Homework Set 2, ECO 311, Spring 2014

Due Date: At the beginning of class on March 31, 2014

Instruction: There are twelve questions. Each question is worth 2 points. You need to submit the answers of only $\boxed{\text{Six}}$ questions which you choose. The maximum point you can get is 12 points. For the purpose of preparing exam, you need to understand $\boxed{\text{all}}$ questions.

I will discuss the homework on the due date. Please do not ask me to go through the homework before the due date. However, you can discuss the homework with your classmates. You need to submit the homework individually though.

Q1: t test and p value

Please use the data file 311_house.dta and provide the stata commands and results. Consider the simple regression

$$\texttt{rprice} = \beta_0 + \beta_1 \texttt{area} + u$$

and the null hypothesis is

$$H_0 : \beta_1 = 30$$

Find (a) the t value for the above null hypothesis (1 point); and (b) the p value for this t test using the two-tailed alternative $H_1 : \beta_1 \neq 30$. What is your conclusion using the significance level 0.05 (1 point).

Q2: confidence interval

Please use the data file 311_house.dta and provide the stata commands and results. Consider the simple regression

$$\texttt{baths} = \beta_0 + \beta_1 \texttt{area} + u$$

Find (a) the 95% confidence intervals for β_1 (0.5 point); (b) the 90% confidence intervals for β_1 (0.5 point); (c) interpret the 90% confidence intervals (1 point).

Q3: fitted value and residual

Regression is used a lot for the purpose of investment. For example, we may want to buy an underpriced house (a bargain) if its actual price is below the price predicted by the regression. On the other hand, we may sell a house if it is overpriced. Please use the data file 311_house.dta for this problem. Consider the simple regression

$$\texttt{rprice} = \beta_0 + \beta_1 \texttt{area} + u$$

Find (a) the predicted real price for a house with 2000 square feet (1 point); and (b) the residual for the first house in the sample. How to interpret this residual? Do you want to buy or sell this house? Why? (1 point).

Q4: goodness of fit

The overall performance of a regression can be evaluated by the R-squared, which is also called the coefficient of determination. Eco 311 emphasizes causality, so we downplay the importance of R-squared. In contrast, R-squared is the focus for the course of applied regression analysis ISA 291. The formula for R-squared is

$$R^2 = 1 - \frac{SSR}{SST}$$

where $SSR = \sum \hat{u}_i^2$ is the sum of squared residual, and $SST = \sum (y_i - \bar{y})^2$ is the total sum of squares. Basically, R-squared can be interpreted as the fraction of the sample variation in y that is explained by x or explained by the regression. Everything else equal, we want to use a regression that produces greater R-squared. Please use the data file 311_house.dta for this problem. Consider two simple regressions:

$$\texttt{rprice} = \beta_0 + \beta_1 \texttt{baths} + u$$

$$\texttt{rprice} = \beta_0 + \beta_1 \texttt{area} + u$$

Notice that the two regressions share the same dependent variable; otherwise the R-squared is not comparable. Please (a) interpret the R-squared in each regression (1 point), and (b) explain which model has better fit (1 point).

Q5: standard error and t value

Most often, we run a regression in the hope that some key regressor is statistically significant. For example, we may hope to find that working out regularly has significant effect on the body weight. In terms of statistics, that means we hope to obtain a big t value, or equivalently, small standard error of $\hat{\beta}_1$, the coefficient of the key regressor. Please use the data file

311_house.dta for this problem. Consider running two simple regressions:

$$\texttt{rprice}_i = \beta_0 + \beta_1 \texttt{baths}_i + u_i, \quad (i = 1, 2, \ldots 100)$$

$$\texttt{rprice}_i = \beta_0 + \beta_1 \texttt{baths}_i + u_i, \quad (i = 1, 2, \ldots 321)$$

So the first regression uses only the first 100 houses (observations), while the second regression uses all 321 houses in the sample. Please explain in detail why the second regression has smaller standard error and bigger t value for $\hat{\beta}_1$ than the first regression. Be specific! (Hint: consider the stata command `reg y x in 1/100`)

Q6: Orthogonal Regressors

One variable is orthogonal to the other if their covariance is zero. Suppose the true model is a multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, E(u|X_1, X_2) = 0.$$

Assume $\texttt{cov}(X_1, X_2) = 0$ (so the two regressors are orthogonal), and we run a simple regression of

$$Y = \beta_0 + \beta_1 X_1 + e$$

where e is the simple-regression error term

1. (1 point) Explain whether $\hat{\beta}_1$ in the simple regression is an unbiased estimate for β_1. You may read formula 3.45 and Table 3.2 of the textbook.

2. (1 point) Use math and compare $\texttt{var}(e)$ to $\texttt{var}(u)$. Which regression has more precise estimate for β_1? Why? Hint

$$\texttt{var}(A + B) = \texttt{var}(A) + \texttt{var}(B) + 2\texttt{cov}(A, B)$$

Q7: Frisch-Waugh Theorem

This exercise intends to show how to apply the Frisch-Waugh Theorem. Read page 76-79 (5th edition) of the textbook

1. (1 point) Use the data file 311_wage1.dta, and report the regression $\texttt{wage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{exper} + u$. Interpret $\hat{\beta}_1$ and $\hat{\beta}_2$

2. (1 point) Suppose the key regressor is educ. Please show the stata code and result that implement the two-step procedure of the Frisch-Waugh Theorem to get $\hat{\beta}_1$

Q8: F Test for Overall Significance of a Regression

Use the data file 311_wage1.dta. You may read page 152-153 (5th edition) of the textbook.

1. (1 point) Report the unrestricted regression $\texttt{wage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{exper} + u$, and find the F test for overall significance of this regression. Specify the null hypothesis for this F test.

2. (1 point) Report the restricted regression that imposes the restriction in the null hypothesis. Find the residual sum squares of both restricted and unrestricted regressions and construct the F test explicitly using formula 4.37 of textbook.

Q9: Testing Hypothesis about a Single Linear Combination of Coefficient

Use the data file 311_wage1.dta.

1. (1 point) Suppose the model is $\texttt{wage} = \beta_0 + \beta_1 \texttt{female} + \beta_2 \texttt{married} + u$. Report the regression result, and report the F statistic for the hypothesis

$$H_0 : \beta_1 = \beta_2$$

   Hint: stata command is $\boxed{\text{test female = married}}$

2. (1 point) Run a transformed regression and get a t statistic for H_0. You need to verify that $\texttt{F statistic} = t^2$. You may read Section 4.4 of the textbook.

Q10: Variance of OLS Estimator

This exercise shows how to apply formula 3.51 of the textbook. Use the data file 311_wage1.dta.

1. (1 point) Find the variance of $\hat{\beta}_1$ in the regression $\texttt{wage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{exper} + u$. Find the estimate for σ^2.

2. (1 point) Please show the stata code and result that obtain $\texttt{SST}_j$ and R_j^2 in formula 3.51, where $j = 1$ for this problem (X_1=educ is the key regressor).

Q11: Quadratic and Interaction Terms

Use the data file 311_house.dta. You may read example 6.2 and 6.3 in the textbook.

1. (1 point) Report the regression with a quadratic term $\texttt{rprice} = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{age}^2 + u$. Explain whether $\texttt{age}^2$ should be included. When a house gets older, does the house price fall at increasing or decreasing rate? Hint: stata command to get the quadratic term is $\boxed{\text{gen agesq} = \text{age}^2}$

2. (1 point) Report the regression with an interaction term $\texttt{rprice} = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{age} * \texttt{bath} + u$. Compare two houses, one with one bathroom and one with two bathrooms. Which one's value falls faster when it gets older? Why?

Q12: Log and Adjusted R Squared

Use the data file 311_house.dta. You may read Table 2.3, page 191-193, 202-205 (5th edition) of the textbook.

1. (1 point) Report regression $\log(\texttt{rprice}) = \beta_0 + \beta_1 \texttt{age} + u$, and regression $\log(\texttt{rprice}) = \beta_0 + \beta_1 \log(\texttt{age}) + u$. Interpret $\hat{\beta}_1$ in each regression

2. (1 point) How to tell which regression fits data better? (Hint: consider adjusted R^2)