FYI Handout #2: Estimating and Interpreting Linear Regressions

1. Estimation by the method of ordinary least squares (OLS)

Regression analysis starts with the specification of an econometric model. The first fundamental decision the applied researcher must make is to select the dependent variable and the explanatory variables. The resulting model must be appropriate to tackle a specific research question. The decision on the dependent and explanatory variables can be based on an economic theory or on previous empirical investigations by other researchers. More often than not, the availability of data plays an important role in the selection of variables.

A general representation of an econometric model is the following:

$$(1) \qquad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i, \qquad i = 1,2,...,n.$$

In this equation, y_i is the dependent variable and $x_{1i}, x_{2i}, ..., x_{ki}$ are k explanatory variables. When $k = 1$ this is a simple regression; when $k > 1$ it is a multiple regression. Notice that both the dependent and the explanatory variables vary with i, which represents the unit of analysis. The analysis will be carried out using a sample of size n. (In general the i index is used for cross sectional data, when different units of analysis are observed in a given time period, while the index t is used for time series data; refer to the FYI Handout #1 on this.)

The parameters of the model are denoted by the Greek letters $\beta_0, \beta_1, \beta_2, ..., \beta_k$. Finding numerical estimates of the parameters is the main objective of the analysis. In order to do this, we'll need data on the dependent and explanatory variables. An example of dataset looks like this table, where y is hourly wage, x_1 years of education, and x_2 years of experience:

i	y	x_1	x_2
1	6.00	7	15
2	10.00	12	20
…	…	…	…
50	11.25	12	30

Notice that the error term ε_i is not included in the dataset. This is because the error term is *unobservable* and represents, as mentioned in the FYI Handout #1, factors that affect the dependent variable that are not considered in the econometric model. The inclusion of the error term in Equation (1) is necessary to make the right-hand side equal to the left-hand side.

To recapitulate, we now have an econometric model that represents the theory and a suitable dataset that will allow us to apply the theory to a particular place and time. The

output of the estimation is, as we mentioned in the FYI Handout #1, an estimated econometric model. Let's represent the estimated model by the following equation:

$$(2) \qquad \hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki}, \qquad i = 1,2,...,n .$$

In this equation $b_0, b_1, b_2, ..., b_k$ represent numerical estimates of the parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$, and $\hat{y}_i$ represents the estimated value of the dependent variable. Be aware of the differences between Equations (1) and (2). While the first one represents a theoretical model, in which the values of the parameters are still to be determined, the second one represents a concrete relationship between the variables in a given place and time.

The question is how we obtain the values of $b_0, b_1, b_2, ..., b_k$ from the dataset? In this course we focus mainly on the method of ordinary least squares or OLS, which is presented below.

First notice that the estimated value of the dependent variable ($\hat{y}_i$) will in general differ from the observed value of the dependent variable (y_i). Let's call the difference between these values the regression *residual*: $e_i = y_i - \hat{y}_i$. Using equation (2), the residual is

$$(3) \qquad e_i = y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki}), \qquad i = 1,2,...,n .$$

Intuitively, the smaller are the residuals, the closer are the estimated values of the dependent variable from its observed values. The problem is to find numerical values for $b_0, b_1, b_2, ..., b_k$ such that the residuals are collectively "small". Different regression methods have different ways to measure how large or small are the residuals. The method of ordinary least squares measures this magnitude by the sum of squared residuals

$$(4) \qquad Q(b_0, b_1, ..., b_k) = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki}) \right]^2 .$$

Therefore, the objective is to find values of $b_0, b_1, b_2, ..., b_k$ such that the function Q is as small as possible. Mathematically, the problem is to minimize Q. To find the solution, the first order conditions are obtained by setting the partial derivatives of Q with respect to the b's equal to zero.[1]

$$(5) \qquad \frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki}) \right] = 0$$

[1] The second order conditions for a minimum are satisfied because this quadratic objective function is strictly convex.

$$\frac{\partial Q}{\partial b_1} = -2x_1 \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki}) \right] = 0$$

.....

$$\frac{\partial Q}{\partial b_k} = -2x_k \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki}) \right] = 0$$

These conditions form a system of $k + 1$ equations (called normal equations) on $k + 1$ unknowns ($b_0, b_1, b_2, ..., b_k$). The solution of the system gives the estimated values of the parameters.

2. Interpreting regression models

Because the parameters of econometric models play such an important role in applied research, we need to make sure we understand what is their meaning. Let's take for that purpose one step back and look at the stylized economic model that the econometric model in Equation (1) implements:

(6) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$,

In this function β_0 is the intercept and $\beta_1, \beta_2, ..., \beta_k$ are the slopes with respect to each of the explanatory variables. The intercept is the value the dependent variable takes when all the explanatory variables are zero. This parameter is not, in general, the most important, as it is uncommon in an economic model that all the explanatory variables take a value of zero. The slopes, on the other hand, are very important for they represent the relationships between the dependent variable with each of the explanatory variables. Assessing the sign and magnitude of (at least some of) these relationships lies at the heart of empirical economic analysis.

In order to interpret the meaning of the slope coefficients in regression models, some basic calculus is useful. Remember that for a continuous function $y = f(x_1, x_2, ..., x_k)$, the total differential is $dy = \frac{\partial y}{\partial x_1} dx_1 + \frac{\partial y}{\partial x_1} dx_1 + ... + \frac{\partial y}{\partial x_k} dx_k$. In the particular case of Equation (6), the total differential is

(7) $dy = \beta_1 dx_1 + \beta_2 dx_1 + ... + \beta_k dx_k$.

Therefore, the slope coefficient β_1 (for example) represents the rate of change of the dependent variable *dy* as the explanatory variable increases by a small amount dx_1, *with all the other explanatory variables held constant*.

To make these concepts clearer suppose that *W* is hourly wage, *ED* years of education, and *EX* years of experience. Then the parameters of the wage determination model

(8) $W = \beta_0 + \beta_1 ED + \beta_2 EX$

have the following meaning: β_0 represents the wage of an individual with neither education nor wage experience, β_1 is the increase in the wage due to an extra year of education (holding years of experience constant), and β_2 is the increase in the wage due to an extra year of experience (holding years of education constant). Notice that *the units of measurement of β_1 and β_2 depend on the units of measurement of the dependent and the explanatory variables*. For example, if the wage is measured in dollars β_1 is measured in dollars per year of education and β_2 in dollars per year of experience.

- Functional forms

If regression analysis were limited to linear economic models it would not be very useful, for many economic models are non-linear. Fortunately, regression analysis can be applied to a large number of non-linear economic models *provided that they are linear in the parameters*. An expression such as $\log y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 (1/x_2)$, for example, is linear in the parameters though not in the explanatory variables. By letting $w \equiv \log y$, $z_2 \equiv x_1^2$, and $z_3 \equiv 1/x_2$, we rewrite the model as $w = \beta_0 + \beta_1 x_1 + \beta_2 z_2 + \beta_3 z_3$. In other words, provided that the parameters enter linearly in the model, it is always possible to transform the dependent and explanatory variable so that the transformed model has the standard linear form. The way the variables enter in a regression model (in logs, polynomials, reciprocals, etc.) defines the model's *functional form.*

A common functional form in economics is the so-called "log-lin" form, in which the dependent variable is expressed in (natural) logs while the explanatory variables enter linearly. For example in the model of wage determination, the wage is often expressed in logs:

(9) $\log W = \beta_0 + \beta_1 ED + \beta_2 EX$

In this case, the total differential is $\frac{dW}{W} = \beta_1 dED + \beta_2 EX$, leading to a different interpretation of the parameters. For example, β_1 now represents *the relative increase* in the wage due to an extra year of education (holding years of experience constant). For example, an estimated parameter of 0.08 tells us that the wage increases *approximately* 8% with an extra year of education. Notice that for non-linear functions, the differential gives only an approximate value of the change in the dependent variable. The accuracy of this approximation depends on the curvature of the function, given by the parameter values. For example, if the parameter β_1 is a large number, it is better to use finite differences to interpret its meaning. If we let W^1 represent the wage with an extra year

of education, we can write the model as $\log W^1 = \beta_0 + \beta_1(ED+1) + \beta_2 EX$. Therefore, $\beta_1 = \log W^1 - \log W = \log \frac{W^1}{W}$, implying that

(9) $$\frac{W^1 - W}{W} = e^{\beta_1} - 1 .$$

For β_1=0.08, this expression equals 0.0833, making β_1 a good approximation of the relative change in W with an extra year of education, but for much larger values of β_1 the approximation becomes less and less accurate.

Other common function form is the "log-log" form. This form arises, for example from the Cobb-Douglas production function $Y = AL^{\alpha}K^{\beta}$. By taking natural logs in both sides the model becomes

(10) $$\log Y = \log A + \alpha \log L + \beta \log K ,$$

whose total differential is $\frac{dY}{Y} = \alpha \frac{dL}{L} + \beta \frac{dK}{K}$. In this case the parameters are interpreted as point elasticities. For example, $\alpha = \frac{dY/Y}{dL/L}$ represents the point elasticity of output with respect to labor.

Other way to introduce curvature in econometric models is expressing variables as polynomials. For example, the wage determination model mentioned above often includes experience squared:

(11) $$\log W = \beta_0 + \beta_1 ED + \beta_2 EX + \beta_3 EX^2 .$$

In this case the total differential is $\frac{dW}{W} = \beta_1 dED + (\beta_2 + 2EX\beta_3)dEX$, implying that the slope of the relative wage with respect to experience changes as experience increases according to the parameter β_3. In empirical work this parameter is usually negative (while β_2 is positive); therefore, the relative increase in the wage with years of experience is larger for younger than for older workers. Besides computing differentials, it is often very useful to sketch the form of the relationship between the dependent and specific explanatory variables to get a clearer grasp of what the parameters mean.

- Dummy variables

Regression models often incorporate qualitative information coded as "dummy variables". These variables take a value of one if the observation has some qualitative characteristic or zero otherwise. For example in the model of wage determination, researchers often want to explore whether wages are different for different subsets of the

population. For that purpose, they add dummy variables, for example for females, blacks, married individuals, etc. For illustration, let's add a dummy for females d_F (=1 for females and 0 for males) as an additional explanatory variable in the model of wage determination:

(12) $\log W = \beta_0 + \beta_1 ED + \beta_2 EX + \beta_3 d_F$

To interpret the meaning of the parameter β_3 consider what the model looks like for females and males:

Females: $\log W^F = \beta_0 + \beta_1 ED + \beta_2 EX + \beta_3$

Males: $\log W^M = \beta_0 + \beta_1 ED + \beta_2 EX$

Therefore, β_3 represents approximately the relative difference in the average wage of female workers and male workers. As mentioned above, that approximation only works for small values of β_3. The general expression that works for small and large values of that parameter is $\frac{W^F - W^M}{W^M} = e^{\beta_3} - 1$.

In the above model the dummy variable introduces a different intercept for males and females (plot it). Dummy variables can also introduce differences in slopes for different categories of individuals. For example in the model

(13) $\log W = \beta_0 + \beta_1 ED + \beta_2 EX + \beta_3 d_F + \beta_4 EX * d_F$

the slope of the wage with respect to years of experience is different for males and females (plot it). In general dummy variables are a powerful tool in regression analysis, allowing the inclusion of different categorical explanatory variables. They can even be used as dependent variables, as we will see later in this course. The only thing you will need to avoid when it comes to estimating models with dummy explanatory variables is to avoid the so-called *dummy variable trap*. For this purpose *the rule is to add as many dummy variables as the number of categories minus one*. For example, for a gender dummy variable add just one variable, male or female, but not both. As another example, if you have race data classified as white, black, and other, just use black and white. The problem is that having as many dummy variables as categories *plus an intercept* in the regression makes the OLS method fail because the linear system the computer needs to solve has more unknowns than equations (prompting the computer to give you an often incomprehensible error message).

3. Omitted variable bias

A subtle problem of interpretation arises when the econometric model fails to include an important explanatory variable that affects the dependent variable. In that case, the

estimates of the parameter of the model may be *biased*. Because this bias arises from the omission of (one or more) explanatory variables, it is known as omitted variable bias.

To make the presentation simple, suppose that there are three variables: the dependent variable *y*, the explanatory variable included in the regression *x*, and the omitted variable *z*. Because the *z* belongs to the regression, the true model is

(14) $y = \beta_0 + \beta_1 x + \beta_2 z$.

However, by mistake or ignorance, we specify the model as

(15) $y = \alpha_0 + \alpha_1 x$.

Let's assume that *z* is linearly related to *x* through the following model

(16) $z = \delta_0 + \delta_1 x$.

Suppose that what we want to learn through our regression analysis is the individual impact of *x* on *y* (so β_1 is our *parameter of interest*). However, if we omit to include *z* in the regression and estimate Equation (15) instead of Equation (14), we will be estimating the parameter α_1 instead. What does this parameter mean? To answer this question, notice that there are two ways to compute the derivative of *y* with respect to *x* in the equations above. One is to compute the simple derivative in Equation (15), $\frac{dy}{dx} = \alpha_1$, and the other is to compute the total derivative using equations (14) and (16), which gives $\frac{dy}{dx} = \beta_1 + \beta_2 \frac{dz}{dx} = \beta_1 + \beta_2 \delta_1$. Mathematically these two expressions for $\frac{dy}{dx}$ are identical. Therefore,

(17) $\alpha_1 = \beta_1 + \beta_2 \delta_1$.

The omitted variable bias is defined as the difference between the parameter when a variable is omitted and the parameter for the true model: $bias = \alpha_1 - \beta_1$. From Equation (17) we find that $bias = \beta_2 \delta_1$. Therefore, an omitted variable bias arises when (1) the omitted variable is associated with the dependent variable ($\beta_2 \neq 0$) and (2) the omitted variable is associated with the included variable ($\delta_1 \neq 0$).

The omitted variable bias can be a serious problem, as it can severely distort the results of the analysis. Therefore, the inclusion of carefully thought out *control variables* is a crucial component of regression analysis. These variables aim at accounting for factors which might not be of central interest but whose omission may lead to serious deficiencies in the analysis.