Econ 140 - Spring 2016
Section 8

GSI: Fenella Carpena

March 17, 2016

Additional Exercises

Question 1. For each of the following functions, state whether it can be linearized. If yes, write the resulting regression function in a form that can be estimated using OLS. If no, explain why.

1. $Y_i = \beta_0 X_{1i}^{\beta_1} e^{\beta_2 X_2 i}$
2. $Y_i = \frac{X_i}{\beta_0 + \beta_1 X_i}$
3. $Y_i = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$
4. $Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} + u_i$

Answer: (1) Yes, $ln(Y_i) = ln(\beta_0) + \beta_1 ln(X_{1i}) + \beta_2 X_{2i}$; (2) Yes, $(1/Y_i) = \beta_0(1/X_i) + \beta_1$; (3) Yes, $ln(Y_i/(1 - Y_i)) = \beta_0 + \beta_1 X_i$; (4) No, it cannot be linearized due to the additive term u_i.

Question 2. Consider the following two regressions of wages on age and gender.

$$\widehat{Earn} = \underset{(21.18)}{323.70} + \underset{(0.55)}{5.15} \cdot Age - \underset{(13.06)}{169.78} \cdot Female$$

$$R^2 = 0.13, \ SER = 274.75$$

and

$$\widehat{ln(Earn)} = \underset{(0.08)}{5.44} + \underset{(0.002)}{0.015} \cdot Age - \underset{(0.036)}{0.421} \cdot Female$$

$$R^2 = 0.17, \ SER = 0.75$$

where $Earn$ is weekly earnings in dollars, Age is measured in years, and $Female$ is a dummy variable equal to 1 if the individual is female, and 0 otherwise.

(a) Interpret each regression carefully. For a given age, how much less do females earn on average? Should you choose the second specification on grounds of the higher regression R^2?

Answer: The first regression (which has a linear specification) suggests that every additional year in age is associated with \$5.15 more weekly earnings, holding gender constant. In this regression, women on average also earn \$169.78 less wages for any given age. The intercept has no useful interpretation since there are likely no observations who are male and have zero age. The regression explains 13% of the variation in weekly earnings.

The second regression (which has a log-linear specification) suggests that for every additional year in age, earnings increase by 1.5%. Women on average earn 42.1% less than men for any given age. Again, the intercept has no useful interpretation since there are likely no observations who are male and have zero age. The regression explains 17% of the variation in log earnings.

Even if the R^2 in the second regression is higher, we should not choose the second specification since the dependent variable in the two regressions is different, thus the R^2 cannot be compared.

(b) Suppose that your professor points out to you that age and ln(earn) profiles typically take on an inverted U-shape. How would you extend the previous regression to test this idea?

Answer: You can add Age^2 in the regression, specifically, regress $ln(Earn)$ on Age and Age^2. Then, an inverted U-shape would mean that the coefficient on Age^2 would be negative.

(c) Now, consider the regression where you add the square of age to your log-linear regression in part (a).

$$\widehat{ln(Earn)} = \underset{(0.18)}{3.04} + \underset{(0.009)}{0.147} \cdot Age - \underset{(0.033)}{0.42} \cdot Female - \underset{(0.0001)}{0.0016} Age^2$$

$$R^2 = 0.28, \ SER = 0.68$$

Interpret the results from the above regression. Why is the Age coefficient here so large relative to its value in the regression in part (a)?

Answer: The coefficient on the variable that was added, Age^2, is statistically significant and has resulted in a substantial increase in the regression R^2. The increase in the Age coefficient is due to the fact that earnings increase more initially than later in life or, mathematically speaking, it compensates for the negative coefficient on Age^2, which lowers earnings as individuals become older.

Question 3. Suppose you have data on weight, age, height and gender for 100 male and female children, between the ages of 9 and 12, who are all at least 4 feet tall. Using this data, you estimate the following relationship

$$\widehat{Weight} = \underset{(3.81)}{45.59} + \underset{(0.46)}{4.32} \cdot Height4$$

$$R^2 = 0.55, \ SER = 15.69$$

where $Weight$ is in pounds, and the $Height4$ variable is inches above 4 feet (so for a child who is 4 feet tall, $Height4$ takes on the value 0, while for a child who is 4 feet and 5 inches tall, $Height4$ takes on the value 5).

(a) Interpret the results of this regression.

Answer: The average weight of children in the sample who are exactly 4 feet tall is 45.59. For every inch above 4 feet, children in the sample gain roughly 4.32 pounds. The regression explains 55 percent of the weight variation for children in the sample.

(b) You remember from the medical literature that females in the adult population are, on average, shorter than males and weigh less. You also seem to have heard that females, controlling for height, are supposed to weigh less than males. To see if this relationship holds for children, you add a binary variable (DFY) that takes on the value one for girls and is zero otherwise. You estimate the following regression function:

$$\widehat{Weight} = \underset{(5.99)}{36.27} + \underset{(7.36)}{17.33} \cdot DFY + \underset{(0.80)}{5.32} \cdot Height4 - \underset{(0.90)}{1.83} \cdot DFY \cdot Height4$$

$$R^2 = 0.58, \ SER = 15.41$$

Are the signs on the new coefficients as expected? Are the new coefficients individually statistically significant? Write down and sketch the regression function for boys and girls separately.

Answer: The regression results show that on average, short girls weigh more than short boys, and also that tall boys weight more than tall girls. If we think that the findings from the medical literature for adults also applies to children, then these findings are perhaps unexpected. The coefficient on DFY is statistically significant at the 95% level, and the same is true for the coefficient on $DFY \cdot Height$. The regression function for boys is $\widehat{Weight} = 36.72 + 5.32 \cdot Height4$, and for girls it is $\widehat{Weight} = 53.60 + 3.49 \cdot Height4$. So if you sketched the regression line, the one for girls would have a higher intercept and flatter slope, relative to the regression line for boys.

(c) Using the regression in part (b), state the hypothesis that the regression function is identical for boys and girls. What test statistic would you use to this hypothesis?

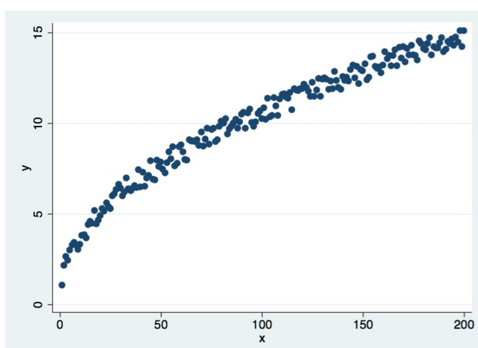
Answer: If we write the population regression function as $Weight = \beta_0+\beta_1 DFY+\beta_2 Height4+\beta_3 DFY \cdot Height4+u$, the hypothesis that the regression function is the identical for boys and girls can be written as $H_0 : \beta_1 = \beta_3 = 0$. To test this hypothesis, we could use an F-test.

(d) Consider the regression in part (b) but now assume that *in addition* to testing whether the relationship between height and weight changes by gender, you also wanted to test if the relationship between height and weight changes by age. Briefly outline how you would specify the regression to test this relationship, where the regression includes the gender binary variable (DFY) and an age binary variable (call it *Older*) that takes on a value of one for eleven to twelve year olds and is zero otherwise. How would the estimated relationship vary between the following four groups: younger girls, older girls, younger boys, and older boys?

Answer: $Weight = \beta_0 + \beta_1 DFY + \beta_2 Height4 + \beta_3 DFY \cdot Height4 + \beta_4 Older + \beta_5 Older \cdot Height4 + u$. The estimates of $\widehat{Weight}$ would vary as follows.

	Younger	Older
Boys	$\hat\beta_0 + \hat\beta_2 Height4$	$\hat\beta_0 + \hat\beta_4 + (\hat\beta_2 + \hat\beta_5)Height4$
Girls	$\hat\beta_0 + \hat\beta_1 + (\hat\beta_2 + \hat\beta_3)Height4$	$\hat\beta_0 + \hat\beta_1 + \hat\beta_4 + (\hat\beta_2 + \hat\beta_3 + \hat\beta_5)Height4$

Question 4. Consider the scatterplot of y and x below. Explain what transformation you would use, and what regression you would estimate to model this pattern. Can you think of two variables that might have an economic relationship shaped like this?



Answer: Note that there is some curvature in the shape of the scatterplot; that is, if we were to fit curve through these points, it would be steep for small values of x, but less steep for large values of x, so we could transformthe x variable into $ln(x)$. The sample regression would then be $\widehat{y} = \beta_0 + \beta_1 \cdot ln(x)$.

There are many economic relationships with this shape. For example, this shape might represent the decreasing marginal product of labor in production, where x is labor and y is output (e.g, if we have only 10 sewing machines in our clothing factory, then going from 9 to 10 workers is more valuable to production than going from 100 to 101 workers).

Question 5. A regression of *wage* (hourly wage, measured in 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.

> **Answer:** The intercept of -0.90 literally means that a person with 0 years of education has a predicted hourly wage of -90 cents an hour.

Suppose that using $ln(wage)$ instead as the response variable, we obtain the following regression:

$$\widehat{ln(wage)} = 0.584 + 0.083 \cdot educ$$

$$n = 526, \ R^2 = 0.186$$

(b) Interpret the slope. Compare the interpretation of the slope in the two regressions when the response variable is $ln(wage)$ vs. $wage$.

> **Answer:** To interpret the slope 0.083, we say that an additional year of education is associated with a 8.3% increase in hourly wage. In part (a), the slope obtained was 0.54, which means that each additional year of schooling is associated with an increase in hourly wage of 54 cents. This 54 cent increase is the same whether it is for the 1st year of education or the 20th year of education. In contrast, the regression above instead imposes a constant percentage effect of education on wage.

(c) Interpret the R^2 in the regression where $ln(wage)$ is the dependent variable.

> **Answer:** The R^2 shows that the variable *educ* explains about 18.6% of the variation in $ln(wage)$ (**NOT** *wage*).