

## FYI Handout #1: Introduction to Regression Analysis

### 1. Economic models

Economic models describe relationships among economic variables with one or more mathematical equations. These equations can be of different type: behavioral functions, technological function, accounting identities, and equilibrium conditions. Take for example the simple Keynesian model of income determination:

$$(1) \quad C = a + bY, \quad 0 < b < 1$$

$$(2) \quad AD = C + I$$

$$(3) \quad Y = AD$$

The first equation is a behavioral equation. It tells us that aggregate consumption ( $C$ ) increases with national income ( $Y$ ). The second equation is an identity that defines aggregate demand ( $AD$ ) as the sum of aggregate consumption and aggregate investment ( $I$ ). The final equation is an equilibrium condition. It indicates that the economy is in equilibrium when national income equals aggregate demand.

The Keynesian model is an example of a linear economic model. As an example of a nonlinear model consider the Cobb-Douglas production function:

$$(4) \quad Q = AL^\alpha K^\beta, \quad \alpha > 0, \quad \beta > 0$$

This is an example of a technological function. It tells us which is the maximum level of output ( $Q$ ) possible with given combinations of labor ( $L$ ) and capital ( $K$ ).

In the two examples of functions shown above (Equations 1 and 4), it is important to distinguish (a) the dependent variable, (b) the independent variable(s), and (c) the parameters. The distinction between dependent and independent variables reflects the direction of causality postulated by the economic theory. For example in the consumption function consumption is the dependent variable and income is the independent variable because changes in income cause changes in consumption. In the production function example output is the dependent variable while labor and capital are the independent variables. The theory postulates that changes in labor and capital cause changes in output, and not the other way around.

The parameters play a very important role in an economic theory: they state the sign and shape of relationships among variables. For example in the consumption function, the parameter  $b$  (which Keynes called the marginal propensity to consume) tells us that increases in income lead to less than proportional increases in consumption. This can be seen more clearly from computing the derivative of  $C$  with respect to  $Y$ :

$$(5) \quad \frac{dC}{dY} = b$$

By the same token, the fact that the parameters  $\alpha$  and  $\beta$  are positive in the Cobb-Douglas production function means that output increases when labor or capital increase, as it can be seen from the partial derivatives of  $Q$  with respect to  $L$  and  $K$ :

$$(6) \quad \frac{\partial Q}{\partial L} = \alpha AL^{\alpha-1} K^{\beta} > 0$$

$$(7) \quad \frac{\partial Q}{\partial K} = \beta AL^{\alpha} K^{\beta-1} > 0$$

Theories can be more or less specific about the values of the parameters. At a minimum, they state their sign (whether they are positive or negative). Theories may also care about the relationship between parameters. For example, a Cobb-Douglas production function is characterized by constant returns to scale when  $\alpha + \beta = 1$ . (Incidentally, the sum of  $\alpha + \beta$  is referred to as the returns to scale parameter.) To see this let's rewrite Equation (4) incorporating this theoretical restriction:

$$(8) \quad Q = AL^{\alpha} K^{1-\alpha}, \quad 0 < \alpha < 1$$

Constant returns to scale means that if all the factors of production increase by the same factor, output also increases by that factor. To verify that this is the case, increase labor and capital by a factor of  $d > 0$  and notice that output also increases by  $d$ :

$$A[(1+d)L]^{\alpha} [(1+d)K]^{1-\alpha} = AL^{\alpha} K^{1-\alpha} (1+d)^{\alpha} (1+d)^{1-\alpha} = AL^{\alpha} K^{1-\alpha} (1+d) = (1+d)Q.$$

In sum, economic theories are expressed as models that include behavioral and technological functions. Functions are characterized both by the choice of dependent and independent variables and by their parameters. Theories can be more or less specific according to the restrictions they impose on the parameters.

## 2. Econometric models

Economic models are general in nature. They provide a simplified way of looking at complex realities. Among their many uses, they provide guidance for the empirical researcher who wants to study a specific economic phenomenon in a given place and time.

There are two main differences between an economic and an econometric model. The first one is that the econometric model is designed with the purpose of applying it to data from a specific place and time. For example, a study on aggregate consumption may refer to a given country such as Australia during a specific time, for example 1970 to 1995. A study on the production function may refer to a group of 20 US telecommunications firms during 1992. To indicate the relationship of the econometric model with specific data, subscripts are added to the economic variables.

The second difference is that econometric models include an *error term* to account for unobservable factors not included in the general economic model. The error term, which is usually denoted by a Greek letter such as epsilon ( $\varepsilon$ ), is added to the right-hand side of the behavioral or technological function under study.

As an example of an econometric model, consider the aggregate consumption function:

$$(9) \quad C_t = a + bY_t + \varepsilon_t, \quad t = 1, 2, \dots, T$$

The subscript  $t$  indicates that the data used in the study is time-series, meaning that the same unit of analysis is observed several times over time. If for example, this model is estimated for Australia with data from the period 1970-95, the unit of analysis is Australia and  $t$  goes from 1970 to 1995. In contrast to aggregate consumption ( $C_t$ ) and national income ( $Y_t$ ), the error term is *unobservable*, meaning that we cannot collect data on it. Its role in the econometric model is to indicate that there are further right-hand side variables explaining consumption that are not included in the model.

As another example of econometric model, consider the Cobb-Douglas production function of Equation (4). A minor inconvenience of this function is that it is nonlinear, in contrast to the Keynesian consumption function. Although there is a well-developed theory for the estimation of nonlinear econometric models, in this course we are going to concentrate on linear econometric models. When faced with a nonlinear model of the form of Equation (4), a usual trick, which not always can be applied, is to transform it into a linear function. In this case, this can be done easily by taking logs at both sides:

$$(10) \quad \log Q = \log A + \alpha \log L + \beta \log K$$

Using lower case letters to denote logs (i.e.  $q \equiv \log Q$ , etc.) the econometric model is

$$(11) \quad q_i = a + \alpha l_i + \beta k_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

In this case, the subscript  $i$  indicates that the data used in the study is cross-sectional, meaning that different units are observed during the same period. If this model is applied to data from 20 US telecommunications firms in 1992 the units of analysis are the 20 firms, which are observed in a single time period. A third type of data usually employed in applied economic studies (besides time-series and cross-sectional) is panel data. Panel data combines time-series and cross-sections in that different units are observed during more than one period. The econometric model for the Cobb-Douglas production function with panel data is

$$(12) \quad q_{it} = a + \alpha l_{it} + \beta k_{it} + \varepsilon_{it}, \quad i = 1, 2, \dots, n, \quad t = 1, 2, \dots, T$$

In this case the variables have both subscripts. For example, if we had information about the 20 telecommunications firms for the years 1990 to 1995,  $i$  would range between 1 and

20 and  $t$  would range between 1990 and 1995. Notice that in this example the dataset includes 120 (20 firms times 6 years) observations.

Two final comments: (1) In econometric models the right-hand side variables (which we referred to as independent variables in the context of economic models) are usually called *explanatory variables*. Also, some authors call the dependent variable response variable. (2) The complete specification of an econometric model requires details about the *distribution* of the error term. We'll attack this topic later on in the course.

### 3. Regression Analysis

Regression analysis, a main statistical tool in applied economics, deals with the specification, estimation, and statistical inference of econometric models. To specify an econometric model means to postulate the relationship to be studied. In general econometric models are derived (or at least inspired) from economic models, as we have seen in the examples of the previous section.

The estimation of econometric models is one of our main topics this semester. In particular, we will concentrate on a technique called *ordinary least squares*, whose details will be developed in a few weeks. Similarly, we will devote an important part of the course to introduce statistical inference and its applications to econometric models.

The end result of regression analysis is an *estimated model*. This is an econometric model where the parameters adopt specific numerical values. For example, for the case of the Keynesian consumption function of Australia, the estimated model is

$$(13) \quad \hat{C}_t = -3985 + 0.6219Y_t, \quad t = 1970, \dots, 1995$$

Notice first that the error term is no longer included in the estimated model. Second notice that we added a hat on top of the dependent variable. The hat means that this is an *estimate* of the dependent variable, which will differ from the actual value of the dependent variable ( $C_t$ ). Finally, notice that the parameters of the econometric model have been replaced by their *estimated values* (-3985 and 0.6219).

### 4. Conclusion

Be aware of the differences between the three types of models studied above:

MODEL	EXAMPLE
Economic model	$C = a + bY, \quad 0 < b < 1$
Econometric model	$C_t = a + bY_t + \varepsilon_t, \quad t = 1, 2, \dots, T$
Estimated econometric model	$\hat{C}_t = -3985 + 0.6219Y_t, \quad t = 1970, \dots, 1995$