Name: ____

Introduction to Econometrics - Midterm exam
Suggested Solution

by Hieu Nguyen

NOTES:

- There are 11 pages for this exam, make sure you check all 11 pages when doing;
- Calculation questions should be explained carefully rather than a number only to get the full grades;
- Make sure your handwriting is readable. Otherwise, you cannot be graded;

GOOD LUCK!

Multiple Choice Questions (30 pts = 10 * 3 pts)

1. What is the meaning of the term ”heteroscedasticity”?

   a. **The variance of the errors is not constant**

   b. The variance of the dependent variable is not constant

   c. The errors are not linearly independent of one another

   d. The errors have non-zero mean

2. Data on one or variables collected at a given point of time

   a. Time series data

   b. **Cross-section data**

   c. Pooled data

   d. Panel data

3. The coefficient of determination (R^2) shows how many %...

   a. **Variation in the dependent variable Y is explained by the variation in independent variable X**

   b. Variation in the independent variable Y is explained by the variation in dependent variable X

   c. Variation in the dependent variable Y explains the variation in independent variable X

   d. Variation in the independent variable Y explains the variation in dependent variable X

4. Rejecting a true hypothesis results in which type of error

   a. **Type I error**

   b. Type II error

   c. Structural error

   d. Hypothesis error

5. Which of the following statements is true of hypothesis testing?

   a. The t test can be used to test 3 coefficient restrictions.

   b. A test of single restriction is also referred to as a joint hypotheses test.

   c. **A restricted model will always have fewer parameters than its unrestricted model.**

   d. OLS estimates maximize the sum of squared residuals.

6. Which of the following statements is true?

   a. If the calculated value of F-test is higher than the F-critical value, we reject the alternative hypothesis and accept the null hypothesis

   b. **The value of F-test is always nonnegative because SSR_r is never smaller than SSR_{ur}**

   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

7. The hypothesis testing with $H_1 : \beta_j \neq 0$, where β_j is a regression coefficient associated with an explanatory variable, represents a one-sided alternative hypothesis.

   a. **true**

   b. false

8. In the following equation, *GDP* refers to gross domestic product (in million USD), *bankcredit* refers to the amount of loans a bank provides to its customers (in million USD), and *FDI* refers to foreign direct investment.

$$log(GDP) = 2.65 + 0.527\, log(bankcredit) + 0.222\, FDI$$

   Which of the following statements is then true?

   a. If GDP increases by 1%, bank credit will increase by 0.527%, given the level of FDI remaining constant.

   b. **If bank credit increases by 1%, GDP will increase by 0.527%, given the level of FDI remaining constant.**

   c. If GDP increases by 1%, bank credit increases by 0.527 million USD, given the level of FDI remaining constant.

   d. If bank credit increases by 1%, GDP will increase by 0.527 million USD, given the level of FDI remaining constant

9. Which of the following correctly identifies an advantage of using adjusted R^2 over R^2?

   a. Adjusted R^2 corrects the bias in R^2

   b. Adjusted R^2 is easier to calculate than R^2

   c. **Adjusted R^2 has the penalty of adding new (irrelevant) independent variable(s) while R^2 doesn't have any**

   d. The adjusted R^2 can be calculated for models having logarithmic functions while R^2 cannot be calculated for such models

10. The term u in an econometric model below is usually referred to as the

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

   a. **error term**

   b. parameter

   c. hypothesis

   d. dependent variable

Theoretical Question (20 pts):

Choose **ONE** out of 2 following questions to answer:

1. Discuss in detail 4 importance specification criteria to decide whether a variable belongs to the regression equation.
2. Discuss in detail 2 types of error in doing hypothesis testing. Provide 1 example for each type.

Suggested solution:

1. When determining whether a variable should be included in a regression equation, there are four important specification criteria to consider:

   - *Theory:* A variable's inclusion should be supported by theoretical justification. This means that the variable's role in the equation should be unambiguous and theoretically sound. If economic theory or domain-specific knowledge strongly suggests that a variable influences the dependent variable, it should be included. For instance, when modeling consumer demand, variables like price and income are theoretically justified as determinants.
   - *t-test:* The significance of a variable's estimated coefficient can be tested using a t-test. The t-test assesses whether the coefficient is statistically different from zero, indicating that the variable has a meaningful impact on the dependent variable. Moreover, it should be significant in the expected direction (positive or negative) as suggested by theory. For example, in a demand model, a positive coefficient on income is expected, as higher income should increase demand.
   - *R^2:* Another criterion is whether the overall fit of the model improves with the inclusion of the variable. This can be evaluated using the R^2 statistic, which measures the proportion of variance explained by the model. If adding a variable significantly increases R^2, it suggests that the variable contributes valuable information to the model. However, the adjusted R^2 should also be considered to account for the addition of multiple variables.
   - *Bias:* The inclusion of a variable can affect the coefficients of other variables in the model. If adding a variable causes significant changes in other coefficients, this suggests that the variable controls for an important factor and reduces omitted variable bias. For instance, in a model predicting wages, including education as a variable might change the coefficients of other demographic factors, indicating its relevance in explaining wage variation.

2. 2 types of error in doing hypothesis testing:

   - Type I Error (False Positive) occurs when we reject the null hypothesis when it is actually true. This is also known as a "false positive" result. The probability of committing a Type I error is denoted by α, which is the significance level of the test (commonly set at 5% or 1%). Example: Suppose a pharmaceutical company tests a new drug to determine if it is more effective than the current standard treatment. The null hypothesis (H_0) is that the new drug has no effect (i.e., it is no more effective than the standard treatment). A Type I error would occur if the test results lead the company to conclude that the new drug is effective when, in reality, it is not. This could result in the company promoting an ineffective drug, potentially causing harm to patients and wasting resources.
   - Type II Error (False Negative) occurs when we fail to reject the null hypothesis when it is actually false. This is also known as a "false negative" result. The probability of committing a Type II error is denoted by β. Example: Consider a criminal trial where the null hypothesis (H_0) is that the defendant is innocent. A Type II error would occur if the jury fails to reject the null hypothesis (and thus finds the defendant not guilty) when the defendant is actually guilty. This would allow a guilty person to go free, which could have serious consequences for society and undermine trust in the justice system.

Practice Exercises (50 pts):

1. From 40 randomly selected large US cities in 1988, the researchers aim to identify what factors can affect the demand for urban transportation by bus. Potential explanatory variables include:

   - *bustravel* ... demand for urban transportation by bus (in thousands of person);
   - *fare* ... bus fare (in USD,);
   - *gasprice* ... price of a gallon of gasoline (in USD);
   - *income* ... average annual income per capita (in USD);
   - *pop* ...population of the city (in thousands);
   - *density* ... population density (persons/sq. mile);
   - *landarea* ... land area of the city (sq. miles).

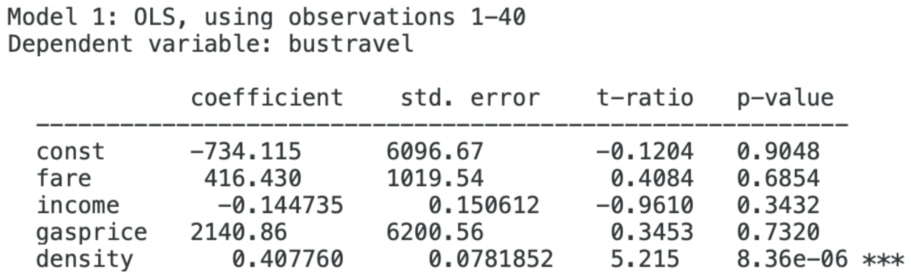
   Answer the following questions:

   a. (2 pts) Construct a regression model in which demand for urban transportation by bus (in thousands of person) is estimated based on bus fare (in USD), average annual income per capita (in USD), price of a gallon of gasoline (in USD), and population density (persons/sq. mile). Name this model as model 1.

   Suggested solution:

   $$\texttt{bustravel} = \beta_0 + \beta_1 \cdot \texttt{fare} + \beta_2 \cdot \texttt{income} + \beta_3 \cdot \texttt{gasprice} + \beta_4 \cdot \texttt{density} + \varepsilon$$

   b. (2 pts) Based on the regression output of model 1 shown below, write down the regression equation. *You are required to write numeric intercept, coefficients, and standard errors in parentheses under the corresponding coefficients*



```
Model 1: OLS, using observations 1-40
Dependent variable: bustravel

             coefficient    std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      -734.115      6096.67       -0.1204   0.9048
  fare        416.430      1019.54        0.4084   0.6854
  income       -0.144735      0.150612   -0.9610   0.3432
  gasprice   2140.86       6200.56        0.3453   0.7320
  density       0.407760      0.0781852   5.215    8.36e-06 ***
```

   Suggested solution:

   $$\widehat{\texttt{bustravel}} = -734.115 + 416.430 \cdot \texttt{fare} - 0.144735 \cdot \texttt{income} + 2140.86 \cdot \texttt{gasprice} + 0.407760 \cdot \texttt{density}$$

   $$(6096.67) \quad (1019.54) \quad (0.150612) \quad (6200.56) \quad (0.0781852)$$

   c. (4 pts) From (b), interpret the estimated coefficients of *fare* and *gasprice*

   Suggested solution:

   - The coefficient of *fare* is estimated to be 416.43, which suggests that for each 1 USD increase in bus fare, the demand for urban transportation by bus (*bustravel*) increases by 416.43 thousand persons, holding other variables constant. This result is counterintuitive, as we would typically expect higher fares to decrease demand (signal for the problem in the regression model!)
   - The coefficient of *gasprice* is 2140.86, indicating that for each 1 USD increase in the price of gasoline, the demand for bus travel increases by 2140.86 thousand persons. This positive relationship aligns with the expectation that higher gasoline prices may encourage more people to use bus transportation.

d. (4 pts) From (b), construct 90% confidence interval for population coefficients of *fare* and 95% confidence interval for *gasprice*

Suggested solution:

- The 90% confidence interval for the coefficient of *fare* can be calculated as:

$$\mathrm{CI}_{90\%} = 416.43 \pm (t_{0.1,35}) \times 1019.54$$

Using a t-table, we find that $t_{0.1,35} \approx 1.6895 \approx 1.690$. Thus:

$$\mathrm{CI}_{90\%} = 416.43 \pm (1.690 \times 1019.54) = (-1306.59, 2139.45)$$

- The 95% confidence interval for the coefficient of *gasprice* is:

$$\mathrm{CI}_{95\%} = 2140.86 \pm (t_{0.05,35}) \times 6200.56$$

Using a t-table, we find that $t_{0.05,35} \approx 2.03$. Thus:

$$\mathrm{CI}_{95\%} = 2140.86 \pm (2.03 \times 6200.56) = (-10446.28, 14728.00)$$

e. (8 pts) From (b), test the null hypothesis at 10% level of significance that *fare* has no effect on *bustravel* against the alternative that it has a negative effect. *Note: You need to state null hypothesis, alternative hypothesis, clearly calculate test value, critical value, decision of rejection or acceptance, and interpretation of the decision.*

Suggested solution:

- (H_0): $\beta_{\text{fare}} = 0$ and (H_1): $\beta_{\text{fare}} < 0$
- t-test statistic: $\frac{\text{Coefficient of fare}}{\text{Standard error of fare}} = \frac{416.43}{1019.54} = 0.4084$
- t-critical value at 10% significance level with 35 degrees of freedom (1-tail test): $t_{0.10,35} \approx 1.306$

Since the absolute value of the calculated t-value (0.4084) is greater than the t-critical value (1.306), we fail to reject the null hypothesis. This suggests that there is no significant evidence at the 10% significance level to conclude that an increase in *fare* has a negative effect on *bustravel.*

f. (8 pts) From (b), I want to test the null hypothesis at 10% level of significance that the total effect of *fare* and *gasprice* is 0 against the alternative that the total effect is different from 0. Specify the way to proceed the t-test for this hypothesis testing. *Note: You need to state null hypothesis, alternative hypothesis, clearly calculate test value, critical value, decision of rejection or acceptance, and interpretation of the decision.*

Suggested solution:

- (H_0): $\beta_{\text{fare}} + \beta_{\text{gasprice}} = 0$ and (H_1): $\beta_{\text{fare}} + \beta_{\text{gasprice}} \neq 0$
- Test statistic:

$$t = \frac{(\beta_{\text{fare}} + \beta_{\text{gasprice}}) - 0}{SE(\beta_{\text{fare}} + \beta_{\text{gasprice}})}$$

To find t, we need to know the $SE(\beta_{\text{fare}} + \beta_{\text{gasprice}})$

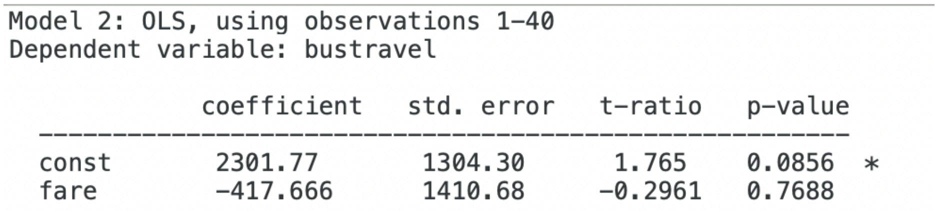
- t-critical value at 10% significance level (two-tailed) with 35 degrees of freedom: $t_{0.1,35} \approx 1.690$

If the absolute value of calculated t-value is greater than 1.690, we reject the null hypothesis. Otherwise, we fail to reject it. This would indicate whether the combined effect of *fare* and *gasprice* is statistically different from zero at the 10% significance level.

g. (8 pts) A person claims that *income, gasprice, density* each has no effect on *bustravel.* Conduct the hypothesis testing for this claim at 5% level of significance.

Note: You may need information of model 2 as below:

Also, given that



```
Model 2: OLS, using observations 1-40
Dependent variable: bustravel

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const       2301.77       1304.30      1.765    0.0856  *
  fare        -417.666      1410.68     -0.2961   0.7688
```

- Model 1 has $R^2 = 0.5375$, $SSR = 1933.175$
- Model 2 has $R^2 = 0.2302$, $SSR = 2431.757$

Suggested solution:

- Setting the hypotheses:

$$H_0\text{: } \beta_{\text{income}} = \beta_{\text{gasprice}} = \beta_{\text{density}} = 0$$
$$H_1\text{: At least one of } \beta_{\text{income}}, \beta_{\text{gasprice}}, \text{ or } \beta_{\text{density}} \text{ is not equal to 0.}$$

- Choosing the Test Statistic: Since we are testing multiple coefficients, we use the F-test for joint significance. The F-test statistic is calculated as:

$$F = \frac{(SSR_{\text{restricted}} - SSR_{\text{unrestricted}})/q}{SSR_{\text{unrestricted}}/(n-k-1)}$$

  where:
  - $SSR_{\text{restricted}}$ is the sum of squared residuals for the restricted model (Model 2, where *income*, *gasprice*, and *density* are excluded),
  - $SSR_{\text{unrestricted}}$ is the sum of squared residuals for the unrestricted model (Model 1, which includes all variables),
  - q is the number of restrictions (3 in this case, for *income*, *gasprice*, and *density*),
  - n is the number of observations (40),
  - k is the number of parameters in the unrestricted model (4 in this case, for *fare*, *income*, *gasprice*, and *density*).
- Calculating the F-Statistic
  Given:

$$SSR_{\text{restricted}} = 2431.757, \quad SSR_{\text{unrestricted}} = 1933.175$$
$$q = 3, \quad n = 40, \quad k = 4$$

$$F = \frac{(2431.757 - 1933.175)/3}{1933.175/(40-4-1)} = \frac{498.582/3}{1933.175/35} = \frac{166.194}{55.2336} \approx 3.01$$

- Finding the Critical Value and Making a Decision
  For an F-test with $q = 3$ and $n - k - 1 = 35$ degrees of freedom at the 5% significance level, we look up the critical value in an F-distribution table. The critical value $F_{0.05,3,35} \approx 2.87$.
- Conclusion: Since the calculated F-statistic (3.01) is greater than the critical value (2.87), we reject the null hypothesis at the 5% significance level. This suggests that at least one of the variables *income*, *gasprice*, or *density* has a statistically significant effect on *bustravel*.

h. (8 pts) Test the overall significance of a regression in model 1, given $R^2 = 0.5375$, $SSR = 1933.175$. *Note: You need to state null hypothesis, alternative hypothesis, clearly calculate test value, critical value, decision of rejection or acceptance, and interpretation of the decision.*

Suggested solution:

- Setting the hypotheses:

$$(H_0)\text{: } \beta_{\text{fare}} = \beta_{\text{income}} = \beta_{\text{gasprice}} = \beta_{\text{density}} = 0 \text{ (the model has no explanatory power)}$$
$$(H_1)\text{: At least one of the slope coefficients is not zero (the model has explanatory power)}$$

- Test Statistic:

$$F = \frac{R^2/q}{(1-R^2)/(n-k-1)}$$

  where:

  - $R^2 = 0.5375$
  - $q = 4$
  - $k = 4$ (number of parameters in the model)
  - $n = 40$ (number of observations)

$$F = \frac{0.5375/4}{(1-0.5375)/(40-4-1)} = \frac{0.5375/4}{0.4625/35} = \frac{0.134375}{0.0132143} \approx 10.17$$

- Critical Value:
  For an F-test with degree of freedom 1 (df1) $= q = 4$ and degree of freedom 2 (df2) $= n-k-1 = 35$ degrees of freedom at the 5% significance level, the critical value $F_{0.05,4,35} \approx 2.61$.
- Compare values and make decision: F-statistic $(10.17) > F - cv(2.61)$, we reject the null hypothesis.
- Interpretation: The decision to reject the null hypothesis suggests that the regression model has significant explanatory power at the 5% level. This means that at least one of the independent variables (*fare*, *income*, *gasprice*, or *density*) contributes to explaining the variability in *bustravel*.

i. (6 pts) One student constructs a regression model (named model 3) in which demand for urban transportation by bus (in thousands of person) is estimated based on bus fare (in USD), average annual income per capita (in USD), price of a gallon of gasoline (in USD), population density (persons/sq. mile), population of the city (in thousands), and land area of the city (sq. miles). You are asked to answer:

- Discuss in detail what classical assumption is violated in model 3.
- What is the consequence of this violation?
- How to solve this violation problem?

Suggested solution:

- In Model 3, the classical assumption that is likely violated is the assumption of *no multicollinearity* among the explanatory variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other. Given that the model includes variables such as population density, population, and land area, which are often related to each other in urban settings, multicollinearity is a probable issue.
- The presence of multicollinearity affects the precision of the estimated coefficients. Specifically, it leads to:
  - *High standard errors* for the affected coefficients, which reduces the statistical significance of those variables.
  - *Unstable coefficient estimates* that can vary greatly with small changes in the model or data, making it difficult to interpret the true effect of each variable.
  - *Reduced reliability* of the regression results, as high multicollinearity inflates the variance of the estimated coefficients, making them less precise.
- There are several methods to address multicollinearity: (*Notes: only one of these following solutions is enough to get full credit.*
  - *Remove one or more correlated variables:* If two variables are highly correlated (e.g., population and land area), consider removing one of them from the model to reduce multicollinearity.
  - *Combine correlated variables:* Create a new variable that combines the information from the correlated variables. For instance, instead of using both population and population density, one might use only one of them or an index that reflects urban density.

Above is suggested solution, detailed grades for students will be based on students' understanding and explanation (with convinced arguments)