Imperial College Business School

MSc Business Analytics

STATISTICS AND ECONOMETRICS

Mock Exam - Solutions

Instructions

Answer all **FOUR** questions.

You are supplied with a formula sheet.

College approved calculators can be used.

Question 1 (20 Marks)

Answer the following questions. Be concise and to the point.

(a) What does R^2 measure? Is a regression useless if its R^2 is low? Explain.

Answer: R^2 measures how well the OLS regression line fits the data in the sample. More specifically, it measures what portion of the variation in y is explained by the variation in the independent variables $x_1, x_2, \ldots, x_k$.

Not necessarily, however, the low R^2 implies that the included regressors do not explain much of the variation in y. That is, there are important omitted factors.

(b) Suppose the true return to education model is $wage = \beta_0 + \beta_1 exper + \beta_2 exper^2 + u$. What is the consequence, if we estimate the model without the quadratic term $exper^2$?

Answer: This is a case of functional form misspecification, which causes the OLS estimators to be biased. It is also a case of omitted variable.

(c) Explain in words the difference between a population regression function and an OLS regression line?

Answer: The population regression function is the true model of the relationship between the dependent and the independent variables. The OLS regression line is the model with estimated parameters, where the estimation uses a sample of data from the population and minimizes the sum of squared residuals.

(d) What is unbiasedness and why is it a desirable property of the OLS estimators?

Answer: $E(\widehat{\beta_j}) = \beta_j$. In words, the expected value of the OLS estimator equals the true parameter value. This means that with repeated samples, the average value of the OLS estimate will approximately equal the true parameter value.

Question 2 (30 Marks)

Suppose you have tested a model of rent rates and student population in a college town

$$\widehat{\log(rent)} = \underset{(.844)}{1.39} + \underset{(.039)}{.066} \log(pop) + \underset{(.081)}{.507} \log(avginc) + \underset{(.0017)}{.0056}\, pctstu + u$$
$$n = 264, \quad R^2 = .458$$

where $rent$ is the average monthly rent paid on rental units in a college town, pop denotes the total city population, $avginc$ denotes the average city income, and $pctstu$ denotes the student population as a percentage of the total population.

(a) Suppose you want to test $H_0: \beta_{\log(avginc)} = 0.5$ against $H_1: \beta_{\log(avginc)} \neq 0.5$. Construct the 95% confidence interval for the parameter $\beta_{\log(avginc)}$. (The 5% critical value for a two-tailed test is 1.96)

Answer: The 95% confidence interval for $\beta_{\log(avginc)}$ is: $.507 \pm 1.96(.081) = [0.348, 0.666]$.

(b) Test the hypothesis in part (a) using the calculated confidence interval for $\beta_{\log(avginc)}$.

Answer: Since .5 is within the 95% confidence interval for $\beta_{\log(avginc)}$, then we fail to reject $H_0: \beta_{\log(avginc)} = 0.5$ against $H_1: \beta_{\log(avginc)} \neq 0.5$ at the 5% level.

(c) Define $\theta = \beta_{\log(pop)} - 2\beta_{pctstu}$. If you want to test $H_0: \theta = 0$ against $H_1: \theta \neq 0$. Rewrite the regression model appropriately, so that you can directly obtain $\hat{\theta}$ and the standard error $se(\hat{\theta})$.

Answer: In general terms, the regression is given by

$$\log(rent) = \beta_0 + \beta_1 \log(pop) + \beta_2 \log(avginc) + \beta_3 pctstu + u$$

If we have $\theta = \beta_1 - 2\beta_3$, then $\beta_1 = \theta + 2\beta_3$. Substituting it into the regression model, one has

$$\log(rent) = \beta_0 + \theta \log(pop) + \beta_2 \log(avginc) + \beta_3(pctstu + 2\log(pop)) + u.$$

Consequently, if we regress $\log(rent)$ on $\log(pop)$, $\log(avginc)$ and $pctstu + 2\log(pop)$, $\hat{\theta}$ is given by the coefficient of $\log(pop)$.

(d) Test the joint significance of $\beta_{\log(pop)}, \beta_{pctstu}$ and $\beta_{\log(avginc)}$ at the 5% level. That is, test $H_0: \beta_{\log(pop)} = 0, \beta_{pctstu} = 0$, and $\beta_{\log(avginc)} = 0$. (The 5% $F_{3,260}$ critical value is c = 2.60.)

Answer: This is a test for overall regression significance. That is, we test whether all slope coefficients are equal to 0.

$$F = \frac{R_u^2/k}{(1-R_u^2)/(n-k-1)} = \frac{.458/3}{(1-.458)/(264-4)} = 73.23$$

Because F > c, we reject H_0. The coefficients $\beta_{\log(pop)}, \beta_{pctstu}$ and $\beta_{\log(avginc)}$ are jointly significant at the 5% level.

Question 3 (30 Marks)

Consider the housing price model:

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + \beta_3 rooms^2 + u$$

where $price$ is the median housing price in a community, nox denotes the amount of nitrogen dioxide in the air in the community, in parts per million, and rooms is the average number of rooms in houses in the community.

(a) What is the effect of $rooms$ on $price$ in this model?

Answer: $\frac{\partial \log(price)}{\partial rooms} = \beta_2 + 2\beta_3 rooms$ or $\%\Delta price = 100(\beta_2 + 2\beta_3 rooms)\Delta rooms$.

(b) Suppose the estimated equation is:

$$\widehat{\log(price)} = \underset{(.57)}{13.39} - \underset{(.115)}{.902} \log(nox) - \underset{(.165)}{.545}\, rooms + \underset{(.013)}{.062}\, rooms^2$$

$n = 506,\ \ R^2 = .603.$

Based on your answer in part (a), what is the predicted difference in median housing prices for a community with $rooms = 5$ and a community with $rooms = 6$?

Answer: $\%\Delta price = 100(\beta_2 + 2\beta_3 rooms)\Delta rooms = 100\big(-.545 + 2(.062)(5)\big) =$ 7.5%.

(c) You want to test the joint hypothesis $H_0\colon \beta_1 = 0, \beta_2 = 0, \beta_3 = 0$ at the 5% level. The p-value associated with the *F* statistic for that test is 0.0234? What would you conclude?

Answer: 0.0234<0.05 and we reject H_0 at the 5% level. In other words, β_1, β_2 and β_3 are jointly statistically significant.

(d) Is $\widehat{\beta_1}$ economically significant? Explain.

Answer: $\widehat{\beta_1} = -0.902$ implies that a 1% increase in nox is predicted to decrease $price$ by 0.902%, holding $rooms$ fixed. This is a practically significant effect.

Question 4 (20 Marks)

Consider the following fixed-effects model for a panel dataset

$$y_{it} = \rho y_{i,t-1} + \beta x_{it} + \delta_i + u_{it} \ ,$$

where y_{it} is a continuous variable, x_{it} is an explanatory variable, δ_i is the fixed effect, u_{it} is the idiosyncratic error. $y_{i,t-1}$ is the lagged dependent variable, which is also used as an explanatory variable in the model.

(a) (10 Marks) Explain how to estimate the model using the first-differenced approach as discussed in the class.

Answer: Using the first-differenced estimator, we first take the difference of adjacent observations from the same i, and thus have

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \beta \Delta x_{it} + \Delta u_{it},$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\Delta x_{it} = x_{it} - x_{i,t-1}$, and $\Delta u_{it} = u_{it} - u_{i,t-1}$. After the transformation, the model is free of the fixed effects, and then we estimate by OLS.

(b) (10 Marks) Discuss why the first-differenced estimator always yields biased estimates for this model.

Answer: $\Delta y_{i,t-1}$ is correlated with Δu_{it}, as the former contains $y_{i,t-1}$ and the latter contains $u_{i,t-1}$, and thus the resulting estimates are biased.