta_8=0, \beta_9=0, \beta_{10}=0, \beta_{11}=0, \beta_{12}=0, \beta_{13}=0$ H_1: H_0 is not true ④ H_0

F-test ②

(6 marks)

(h) For houses with 2500 square feet of living area and lots of 6000 square feet that were 10 years old and near the mall, mean price in 2015 was **greater than** mean price in 2009.

$E(price_i | d15_i=1) - E(price_i | d15_i=0) = \beta_7 + \beta_8 hsize_i + \beta_9 hsize_i^2 + \beta_{10} lot_i + \beta_{11} age_i + \beta_{12} hsize_i age_i + \beta_{13} mall_i$

at $hsize_i=25$; $lot_i=60$; $age_i=10$; $mall_i=1$

$H_0: \beta_7 + \beta_8 25 + \beta_9 (25)^2 + \beta_{10} 60 + \beta_{11} 10 + \beta_{12}(25)(10) + \beta_{13} = 0$ (or ≤ 0) ②

~~H_1: H_0 is not true~~ ~~F-test~~. see next page

$H_1: \beta_7 + \beta_8 25 + \beta_9 (25)^2 + \beta_{10} 60 + \beta_{11} 10 + \beta_{12} (25)(10) + \beta_{13} > 0$ (4)

right tail t-test (2)

(6 marks)

(i) Mall location did not affect the housing price. Suppose the mall was built in 2012.

$H_0: \beta_{13} = 0$ (2)

$H_1: \beta_{13} \neq 0$ (2). (4)

a two-tailed t-test. (2)