Econ 140 - Spring 2016
Section 4

GSI: Fenella Carpena

February 11, 2016

1 Population Regression vs. Sample Regression, Ordinary Least Squares (OLS) Estimation

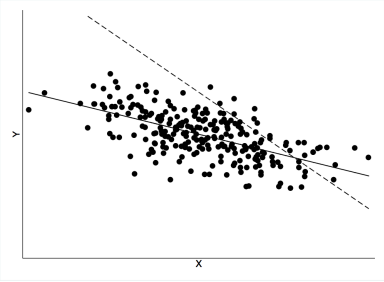
Exercise 1.1. (Stock & Watson, Review the Concepts 4.1) Explain the difference between $\widehat{\beta}_1$ and β_1; between the residual $\widehat{u}_i$ and the regression error u_i; and between the OLS predicted value $\widehat{Y}_i$ and $E(Y_i|X_i)$?

- β_1 is the slope of the population regression line. It is a constant and is unknown. $\widehat{\beta}_1$ is the slope of the sample regression line. It is our estimate of β_1 using our sample. Note that $\widehat{\beta}_1$ is a random variable because the value of $\widehat{\beta}_1$ depends on the sample we get.
- u_i is the population regression error. Note that u_i is unobserved and unknown, since $u_i = Y_i - \beta_0 - \beta_1 X_i$ and β_0, β_1 are unknown. $\widehat{u}_i$ is the residual. In contrast to u_i, $\widehat{u}_i$ is observed. $\widehat{u}_i$ is our estimator of u_i.
- $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$ is the conditional expectation function, it is unknown. In contrast, $\widehat{Y}_i$ is the predicted value based on the sample regression line, and it is known.

Exercise 1.2 (Final Exam, Fall 2014) Consider the population regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

(a) What minimization problem does OLS solve in order to estimate β_0 and β_1? Express the problem mathematically (there is no need to solve the minimization problem problem).
$\min_{b_0,b_1} \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$

(b) The figure below shows data from a sample of 250 observations of X and Y . One of the lines is the sample regression line, $\widehat{\beta}_0 + \widehat{\beta}_1 X_i$; the other is the population regression line $\beta_0 + \beta_1 X_i$. Is the sample regression line solid or dashed? Explain.



The sample regression line is the solid line. This is because the sample regression line minimizes the sum of squared residuals and we can see that the solid line fits the scatter plot better than dashed line.

2 Interpreting OLS Regression Coefficients

Exercise 2.1 A regression of *wage* (hourly wage, measured in 1979 dollars per hour) and *educ* (years of schooling) using data from a random sample of 526 American workers yields the following:

$$\widehat{wage} = -0.90 + 0.54 \cdot educ$$

(a) Interpret the intercept of this regression.
The intercept of -0.90 literally means that a person with 0 years of education has a predicted hourly wage of -$0.90 dollars per hour. This is non-sensical, so it would be difficult to attach any real-world meaning to it.

(b) It turns out that all workers in the data have at least 8 years of education. Does this help reconcile your answer in part (a)?
Yes. Since there are no workers in the sample who have very low levels of education, it is not surprising that the regression line does poorly at predicting wages when an individual has 0 years of schooling. In this case, the intercept lies far from the data (a case of extrapolation) so our estimate of wages for those with little schooling is not well determined.

(c) Interpret the slope of this regression.
The slope estimate of 0.54 implies that one more year of education is associated with an increase in the hourly wage of 0.54 cents per hour.

(d) A worker has 16 years of education. What is the regression's prediction for the worker's hourly wage?
$\widehat{wage} = -0.90 + 0.54 \cdot 16 = 7.74$

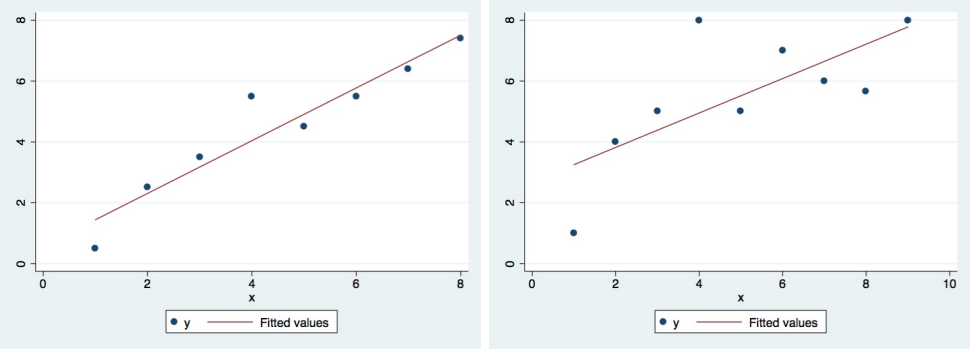
(e) Suppose a worker obtains 4 more years of schooling. What is the regression's prediction for the change in the worker's hourly wage?
The predicted wage would increase by $4 \cdot 0.54 = 2.16$

3 Measures of Fit

Exercise 3.1 (Stock & Watson, Review the Concepts 4.3) SER and R^2 are "measures of fit" for a regression. Explain how the SER measures the fit of a regression. What are the units of SER? Explain how R^2 measures the fit of a regression. What are the units of R^2?

- The SER is the standard error of the regression, which is an estimator of the SD of the population error u_i. The formula is $SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2}$. Since $\widehat{u}_i$ is the difference between our prediction $\widehat{Y}_i$ and the true data point Y_i, looking at the spread of the $\widehat{u}_i$'s (as measured by the SER) gives us a measure of the fit of the regression.
- The units of $\widehat{u}_i$ are the same as that of Y_i, so the units of SER as the same as Y_i.
- R^2 is the fraction of the variation in Y that is explained by X. The formula for R^2 is $R^2 = ESS/TSS$, where ESS is the explained sum of squares, and TSS is the total sum of squares, $ESS = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$, $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$. To see how the R^2 measures the fit of the regression, it is useful to look at two extreme cases.
  - Case 1: Suppose that X explains all of the variation in Y, so $\widehat{Y}_i = Y_i$ for all i. Then, $ESS = TSS$, so $R^2 = 1$.
  - Case 2: Suppose that X explains none of the variation in Y, so that $\widehat{\beta}_1 = 0$. This means $\widehat{Y} = \widehat{\beta}_0 = \overline{Y}$, so $ESS = 0 \implies R^2 = 0$.
- R^2 is scale-free (has no units).

Exercise 3.2 (Stock & Watson, Review the Concepts 4.4) Sketch a hypothetical scatterplot of data for an estimated regression with $R^2 = 0.9$. Sketch a hypothetical scatterplot of data for a regression with $R^2 = 0.5$.
Answers will vary depending on how you sketched your scatterplot, but here are two plots with $R^2 = 0.9$ (left graph) and $R^2 = 0.5$ (right graph). As you can see, the plot with $R^2 = 0.9$ has data points that are closer to the regression line than the plot with $R^2 = 0.5$.



Exercise 3.3 Suppose that the R^2 from the regression in Exercise 2.1 is equal to 0.242. How would you interpret this R^2?
24.2% of the variation in wages is explained by education.

4 Least Squares Assumptions

Exercise 4.1. (Stock & Watson, Chapter 4, Review the Concepts, Exercise 2) For each least squares assumption, provide an example in which the assumption is valid and then provide an example in which the assumption fails.

- Assumption 1: $E(u_i|X_i) = 0$. Valid when X_i is randomly assigned, because random assignment implies X_i and u_i are independent, hence $E(u_i|X_i) = 0$. Fails when u_i and X_i are correlated.
- Assumption 2: (X_i, Y_i) are i.i.d. Valid when observations are drawn using random sampling. Fails when data is not from random sampling (e.g., we collect data from the first 10 people that choose to respond to our survey). Also fails when we have time series data: e.g., if X is US dollar exchange rate, and Y is US exports over time, we might think that the exchange last month will affect exports this month, in which case $cov(X_1, Y_2) \neq 0$ violating independence between any pair of X_i and Y_j.
- Assumption 3: Outliers are unlikely, i.e., finite fourth moment $0 < E(X^4) < \infty$, $0 < E(Y^4) < \infty$. Valid when X and Y have a finite range since if the range is finite, the kurtosis (fatness of the tails) will also be finite. Fails when there are data entry errors that cause outliers.

Exercise 4.2. (From Section Notes 4 in bCourses (All Students) > Files > Discussion Section) A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam while others have 120 minutes. Each student is randomly assigned one of the exam times based on the flip of a coin. Let Y_i denote the number of points scored on the exam by the i^{th} student ($0 \leq Y_i \leq 100$). Let X_i denote the amount of time the student has to complete the exam ($X_i = \{90, 120\}$), and consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

1. Explain what the term u_i represents. Why will different students have different values of u_i?
   u_i represents factors other than time that influence the student's performance on the exam other than time allowed during the exam, including amount of time studying, aptitude for the material, and etc. Some students will have studied more than average, other less; some students will have higher than average aptitude for the subject, others lower, and so forth.

2. Explain why $\mathbb{E}[u_i|X_i] = 0$ for this regression model.
   Because X_i is randomly assigned, u_i is independent of X_i. There is no unobserved factor that is correlated with the amount of time a student receives to take the exam. Because u and X are independent $E(u_i|X_i) = E(u_i) = 0$.

3. Are the other least squares assumptions satisfied?
   Assumption #2 is satisfied if this year's class is typical of other classes, that is, students in this year's class can be viewed as random draws from the population of students that enroll in the class. Assumption #3 is satisfied because $0 \leq Y_i \leq 100$ and X_i can take on only two values (90 and 120).

5 Regressions in Stata

Exercise 6.1 The table below shows regression output from Stata.

```
Linear regression                                      Number of obs =      74
                                                       F(  1,    72) =   17.28
                                                       Prob > F      =  0.0001
                                                       R-squared     =  0.2196
                                                       Root MSE      =  2623.7

------------------------------------------------------------------------------
             |               Robust
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -238.8943   57.47701    -4.16   0.000    -353.4727    -124.316
       _cons |   11253.06   1376.393     8.18   0.000     8509.272    13996.85
------------------------------------------------------------------------------
```

Identify the following from the above regression output.

(a) Dependent and independent variables
   Dependent variable: price, independent variable: mpg

(b) Sample size
   $n = 74$

(c) R^2
   $R^2 = 0.2196$

(d) SER
   $SER = 2623.7$ (Root MSE in the above table)

(e) $\widehat{\beta}_0$ and $\widehat{\beta}_1$
   $\widehat{\beta}_0 = 11253.06, \widehat{\beta}_1 = -238.8943$, so the sample regression line is $\widehat{price} = 11253.06 - 238.8943 \cdot mpg$.