ØAMET4100 · Spring 2019
Worksheet 4A

Instructor: Fenella Carpena

January 31, 2019

1 Nonlinear Regression Functions

Exercise 1.1 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.

(a) $Y_i = \frac{X_i}{\beta_0+\beta_1 X_i}$

(b) $Y_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$

(c) $Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} + u_i$

Exercise 1.2 Suppose that for your master's thesis, you conduct a study to explore the "gender gap" in earnings in top corporate jobs in the US. To do so, you examine total compensation among top executives in a large set of public corporations in the 1990s, and you estimate the following regression.

$$\widehat{\ln(Earn)} = \underset{(0.08)}{5.44} - \underset{(0.036)}{0.421} \cdot Female + \underset{(0.002)}{0.015} \cdot Age \qquad R^2 = 0.17$$

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 the estimated coefficient on $Female$ and the R^2.

(b) Does the above regression suggest that female top executives earn less than male top executives? Explain.

(c) Does the above regression suggest that there is gender discrimination? Explain.

(d) Your thesis supervisor points out to you that age and $ln(Earn)$ profiles typically take on an inverted U-shape. Describe how you would extend the above regression to test this idea.

(e) You add two new variables to the regression, namely, the market value of the firm (a measure of firm size, in millions of dollars) and stock return (a measure of firm performance, in percentage points), and you obtain the following results.

$$\ln(\widehat{Earn}) = \underset{(0.03)}{3.86} - \underset{(0.04)}{0.28} \cdot Female + \underset{(0.001)}{0.010} Age + \underset{(0.004)}{0.37} \ln(MarketValue) + \underset{(0.003)}{0.004} Return \qquad R^2 = 0.33$$

Explain what the coefficient on $\ln(MarketValue)$ means.

(f) The coefficient on $Female$ is now -0.28. Explain why it has changed from the first regression.

(g) Are large firms more likely than small firms to have female top executives? Explain.

Exercise 1.3 You work as a strategist at an advertising company, and one of your clients—a large international clothing retailer—has asked you to help optimize its advertising efforts. The client gives you data on their sales and advertising spending for each quarter over the last ten years. Now, you would like to estimate the relationship between sales in each quarter (Y) and the level of advertising in each quarter (X).

(a) You suspect that, irrespective of advertising, sales in the fourth quarter of each year are generally higher because of holiday shopping.

(i) Write a regression model that would allow you to examine this holiday season effect. Precisely define all the variables in the model.

(ii) How would you use the regression model you wrote down in part (i) to test for a holiday season effect?

(b) You suspect that, in addition to sales in the fourth quarter being higher, the effect of advertising on sales might differ in the fourth quarter. How would you modify your model in part (a) to accommodate this? Precisely define all the variables in the model. Explain the meaning of each of the parameters in the modified model, including the intercept(s).

2 Regression Table

Exercise 2.1 The Department of Education is currently deciding whether to hire additional teachers for elementary schools across the country. Hiring more teachers will reduce the number of students per teacher (the student-teacher ratio), but doing so will cost a lot of money. The Department of Education would like your advice: if class sizes are reduced, what will be the effect on students' performance?

The attached table, which is Table 8.3 from the textbook, analyzes the relationship between test scores and student-teacher ratios across different school districts in the country. The dependent variable is the average test score in each district. The independent variables in the table include the following:

- STR: the student-teacher ratio in a district
- % English learners: the percentage of elementary students in the district who are English learners (e.g., immigrant students)

- HiEL: a dummy variable for a district with a high percentage of English learners, defined as % English learners $\geq$ 10%
- % Eligible for subsidized lunch: the percentage of elementary students in the district who are eligible for subsidized lunch at school (e.g., poor students)
- Average district income, in logarithm

The following exercises ask about the regression results in Table 8.3.

(a) Comparing columns (1) and (2) how does the coefficient of *Student-teacher ratio (STR)* change after incuding *ln(Average district income)* into the regression? Why do you think this change occurs?

(b) Looking at column (4), can we reject the hypothesis that the effect of STR is the same for districts with low and high percentages of English learners (i.e for $HiEL = 1$ and for $HiEL = 0$) at the 5% significance level?

(c) Consider the results in column (5). What can you say about the effect of STR on *testscore*? Is there a linear relationship between STR and *testscore*?

(d) Looking at column (6), which variables allow us to test if the regression function relating *testscore* and STR differ between districts with low and high percentages of English learners? Based on the results, are the regression functions different for districts with high and low percentages of English learners?

(e) Again looking at column (6), what is the predicted impact on *testscore* of a one-unit change in STR for an $HiEL = 1$ district with a baseline STR=20? For an $HiEL = 0$ district with a baseline STR=20?

TABLE 8.3 Nonlinear Regression Models of Test Scores

Dependent variable: average test score in district; 420 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (*STR*)	−1.00** (0.27)	−0.73** (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33** (24.86)	83.70** (28.50)	65.29** (25.26)
STR^2					−3.42** (1.25)	−4.38** (1.44)	−3.47** (1.27)
STR^3					0.059** (0.021)	0.075** (0.024)	0.060** (0.021)
% English learners	−0.122** (0.033)	−0.176** (0.034)					−0.166** (0.034)
% English learners ≥ 10%? (Binary, *HiEL*)			5.64 (19.51)	5.50 (9.80)	−5.47** (1.03)	816.1* (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3* (50.2)	
$HiEL \times STR^2$						6.12* (2.54)	
$HiEL \times STR^3$						−0.101* (0.043)	
% Eligible for subsidized lunch	−0.547** (0.024)	−0.398** (0.033)		−0.411** (0.029)	−0.420** (0.029)	−0.418** (0.029)	−0.402** (0.033)
Average district income (logarithm)		11.57** (1.81)		12.12** (1.80)	11.75** (1.78)	11.80** (1.78)	11.51** (1.81)
Intercept	700.2** (5.6)	658.6** (8.6)	682.2** (11.9)	653.6** (9.9)	252.0 (163.6)	122.3 (185.5)	244.8 (165.7)
***F*-Statistics and *p*-Values on Joint Hypotheses**							
(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR, HiEL \times STR^2, HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\bar{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.