ØAMET4100 · Spring 2019
Lecture Note 4A

Instructor: Fenella Carpena

January 31, 2019

This lecture note provides a review of non-linear regression functions (Stock & Watson, Chapter 8). This lecture note is not intended to be a comprehensive review of lecture or the textbook, since there is a lot more material than we have time to cover. However, I have tried to focus on the concepts which I believe are necessary to be successful in our class.

1 Non-Linear Regression Functions: Overview

So far, we have discussed a population regression line that is linear. In a linear population regression, the slope is constant, so the effect of X on Y does not depend on X. This linearity may not appropriately capture the relationship between X and Y. For example, some relationships between X and Y might have some curvature, and we would like our regression to capture such a non-linear relationship.

Now, we consider the case where the population regression function is a nonlinear function of the independent variables, that is, $E(Y|X_{1i}, ..., X_{ki})$ is a nonlinear function of one or more of the X's. A function $f(X)$ is linear if the slope of $f(X)$ is the same for all values of X. However, if the slope of $f(X)$ depends on the value of X, then $f(X)$ is nonlinear.

Since the three models we describe below—polynomials, logarithms, and interactions—are still linear functions of the unknown population parameters (i.e., β's) of the population regression model, we can estimate these non-linear models using OLS and earlier methods used.

In general, when deciding on a regression model, we want economic theory to guide our decision as to what the relevant variables are for inclusion. Economic intuition should also guide our decision as to what is the appropriate functional form for the model.

2 Polynomials

2.1 Quadratic regression model

Consider the regression model: $Testscore_i = \beta_0 + \beta_1 Income_i + u_i$

If we believe instead that a more accurate depiction of the relationship between $Testscore$ and $Income$ is quadratic (i.e. $Testscore$ is non-linear in $Income$), we would use the **quadratic regression model**:

$$Testscore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i$$

We can test the hypothesis that the model is linear vs. quadratic by testing $H_0 : \beta_2 = 0.$ vs. $H_1 : \beta_2 \neq 0.$

2.2 Higher-order Polynomial Terms

We can also imagine a model of Y with higher order X terms:

$$Y_i = f(X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_m X_i^m + u_i \quad (1)$$

In general, if you include polynomial terms in the RHS, you should have a good motivation for doing so. For example, increasing/decreasing returns to scale is a common economic motivation for including a quadratic term (X^2).

2.3 Interpretation with Polynomial Terms

The mathematical interpretation of coefficients on polynomial terms is fairly straightforward. Suppose our model is the following:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

where the second term is quadratic in X_i. Taking the derivative of Y with respect to X_i, we see that the effect of a change in X_i on Y depends on the value of X_i itself:

$$\frac{dY_i}{dX_i} = \beta_1 + 2\beta_2 X_i$$

An example of this might be a regression of $Testscore$ on $income$ and $income^2$. This model specification implies that the effect of a change in $income$ on $Testscore$ will vary depending on the starting income level (i.e., the effect of a \$1,000 increase in income on test scores differs when we are going from \$10,000 to \$11,000 versus if we are going from \$30,000 to \$31,000).

2.4 Testing Model Specification

Which degree polynomial should you use for a polynomial model as in Equation (1) above? One approach is to use sequential hypothesis testing to see if the higher degree polynomials can be omitted from the regression (i.e., testing whether the coefficient estimates for these variables differ from zero). To do this, we can do the following steps:

1. Pick a maximum value of m and estimate the polynomial regression for that m.
2. Use the t-statistic to test the hypothesis that the coefficient on X^m (*i.e.* β_m) is zero. If you reject this hypothesis, then X^m belongs in the regression.
3. If you do not reject $\beta_m = 0$ in step 2, then eliminate X^m from the regression and estimate a polynomial regression of degree $m-1$. Test whether the coefficient on X^{m-1} is zero. If you reject the test, use the polynomial of degree $m-1$.
4. If you do not reject $\beta_{m-1} = 0$ in step 3, continue this procedure until the coefficient on the highest power in your polynomial is statistically significant.

Note that in the steps outlined here, there is one missing ingredient: the initial degree m of the polynomial. In many applications involving economic data, the nonlinear functions are smooth (i.e., they do not have jumps or spikes). In this case, it is typical to begin with $m = 2$, 3, or 4 in step 1.

3 Logarithms

We can specify nonlinear models using the natural logarithm of Y and/or X.

Observe that $d\ln(X) = \frac{dX}{X} \approx \%\Delta X/100$ for **small** ΔX. Alternatively said, when ΔX is **small**, $\ln(X + \Delta X)$ - $\ln(X)$ is approximately equal to the percentage change in X divided by 100.

We have three cases when the logarithms might be used and the interpretation of the model/coefficients will vary by case:

- **Linear-Log Model**: $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$: A 1% increase in $X \Rightarrow 0.01 \times \beta$-unit increase in Y.
- **Log-Linear Model**: $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$: A 1-unit increase in $X \Rightarrow 100 \times \beta\%$ increase in Y.
- **Log-Log Model**: $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$: A 1% increase in $X \Rightarrow$A $\beta\%$ increase in Y.

How can we compare the linear-log model and the log-log model? We cannot use the $\overline{R}^2$ because the dependent variables in these two models are different. Because of this, it is best to use economic theory to guide your decision in choosing which model is best.

3.1 Interactions Between Independent Variables:

When using binary and/or continuous variables in our regression, there are three types of interactions we might encounter.

- **Interaction between two binary variables**: $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$
- **Interaction between a continuous and a binary variable**: When we use the interaction term $X_i \times D_i$, there are three possibilities for the population regression function:
  - Scenario 1: *Different intercept, same slope*: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$
  - Scenario 2: *Different intercept, different slope*: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$
  - Scenario 3: *Same intercept, different slope*: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$
- **Interaction between two continuous variables**: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$. Here, including the interaction term allows the effect on Y of a change in X_1 to depend on the value of X_2, and conversely, allows the effect of a change in X_2 on Y to depend on the value of X_1.