CHAPTER 7: SPECIFICATION – FUNCTIONAL FORM

The previous chapter dealt with specifying the correct independent variables. In this chapter we deal with **specifying the correct functional form**. Up until now we have relied exclusively on the linear model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Again, economic theory should be used to establish the functional form of the relationship between the dependent variable and each independent variable. There are several commonly used function forms we will examine.

Before we do that, we should note that all of these forms will **always** include a constant (intercept) term. If we omit the constant term, we force the regression line to go through the origin, which might be a serious misspecification.

The Linear Form

This choice of equation is **linear in the independent variables:**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

This choice should be made when the economic theory suggests:

- A constant slope between X_1 and Y and X_2 and Y:

$$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1, \qquad \frac{\partial Y_i}{\partial X_{2i}} = \beta_2$$

- The elasticity of Y with respect to X_1 and X_2 is not constant:

$$\eta_{Y,X_1} = \frac{\partial Y_i}{\partial X_{1i}} \frac{X_{1i}}{Y_i} = \beta_1 \frac{X_{1i}}{Y_i}, \qquad \eta_{Y,X_2} = \frac{\partial Y_i}{\partial X_{2i}} \frac{X_{2i}}{Y_i} = \beta_2 \frac{X_{2i}}{Y_i}$$

The Double-Log Form

This choice of equation is **linear in the coefficients** but **not linear in the independent variables:**

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \varepsilon_i.$$

This choice should be made when the economic theory suggests:

- The slope between X_1 and Y and X_2 and Y is not constant:

$$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1 \frac{Y_i}{X_{1i}}, \qquad \frac{\partial Y_i}{\partial X_{2i}} = \beta_2 \frac{Y_i}{X_{2i}}$$

- The elasticity of Y with respect to X_1 and X_2 is constant:

$$\eta_{Y,X_1} = \frac{\partial Y_i}{\partial X_{1i}} \frac{X_{1i}}{Y_i} = \beta_1, \qquad \eta_{Y,X_2} = \frac{\partial Y_i}{\partial X_{2i}} \frac{X_{2i}}{Y_i} = \beta_2$$

- All observations on X and Y are strictly positive

The Semi-Log Form

Two popular choices of equation for this form:

$$Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$\ln Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

In the first case, X_2 is linearly related to *Y*, while X_1 is nonlinearly related to *Y*.

In the second case, the coefficients have a very useful interpretation – If X_1 increases by one unit, β_1 represents the percentage change in *Y*.

Why would this be a useful interpretation for a coefficient? Suppose *Y* is an employee's salary and *X* is years of experience. Each additional **year** of experience may be associated with a given **percentage** increase in salary.

The Polynomial Form

These types of forms are more general the linear form. The dependent variable is expressed as a function of the independent variables, some of which have been raised to powers greater than one**:**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (X_{1i})^2 + \beta_3 X_{2i} + \varepsilon_i.$$

This choice should be made when the economic theory suggests:

- The slope between X_1 and Y is not constant:

$$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1 + 2\beta_2 X_{1i}, \qquad \frac{\partial Y_i}{\partial X_{2i}} = \beta_3$$

  Note: It is difficult to interpret individual regression coefficients.

- The elasticities are not constant

The Inverse Form

This form expresses the dependent variable as a function of the reciprocal (inverse) of one or more of the independent variables:

$$Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_{1i}}\right) + \beta_2 X_{2i} + \varepsilon_i$$

This choice should be made when the economic theory suggests:

- The impact of a variable is expected to approach zero as it gets very very large
- The slope between X_1 and Y is not constant:

$$\frac{\partial Y_i}{\partial X_{1i}} = -\frac{\beta_1}{X_{1i}^2}, \qquad \frac{\partial Y_i}{\partial X_{2i}} = \beta_2$$

Notice X_{1i} cannot be zero

Problems with Incorrect Functional Form

Once again, economic theory should be used to establish the functional form of the relationship between the dependent variable and each independent variable.

Avoid using goodness of fit over the sample to make your decision about functional form:

- Cannot use $\bar{R}^2$ to compare models when Y has been transformed because the TSS will be different:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$\ln Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- The "best" fitting functional form might not be accurate outside of your sample

Dummy Variables

So far we have assumed that each explanatory variable is numerical. However, there are **many** occasions in which categorical variables need to be included in the model.

How do we quantify this qualitative information?

Consider an independent variable that takes on only two values, 0 or 1. This type of variable is called a *dummy variable*.

Consider first a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

and introduce a dummy variable X_{2i}

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

What is the effect of adding this dummy variable X_{2i} to the model?

What if the qualitative variable had more than two categories?

Suppose we were interested in measuring the effect of experience (years) and educational attainment on salaries. Specifically, we want to distinguish high school (HS) graduates from graduates with Bachelor (BA) and Master (MA) degrees.

What do we do?

Notice that TWO dummy variables can be used to identify each of the three educational achievement levels:

In general, we need one less dummy variable than categories. If you fail to adhere to this, you will have fallen into the *Dummy Variable Trap* and violated Assumption 6.

We can also use dummy variables to allow for differences in the slope coefficient by adding an **interaction term**.

Let's consider an example where we want to relate a person's hourly wage to their gender and experience.

A model that allows for intercept differences is specified as

If experience is expected to increase the wages of males and females differently, then you would need to allow for a slope dummy as well:

Example: How much does a vote cost?

This is a real-life (data) example. Suppose we want to relate a candidate's vote share (*Y*) to his/her campaign expenditures (*X*) in Canadian federal elections. What other things might matter in this model?

- Incumbency status of candidate
  - $Incumbency_i$ = 1 if the candidate is an incumbent; 0, otherwise
- Gender of candidate
  - $Male_i$ = 1 if the candidate is male; 0 if the candidate is female

- Party affiliation of candidate (major parties only)
  There are four major parties in Canadian federal politics: Liberals, Conservatives, Bloc Quebecois (BQ), and New Democratic Party (NDP)

  - $Liberal_i$ = 1 if the candidate is affiliated with the Liberal party; 0, otherwise
  - $Conservative_i$ = 1 if the candidate is with the Conservative party; 0, otherwise
  - BQ_i = 1 if the candidate is affiliated with the Bloc Quebecois; 0, otherwise
  - NDP_i = 1 if the candidate is with the New Democratic Party; 0, otherwise

- Other factors

Regression equations

For illustration purposes let's consider a few models. In all the models assume that all relevant *other factors* are included as explanatory variables.

Data: 2006 Canadian federal election

Model 1

$$\text{vote share}_i = \beta_0 + \beta_1 \text{ spending}_i + \beta_2 \text{ incumbency}_i + \beta_3 \text{ male}_i + \beta_4 \text{ Liberal}_i + \beta_5 \text{ Conservative}_i + \beta_6 \text{ BQ}_i + \beta_7 \text{ NDP}_i + \varepsilon_i$$

What happens when we try to estimate this model?

Suppose we estimate:

$$\widehat{\text{vote share}}_i = \mathbf{27.62} + \mathbf{0.19}\ \text{spending}_i + \mathbf{17.40}\ \text{incumbency}_i + 0.12\ \text{male}_i + \mathbf{5.57}\ \text{Conservative}_i + \mathbf{8.72}\ \text{BQ}_i + \mathbf{1.83}\ \text{NDP}_i$$

Model 2

Suppose you believe that the marginal effect of campaign spending on vote share is different for male and female candidates. How do you account for this?

$$\text{vote share}_i = \beta_0 + \beta_1 \text{ spending}_i + \beta_2 \text{ incumbency}_i + \beta_3 \text{ male}_i + \beta_4 (\text{spending}_i \times \text{male}_i) + \beta_5 \text{ Conservative}_i + \beta_6 \text{ BQ}_i + \beta_7 \text{ NDP}_i + \varepsilon_i$$

Estimation yields:

$$\widehat{\text{vote share}}_i = \mathbf{25.08} + \mathbf{0.26} \text{ spending}_i + \mathbf{14.35} \text{ incumbency}_i + \mathbf{2.48} \text{ male}_i - \mathbf{0.09} (\text{spending}_i \times \text{male}_i) + \mathbf{5.81} \text{ Conservative}_i + \mathbf{8.83} \text{ BQ}_i + \mathbf{1.79} \text{ NDP}_i$$