Nonlinear Regression: Test Scores and Class Size

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.

Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.

The textbook comes with online resources and study guides. Other references will be given from time to time.

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 8, Exercise 3.

After reading this chapter's analysis of test scores and class size, an educator comments, "In my experience, student performance depends on class size, but not in the way your regressions say. Rather, students do well when class size is less than 20 students and do very poorly when class size is greater than 25. There are no gains from reducing class size below 20 students, the relationship is constant in the intermediate region between 20 and 25 students, and there is no loss to increasing class size when it is already greater than 25." The educator is describing a threshold effect, in which performance is constant for class sizes less than 20, jumps and is constant for class sizes between 20 and 25, and then jumps again for class sizes greater than 25. To model these threshold effects, define the binary variables:

$$STRsmall = 1 \text{ if } STR < 20 \text{ and } STRsmall = 0 \text{ otherwise;}$$
$$STRmoderate = 1 \text{ if } 20 \leq STR \leq 25 \text{ and } STRmoderate = 0 \text{ otherwise;}$$
$$STRlarge = 1 \text{ if } STR > 25 \text{ and } STRlarge = 0 \text{ otherwise.}$$

a. Consider the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRlarge_i + u_i$. Sketch the regression function relating $TestScore$ to STR for hypothetical values of the regression coefficients that are consistent with the educator's statement.

b. A researcher tries to estimate the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRmoderate_i + \beta_3 STRlarge_i + u_i$ and finds that the software gives an error message. Why?

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 8, Exercise 3.

After reading this chapter's analysis of test scores and class size, an educator comments, "In my experience, student performance depends on class size, but not in the way your regressions say. Rather, students do well when class size is less than 20 students and do very poorly when class size is greater than 25. There are no gains from reducing class size below 20 students, the relationship is constant in the intermediate region between 20 and 25 students, and there is no loss to increasing class size when it is already greater than 25." The educator is describing a threshold effect, in which performance is constant for class sizes less than 20, jumps and is constant for class sizes between 20 and 25, and then jumps again for class sizes greater than 25. To model these threshold effects, define the binary variables:

$$STRsmall = 1 \text{ if } STR < 20 \text{ and } STRsmall = 0 \text{ otherwise;}$$
$$STRmoderate = 1 \text{ if } 20 \leq STR \leq 25 \text{ and } STRmoderate = 0 \text{ otherwise;}$$
$$STRlarge = 1 \text{ if } STR > 25 \text{ and } STRlarge = 0 \text{ otherwise.}$$

a. Consider the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRlarge_i + u_i$. Sketch the regression function relating $TestScore$ to STR for hypothetical values of the regression coefficients that are consistent with the educator's statement.

b. A researcher tries to estimate the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRmoderate_i + \beta_3 STRlarge_i + u_i$ and finds that the software gives an error message. Why?

Problems and Applications

a. Consider the regression $TestScore_i = \beta_0 + \beta_1\, STRsmall_i + \beta_2\, STRlarge_i + u_i$. Sketch the regression function relating $TestScore$ to STR for hypothetical values of the regression coefficients that are consistent with the educator's statement.

Since the regression function is given, the question is effectively asking what values or range of values of the coefficients are consistent with the educator's statement. First, the signs are $\beta_1 \geq 0$ and $\beta_2 \leq 0$. The population regression line of $TestScore$ on STR has three horizontal segments. A higher segment for values of $STR < 20$, an intermediate segment for $20 \leq STR \leq 25$, and a lower segment for $STR > 25$.

Problems and Applications

a. Consider the regression $TestScore_i = \beta_0 + \beta_1 \, STRsmall_i + \beta_2 \, STRlarge_i + u_i$. Sketch the regression function relating $TestScore$ to STR for hypothetical values of the regression coefficients that are consistent with the educator's statement.

Since the regression function is given, the question is effectively asking what values or range of values of the coefficients are consistent with the educator's statement. First, the signs are $\beta_1 \geq 0$ and $\beta_2 \leq 0$. The population regression line of $TestScore$ on STR has three horizontal segments. A higher segment for values of $STR < 20$, an intermediate segment for $20 \leq STR \leq 25$, and a lower segment for $STR > 25$.

Problems and Applications

b. A researcher tries to estimate the regression $TestScore_i = \beta_0 + \beta_1\, STRsmall_i + \beta_2\, STRmoderate_i + \beta_3\, STRlarge_i + u_i$ and finds that the software gives an error message. Why?

This is the "dummy variable trap." The error is due to the perfect multicollinearity among the three binary regressors $STRsmall$, $STRmoderate$, and $STRlarge$. Because the intercept is a linear function of the three regressors, the regression cannot be estimated by least squares. The solution is to either drop one of the binary regressors or drop the intercept.

Problems and Applications

b. A researcher tries to estimate the regression $TestScore_i = \beta_0 + \beta_1\, STRsmall_i + \beta_2\, STRmoderate_i + \beta_3\, STRlarge_i + u_i$ and finds that the software gives an error message. Why?

This is the "dummy variable trap." The error is due to the perfect multicollinearity among the three binary regressors $STRsmall$, $STRmoderate$, and $STRlarge$. Because the intercept is a linear function of the three regressors, the regression cannot be estimated by least squares. The solution is to either drop one of the binary regressors or drop the intercept.