Handout 3 - Regression Analysis

Regression analysis is a technique used to analyze data consisting of a *dependent variable* (or response variable) and one or more *independent variables* (or explanatory variables). The dependent variable is modeled as a function of independent variables, fixed coefficients known as *parameters* and an *error term.* The error term represents unexplained variation in the dependent variable and is treated as a random variable. The parameter values are estimated to provide the "best fit" to the data. The most commonly used method to estimate parameter values is **least squares**, where parameter values are chosen to minimize the squared difference between the true and fitted values summed over all observations.

1 Univariate Analysis

The simplest form of regression analysis is a **univariate regression** or a model with one independent variable. Assuming a linear relationship between the independent and dependent variables, the general equation can be written as:

$$W_i = \alpha + \beta X_i + \epsilon_i$$

In this equation, W_i is the dependent variable value for person i, X_i is the independent variable value for person i, α and β are parameter values, and ϵ_i is the random error term. The parameter α is called the **intercept** or the value of W when $X = 0$. The parameter β is called the **slope** or the change in W when X increases by one. Since there is only one independent variable, regression analysis estimates the parameters that provide the "best fitting" line when the dependent variable is graphed on the vertical axis and independent variable is on the horizontal axis.

Suppose we have the following data on wages and education levels.

Individual	**Wage**	**Education**
A	8	11
B	9	11
C	8.5	12
D	9.5	12
E	9	13
F	10	13
G	9.5	14
H	10.5	14
I	10	15
J	11	15
K	10.5	16
L	11.5	16

To estimate the relationship between wages and education we can use the data above to estimate the following equation:

$$W_i = \alpha + \beta Ed_i + \epsilon_i$$

where W_i and Ed_i are the wage and education for person i, respectively.

Regression analysis finds estimates $\hat{\alpha}$ and $\hat{\beta}$ for α and β that minimize the squared difference between the true and fitted value for all individuals.[1] Therefore, we want to minimize

$$\Sigma_i(W_i - \hat{W}_i)^2 \equiv \Sigma_i(W_i - \hat{\alpha} - \hat{\beta}Ed_i)^2$$

where $\hat{W}_i$ is the fitted wage value for person i.

We can graph the data with wage on the vertical axis and education on the horizontal axis. The parameters, $\hat{\alpha}$ and $\hat{\beta}$, are the values that provide the "best fitting" line given the data.

[1] A hat (ˆ) on top of a variable indicates an estimated value rather than true value

In this case, the best fitting line slopes upward implying a positive relationship between education and wages. Note that the best fitting line does not need to go through all or even any of the data points.

The best fitted line graphed above corresponds to the following equation:

$$\hat{W}_i = 3 + 0.50Ed_i$$
$$(1.26)(0.09)$$

The estimate of the intercept ($\hat{\alpha}$) is 3 implying that a person with no education has an estimated wage of \$3 per hour. The estimate of the slope parameter ($\hat{\beta}$) is 0.50 implying an additional year of education corresponds to an estimated hourly wage increase of \$0.50.

2 Multivariate Analysis

Even though univariate analysis is useful to demonstrate basic regression analysis, economic theory typically suggests there exists multiple factors that influence the variable of interest. For example, other factors that may influence wage include experience, age, location, job type, sex, race, etc. If any of these previously excluded variables vary systematically with education, our previous estimated relationship between education and wage rates will be incorrect. These additional variables need to be included in the model to get correct parameter estimates. Economic theory is used to decided which variables should be included and excluded from the model and also suggests the direction of causation.

Suppose we want to model wages and economic theory suggests we include education, experience, sex and race into the model. Assume that the available data set includes only whites and blacks. Previous economic research suggests that wages will be increasing in education, increasing in experience, higher for males and higher for whites. Given data, we can estimate the following equation,

$$W_i = \alpha + \beta_1 Ed_i + \beta_2 Exp_i + \beta_3 Male_i + \beta_4 White_i \tag{1}$$

where $Male_i$ and $White_i$ are **dummy variables** that take the value of 1 if the person has the characteristic and 0 otherwise.

The parameters $(\alpha, \beta_1, \beta_2, \beta_3, \beta_4)$ can be estimated using **multiple regression analysis**. Multiple regression analysis is analogous to the univariate analysis where the estimated parameter values are those that construct the best straight-line relationship between the dependent variable and the **set** of independent variables. The parameter values provide the effect on the dependent variable from a one-unit change (or state change for dummy variables) in the corresponding independent variable, holding all other independent variables constant. A common phrase used in economics is **ceteris paribus**, which is Latin for "with other things the same." Therefore, $\hat{\beta}_3$ is the estimated effect on hourly wage of being male, holding education, experience and race constant. Other parameters can be interpreted in a similar manner.