Quantitative Economics

Tutorial 3a: Introduction to Regression with the Population

This tutorial reviews some of the fundamentals of conditional expectation function, regressions etc and consolidates the material on handling elementary probability/random variables etc studied in week 1.

1. Given the CEF decomposition
$$Y = \mathbb{E}\left[Y|X\right] + e$$
show that
$$Var\left(Y\right) = Var\left(\mathbb{E}\left[Y|X\right]\right) + \mathbb{E}\left(e^2\right)$$
and that
$$Var\left(Y\right) = Var\left(\mathbb{E}\left[Y|X\right]\right) + \mathbb{E}\left[Var\left(Y|X\right)\right]$$

2. Galton (1886) studied the relationship between the height of parents (X) and their grown up children (Y).

   (a) Assuming that the distribution of height in the adult population is approximately the same in each generation with a mean of μ and a variance of σ^2, show that in the linear regression model
   $$Y = \alpha + \beta X + e$$
   i. the coefficient β is equal to the intergenerational correlation coefficient of height and is therefore $|\beta| < 1$,

   ii. the coefficient α is equal to $(1-\beta)\,\mu$ and that therefore
   $$\mathbb{E}[Y|X] = (1-\beta)\mu + \beta X.$$

   (b) Interpret this model. Given that $|\beta| < 1$,what does it mean about the intergenerational dynamics of height?

   (c) Consider the reverse regression
   $$X = \alpha^* + \beta^* Y + e^*$$
   i. Show that the coefficient β^* is equal to β and that the coefficient α^* is equal to α.

   ii. Why is it not the case that
   $$X = -\frac{\alpha}{\beta} + \frac{1}{\beta}Y - \frac{1}{\beta}e \quad ?$$
   iii. The result in c.(i) depends crucially on the assumption that the distributions of Y and X are stable. Explain.

(d) Suppose we made everyone studying QE take a quiz at the start of the next lecture; then took the bottom 20% and gave them some special extra on-line QE tutorials. How would you expect them to score in the next quiz? What if we gave the top 20% the on-line support instead? How should we assess whether the on-line course is effective?

3. Download the data on class sizes and test scores (either the Excel spreadsheet `ClassSize.xls` or if you prefer the R dataset `ClassSize.RData`). The data record 5th grade class sizes and the performance of pupils in a standardised math(s) test administered in the 5 grade in Californian schools - they are based on, but not identical to, the data in Stock and Watson, Chapter 4. You may treat these data as the population of interest. Using either Excel or R as you prefer:

   (a) Draw the scatter plot of the data (class size on the horizontal axis)

   (b) Calculate the conditional mean of the test score at each value of the class size variable and plot these conditional means on your scatterplot

   (c) Calculate the coefficients of the regression of test scores on class size.

   (d) Interpret the coefficient on Class Size.

   (e) Plot the linear regression on your scatter plot.

   (f) An implication of the fact that the linear regression approximates the CEF is that regression coefficients can be calculated by using $\mathbb{E}[Y|X]$ as the dependent variable. Using the data demonstrate that this is the case.

Quantitative Economics

TUTORIAL 3B: INTRODUCTION TO REGRESSION WITH A SAMPLE

(SET BY PATRICIA RICE)

This tutorial reviews the basic concepts and methods used in least squares estimation, provides some practice with techniques of statistical inference (from weeks 2 and 3) in the context of sample regressions and considers the interpretation of regression results.

1. Consider the following model:

$$Y_i = \beta_0 + u_i.$$

Derive the Least Squares estimator for β_0. Now consider the following model:

$$Y_i = \beta_1 X_i + u_i.$$

Derive the LS estimator for β_1.

2. The OLS slope estimator is not defined if there is no variation in the data for the explanatory variable. You are interested in estimating a regression relating earnings to years of schooling. Imagine that you had collected data on earnings for different individuals, but that all these individuals had completed a college education (16 years of education). Sketch what the data would look like and explain intuitively why the OLS coefficient does not exist in this situation.

3. Assume that there is a change in the units of measurement on both Y, the dependent variable and X, the sole regressor. The new variables are $Y^* = aY$ and $X^* = bX$. What effect will this change have on the regression slope? How about if there was a change in the units for X alone?

4. You have analysed the relationship between the weight and height of individuals. Although you are quite confident about the accuracy of your measurements, you feel that some of the observations are extreme, say, two standard deviations above and below the mean. Your therefore decide to disregard these individuals. What consequence will this have on the standard deviation of the OLS estimator of the slope?

5. It is proposed to test the joint hypothesis $\beta_1 = \beta_2 = 0$ by sequentially testing the individual hypotheses $\beta_1 = 0$ and $\beta_2 = 0$ at the 5 percent significance level. Assuming that the OLS estimators $\hat{\beta}_1, \hat{\beta}_2$ are independent , show that using this 'one at a time' procedure, the probability of incorrectly rejecting the joint hypothesis $\beta_1 = \beta_2 = 0$ is greater than 5 percent

Hint : The decision rule for the "one at time" test is:

Reject H_0: $\beta_1 = \beta_2 = 0$ if $|t_1| > 1.96$ and/or $|t_2| > 1.96$

Where t_1 and t_2 are the t-statistics for the individual hypotheses. Compute the probability $[|t_1| > 1.96$ and/or $|t_2| > 1.96]$

6. A subsample from the Current Population Survey is taken, on weekly earnings of individuals, their age, and their gender. You have read in the news that women make 70 cents to the $1 that men earn. To test this hypothesis, you first regress earnings on a constant and a binary variable, which takes on a value of 1 for females and is 0 otherwise. The results were:

$$\widehat{Earn} = 570.70 - 170.72 \times \textit{Female}, \qquad R^2 = 0.084,\ \textit{SER} = 282.12.$$

(a) There are 850 females in your sample and 894 males. What are the mean earnings of males and females in this sample? What is the percentage of average female income to male income?

(b) You decide to control for age (in years) in your regression results because older people, up to a point, earn more on average than younger people. This regression output is as follows:

$$\widehat{Earn} = 323.70 - 169.78 \times \textit{Female} + 5.15 \times \textit{Age}, \qquad R^2 = 0.135,\ \textit{SER} = 274.45.$$

Interpret these results carefully. How much, on average, does a 40-year-old female make per year in your sample? What about a 20-year-old male? Does this represent stronger evidence of discrimination against females?

7. There has been much debate about the impact of minimum wages on employment and unemployment in the US. While most of the focus has been on the employment-to-population ratio of teenagers, you decide to check if aggregate state unemployment rates have been affected. Your idea is to see if state unemployment rates for the 48 contiguous U.S. states in 1985 can predict the unemployment rate for the same states in 1995, and if this prediction can be improved upon by entering a binary variable for "high impact" minimum wage states. One labour economist labelled states as high impact if a large fraction of teenagers was affected by the 1990 and 1991 federal minimum wage increases. Your first regression results in the following output:

$$\widehat{Ur_i^{95}} = \underset{(0.56)}{3.19} + \underset{(0.07)}{0.27} \times \widehat{Ur_i^{85}}, \quad R^2 = 0.21,\ \text{SER} = 1.031$$

(a) Sketch the regression line and add a 45° line to the graph. Interpret the regression results. What would the interpretation be if the fitted line coincided with the 45° line?

(b) Adding the binary variable *DhiImpact* by allowing the slope and intercept to differ, results in the following fitted line:

$$\widehat{Ur_i^{95}} = \underset{(0.66)}{4.02} + \underset{(0.09)}{0.16} \times \widehat{Ur_i^{85}} - \underset{(0.89)}{3.25} \times \textit{DhiImpact} + \underset{(0.11)}{0.38} \times (\textit{DhiImpact} \times \widehat{Ur_i^{85}}),$$

where $R^2 = 0.31$, SER=0.987. Compute the 'rule of thumb' F-statistic for the null hypothesis that both parameters involving the high impact minimum wage variable are zero. Can you reject the null hypothesis that both coefficients are zero? Sketch the two regression lines together with the 45° line and interpret the results again.

(c) To check the robustness of these results, you repeat the exercise using a new binary variable for the so-called mining state (*Dmining*), i.e., the eleven states that have at least three percent of their total state earnings derived from oil, gas extraction, and coal mining, in the 1980s. This results in the following output:

$$\begin{array}{l}\widehat{Ur_i}^{95} = \underset{(0.65)}{4.04} + \underset{(0.09)}{0.15} \times \widehat{Ur_i}^{85} - \underset{(0.90)}{2.92} \times Dmining + \underset{(0.10)}{0.37} \times (Dmining \times \widehat{Ur_i}^{85}),\end{array}$$

where $R^2 = 0.31$, SER=0.997. How confident are you that the previously found effect is due to minimum wages?

8. The following table describes a dataset recording the maths test scores and a number of pupil, school and teacher characteristics for a sample of 6028 children:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  maths_test |      6028    61.81252    3.989951       48.7       77.4
       black |      6028     .325647    .4686549          0          1
other_non_wh |      6028    .0066357    .0811958          0          1
 summer_baby |      6028    .2720637    .4450594          0          1
  free_lunch |      6028    .4732913    .4993276          0          1
urban_school |      6028    .1989051    .3992096          0          1
suburban_sch |      6028    .2463504    .4309207          0          1
 small_class |      6028    .3214997    .4670908          0          1
 teacher_exp |      6028    13.93149    8.625292          0         38
  higher_deg |      6028    .4308228    .4952325          0          1
```

The variable definitions are as follows:

`maths_test` - Maths test score out of 100
`black` - Race of pupil: black = 1, 0 otherwise
`other_non_wh` - Race of pupil: non-white and not black = 1, 0 otherwise
`summer_baby` - Pupil born in June-August = 1, 0 otherwise
`free_lunch` - Pupil receives free school lunches = 1, 0 otherwise
`urban_school` - Urban school = 1, 0 otherwise
`suburban_sch` - Suburban school = 1, 0 otherwise
`small_class` - Small class size = 1, 0 otherwise
`teacher_exp` - Teaching experience of class teacher (completed years)
`higher_deg` - Class teacher has a higher degree = 1, 0 otherwise.

The next table describes the results of a regression of maths test scores on pupil, school and class teacher characteristics.

```
      Source |       SS       df       MS
-------------+------------------------------
       Model |  12851.2921     9  1427.92134
    Residual |  83096.7914  6018  13.8080411
-------------+------------------------------
       Total |  95948.0835  6027  15.9197086
-------------------------------------------
maths_test   |      Coef.   Std. Err.
-------------+-----------------------------
    constant |   62.88789   .1190653
       black |  -1.530203    .159659
other_non_wh |   .9023426   .5930611
 summer_baby |  -.5747223   .1076502
  free_lunch |  -1.794279   .1102697
urban_school |  -.0316958   .1881001
suburban_sch |  -.2144412   .1295874
 small_class |   .6583383   .1030811
 teacher_exp |     .00399   .0056772
  higher_deg |   .4977059    .098805
----------------------------------------
Number of obs =    6028
```

(a) Compute the following sample statistics for the regression model

- R-squared and adjusted R-squared
- Standard error of the regression

(b) Test the following hypotheses

- Teacher experience has no effect on math test scores
- Small class size increases math test scores

In each case, explain fully the null and alternative hypothesis, test statistic, decision rule and conclusion

(c) Construct a 99% confidence interval for the effect of being a "summer baby" on math test scores

(d) Assess the evidence for ethnic differences in math test scores.

(e) Outline the policy implications which you can draw from this regression .