

## LECTURE 1

### 1. INTRODUCTION

#### 1.1. What is Econometrics?

Economic theory studies the relationship between economic variables. Econometrics tries to quantify these relationships, by analyzing data describing economic phenomena through the use of statistical techniques.

#### 1.2. What Kind of Data?

Economic data are typically **nonexperimental**, i.e. empirical economists cannot specify or choose the level of a stimulus (e.g. individual's income) and then record the agents' responses (e.g. demand for a good).

There are different types of economic data. The main types are: **cross section data**, when we observe a number of individuals, or firms, or in general units, at a single point in time; **time series data**, when we observe one or more variables over time; and **panel or longitudinal data**, when we observe a number of units, such as individuals, firms, countries, etc., for a number of time periods.

#### 1.3. Steps of the Econometric Analysis

There are typically six steps in an econometric analysis:

- (1) Pose the question(s) of interest.
- (2) Collect relevant data.
- (3) Write down a statistical model.
- (4) Estimate the model.
- (5) Answer the question(s).
- (6) Question the validity of the model.

## 1.4. Posing a Question

Here are some interesting questions:

- (a) How are wages affected by education?
- (b) How is the consumption of a good affected by its price and consumer's income?
- (c) If the price of a stock is \$150 today what will it be in a month?

Posing the question identifies a **dependent variable** or **regressand** or **left hand side (LHS) variable**  $Y$  (wages in (a), consumption in (b), future price in (c)), that we are trying to explain or predict, and a set of **independent variables** or **regressors** or **explanatory variables** or **right hand side (RHS) variables**  $X_1, X_2, \dots, X_K$  (education in (a), price and income in (b), today's stock price in (c)), that we hope will help us explain and/or predict  $Y$ .

## 1.5. Collecting Data

Most of these lectures concern the analysis of cross-sectional data. Suppose we collect a sample of  $Y$  and  $X_1, X_2, \dots, X_K$  of size  $n$ , i.e. we have  $n(K+1)$ -tuples of observations  $\{(Y_i, X_{i1}, X_{i2}, \dots, X_{iK})\}_{i=1}^n$ . There are different ways of sampling in a cross sectional setting. The most common one is **random sampling**, when the observations  $(Y_i, X_{i1}, X_{i2}, \dots, X_{iK})$  are independent and identically distributed (i.i.d.). We may think of this type of sampling as drawing units independently from a *single* population (pool) that is characterized by some joint distribution for the vector of random variables  $(Y, X_1, X_2, \dots, X_K)$ . Equivalently, we may think of it as having  $n$  subpopulations (pools), *all* of which have the same joint distribution for  $(Y, X_1, X_2, \dots, X_K)$ .

## 1.6. Example: Determinants of Workers' Wages

Suppose we are interested in the determinants of workers' wages. We have at our disposal data from the Current Population Survey (CPS) of March 1995. The sample consists of 1289 individuals. These are workers with complete records on relevant variables and aged between 18 and 65. Self-employed individuals or individuals employed by the Armed Forces or individuals working without payment are excluded from the sample. For these individuals we observe:

- Hourly wage (in US\$) (or if salaried, hourly wage computed as the ratio of weekly labor earnings to the usual hours of work per week)
- Education: number of years of school attended by the individual—topcoded at 20
- Experience (potential): defined as Age – Years of Schooling
- Age
- Female: indicator or dummy variable that equals 1 if the individual is female and is 0 otherwise.

- Nonwhite: indicator or dummy variable that equals 1 if the individual is nonwhite (e.g. black, Hispanic, etc.) and is 0 otherwise.

- Union member: indicator or dummy variable that equals 1 if the individual is a union member and is 0 otherwise.

The following table provides summary statistics for our sample:

Table 1

**Summary Statistics**

Variable	Average	Standard		
		Deviation	Minimum	Maximum
Wage	12.4	7.9	0.8	64.1
Education	13.2	2.8	0.0	20.0
Experience	18.8	11.7	0.0	56.0
Age	37.9	11.5	18	65.0
Female	0.5	0.5	0.0	1.0
Non-white	0.2	0.4	0.0	1.0
Union member	0.2	0.4	0.0	1.0

We are interested in answering questions of the form: Is there wage discrimination by sex and/or by ethnicity and/or by union status?

In order to answer this type of question, one may compare average wages between men and women in the sample, average wages between whites and nonwhites, and average wages between union and non-union members in order to answer these questions. The following table contains average wages by each group and performs simple *t*-tests of equality of means by gender, ethnicity and union status.

Table 2

**Average Wage by Group**

Group	Average	Standard		Sample Size	Test Statistic
		Deviation	Difference		
Men	14.1	8.4		648	
Women	10.6	6.9	3.5	641	8.2
White	12.8	8.1		1092	
Non-white	10.0	5.8	2.8	197	5.8
Union	14.2	5.9		205	
Non-Union	12.0	8.2	-2.2	1084	4.6

But simple differences in averages may be misleading. For example, the difference in the average wage between men and women almost surely reflects the fact that union membership is higher among men (19%) than among women (13%), and the fact that there are more whites among

men (86%) than among women (83%). These percentages come from the following composition of the sample

Table 3

**Composition of the Sample**

Gender	Race	Non-union	Union
Men	White	461	98
	Non-white	63	26
Women	White	471	62
	Non-white	89	19

And conversely, the differences in average wages by union status and race surely reflect the difference in the average wage by gender. In addition, the groups differ in education levels attained and years of (potential) work experience. Both higher education and higher experience are generally associated with higher wages. Indeed, as can be seen from the table below, in this sample, women are on average less educated and less experienced than men; whites are more educated and more experienced than non-whites; and union members have about the same education as non union members but are much more experienced.

Table 4

**Average Education and Experience by Group**

Group	Average Education	Average Experience
Men	13.2	19.1
Women	13.1	18.5
White	13.3	19.0
Non-white	12.6	17.7
Union	13.2	23.0
Non-Union	13.1	18.0

With so many attributes varying simultaneously across individuals, we would like to have a way of separating the wage variation associated with one attribute, say gender, from another attribute, such as education. To this end we first postulate a model that describes the relationship among wages and the attributes of workers. A simple model is a function of the additive form:

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \dots + X_{iK}\beta_K + \varepsilon_i$$

where  $Y_i$  represents the wage for individual  $i$  and the  $X_{ik}$ 's represent the other variables. The coefficients  $\beta_1, \dots, \beta_K$  represent the marginal effects of each  $X_{ik}$  on  $Y_i$ . The variable  $\varepsilon_i$  is unobservable and is typically thought of capturing the effects of other variables that affect the dependent variable but are unobservable by the econometrician. In our example such a variable may be individuals'

unobserved ability. In the sequel, we will examine a method for estimating the coefficients  $\beta_1, \dots, \beta_K$ , called Ordinary Least Squares, and the assumptions under which this method is appropriate to use.

## 2. The Classical Linear Regression (CLR) Model

Before we build the statistical model that relates  $Y_i$  and  $X_i$ , it is useful to define some notation. From now on  $X_i$  is a  $1 \times K$  (row) vector,  $\beta$  is a  $K \times 1$  (column) vector,  $X$  is a  $n \times K$  matrix, and  $Y$  and  $\varepsilon$  are  $n \times 1$  (column) vectors.

$$X_i \equiv \begin{bmatrix} X_{i1} & X_{i2} & \dots & X_{iK} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

$$X \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1K} \\ X_{21} & X_{22} & \dots & X_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nK} \end{bmatrix} \quad Y \equiv \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The **Classical Linear Regression** model is the most basic tool for studying the relationship between a dependent variable and a set of independent variables. Suppose we have  $n$  observations and let  $Y_i$  be the  $i$ 'th observation of the dependent variable and  $X_i \equiv (X_{i1}, \dots, X_{iK})$  the  $i$ 'th observation of the  $K$  regressors. Here both  $Y_i$  and  $X_i$  will be considered as stochastic i.e. as random variables, since data in economics do not come from controlled experiments. The CLR model is the following set of restrictions (assumptions) on the joint distribution of the dependent and the independent variables.

**Assumption 1.1 (Linearity)**  $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_K X_{iK} + \varepsilon_i = X_i \beta + \varepsilon_i$

for all  $i = 1, 2, \dots, n$ . In other words, the CLR model assumes that each  $Y_i$  is generated as a *linear* combination of some observed explanatory variables  $(X_{i1}, \dots, X_{iK})$  with coefficient vector  $\beta$  that is common across all  $i$ , and a scalar unobserved **error term** (or **disturbance** or **shock**),  $\varepsilon_i$ , that is commonly thought of as capturing the effect of unobservable variables not included in the set of explanatory variables  $(X_{i1}, \dots, X_{iK})$ .

**Assumption 1.2 (Strict Exogeneity)**  $E(\varepsilon_i | X) = E(\varepsilon_i | X_1, X_2, \dots, X_n) = 0$

for all  $i = 1, 2, \dots, n$ . In other words, the CLR model assumes that the expectation of the  $i$ 'th error term is zero conditional on the regressors for *all* observations (not just the  $i$ 'th observation). By the **Law of Iterated Expectations**,<sup>1</sup> this assumption implies that the

---

<sup>1</sup>The Law of Iterated Expectations states that  $E(E(y|x)) = E(y)$ .

unconditional mean of the error terms is also zero, i.e.  $E(\varepsilon_i) = E(E(\varepsilon_i|X)) = 0$ . Furthermore, it implies that each regressor is **orthogonal**<sup>2</sup> to the error terms for *all* observations. i.e.  $E(X_{ik}\varepsilon_j) = 0$  for all  $i, j = 1, \dots, n$  and for all  $k = 1, \dots, K$ . Note that since the mean of the error term is zero,  $E(X_{ik}\varepsilon_j) = Cov(X_{ik}, \varepsilon_j)$ , that is, errors and regressors are uncorrelated. For time series models where  $i$  is time, the implication of strict exogeneity can be rephrased as the regressors being orthogonal to the past, current, and future error terms. Thus, in a time series setting this assumption rules out lag(s) of the dependent variable as regressors. In other words, time series models with the lagged dependent variable appearing on the RHS violate the strict exogeneity assumption and therefore cannot be considered as CLR models.

**Assumption 1.3 (Spherical Errors)**

- $E(\varepsilon_i^2|X) = \sigma^2 > 0$

for all  $i = 1, 2, \dots, n$ , where  $\sigma^2$  is an unknown positive scalar. This property of the observations is called **conditional homoskedasticity**. Note that under the strict exogeneity assumption  $E(\varepsilon_i^2|X)$  is just  $Var(\varepsilon_i|X)$  since  $Var(\varepsilon_i|X) = E(\varepsilon_i^2|X) - [E(\varepsilon_i|X)]^2$ . Furthermore, by the strict exogeneity assumption,  $Var(\varepsilon_i) = Var(E(\varepsilon_i|X)) + E(Var(\varepsilon_i|X)) = E(Var(\varepsilon_i|X)) = E(E(\varepsilon_i^2|X)) = \sigma^2$ , that is, the error term is also **unconditionally homoskedastic**.

- $E(\varepsilon_i\varepsilon_j|X) = 0$

for all  $i \neq j$  where  $i, j = 1, 2, \dots, n$ . Here we are assuming that the errors are uncorrelated across observations conditional on  $X$ , since by the strict exogeneity assumption  $E(\varepsilon_i\varepsilon_j|X) = Cov(\varepsilon_i, \varepsilon_j|X)$ . Note that  $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i\varepsilon_j) = E[E(\varepsilon_i\varepsilon_j|X)] = 0$  by the Law of Iterated Expectations, which implies that the errors are uncorrelated across observations unconditionally as well.

**Assumption 1.4 (Full Rank)** The  $n \times K$  matrix  $X$  has rank  $K$  with probability 1.

This condition, that  $X$  has rank equal to the number of its columns, means that there does not exist a perfect linear relationship between the columns of  $X$  (no **perfect collinearity**). A necessary (but not sufficient) condition is that the number of rows is at least as large as the number of columns, i.e.  $n \geq K$ , or in other words that we have at least as many observations as the number of parameters we are trying to estimate.

It is important to notice that we can write the CLR model dispensing altogether with the error term  $\varepsilon_i$ . In particular, we can rewrite assumptions 1.1-1.3 as

$$E(Y_i|X) = X_i\beta \tag{1.1'}$$

$$V(Y_i|X) = \sigma^2 \tag{1.2'}$$

$$Cov(Y_i, Y_j|X) = 0 \tag{1.3'}$$

for all  $i$  and for all  $i \neq j$ .

---

<sup>2</sup>If the cross moment  $E(xy)$  of two random variables is zero, then we say that  $x$  is orthogonal to  $y$  (or equivalently that  $y$  is orthogonal to  $x$ ).

In the CLR model  $\beta \equiv (\beta_1, \dots, \beta_K)'$  is the  $(K \times 1)$  vector of unknown parameters of interest. Note that for each  $k = 1, \dots, K$ ,

$$\beta_k = \frac{\partial E(Y_i|X)}{\partial X_{ik}} = E\left(\frac{\partial Y_i}{\partial X_{ik}}|X\right)$$

i.e.  $\beta_k$  measures the **marginal effect of the  $k$ 'th variable on  $E(Y_i|X)$** , i.e. by how much  $E(Y_i|X)$  will change if  $X_{ik}$  changes by one unit. Note that commonly the first element of  $X_i$  will be 1 for all  $i$ , i.e.  $X_{i1} \equiv 1$ , which means that the matrix  $X$  has as a first column a vector ones. This implies that we are allowing for an intercept in the model. But we don't have to. However, if we don't allow for an intercept, we are imposing a restriction on the model, which may be best understood in the two-variable case. In the two-variable model  $X_{i1} \equiv 1$  means that for all  $i$ ,  $E(Y_i|X) = \beta_1 + \beta_2 X_{i2}$ , i.e. the conditional means of the  $Y_i$ 's lie on a straight line with intercept  $\beta_1$  and slope equal to  $\beta_2$ . Imposing the restriction that the intercept is zero (i.e. that  $\beta_1 = 0$  which implies that the model is  $E(Y_i|X) = \beta_2 X_{i2}$ ) means that we are imposing the restriction that the line passes through the origin.

We can write assumption 1.1-1.3 (or 1.1'-1.3') of the CLR model in matrix notation as follows:

$$\begin{aligned} Y &= X\beta + \varepsilon \\ E(\varepsilon|X) &= 0 \\ V(\varepsilon|X) &= V(Y|X) = \sigma^2 I_n \end{aligned}$$

or dispensing with  $\varepsilon$  altogether,

$$\begin{aligned} E(Y|X) &= X\beta \\ V(Y|X) &= \sigma^2 I_n \end{aligned}$$

Here,  $E(Y|X)$  is the  $n \times 1$  vector of conditional means,  $E(Y|X) \equiv \begin{bmatrix} E(Y_1|X) \\ E(Y_2|X) \\ \vdots \\ E(Y_n|X) \end{bmatrix} = \begin{bmatrix} X_1\beta \\ X_2\beta \\ \vdots \\ X_n\beta \end{bmatrix}$  and

$V(Y|X)$  is the  $n \times n$  conditional variance-covariance matrix of  $Y$  given  $X$ ,

$$\begin{aligned} V(Y|X) &\equiv \begin{bmatrix} \text{Var}(Y_1|X) & \text{Cov}(Y_1, Y_2|X) & \dots & \text{Cov}(Y_1, Y_n|X) \\ \text{Cov}(Y_1, Y_2|X) & \text{Var}(Y_2|X) & \dots & \text{Cov}(Y_2, Y_n|X) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_1, Y_n|X) & \text{Cov}(Y_2, Y_n|X) & \dots & \text{Var}(Y_n|X) \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I_n \end{aligned}$$

Like any variance-covariance matrix,  $V(Y|X)$  is a symmetric Positive Definite matrix that has the  $n$  conditional variances on the diagonal and the  $n(n-1)$  conditional covariances off the diagonal.

## 2.1. Nonlinearities in the Population Regression Function

A critical assumption underlying the CLR model is that  $E(Y_i|X) = X_i\beta$ , i.e. that the conditional mean of the dependent variable is a linear function of each of the  $X_{ij}$ 's. This implies that *the average marginal effect of each independent variable (i.e. the derivative of  $Y_i$  with respect to each  $X_{ij}$ ) is constant for all of its own levels as well as for all the levels of the other independent variables*, i.e.  $\frac{\partial E(Y_i|X)}{\partial X_{ij}}$  is constant (and equal to  $\beta_j$ ) and not a function of  $X_{ij}$  or of  $(X_{i1}, \dots, X_{ij-1}, X_{ij+1}, \dots, X_{iK})$ . However, we may think that this linearity assumption may not hold on grounds of either economic theory or by merely examining the scatterplot of the dependent variable and each one of the independent variables.

### Example:

(For simplicity, from now on I will drop the subscript  $i$  that denotes an individual's identity.) Suppose we want to analyze wages  $W$ . Two important variables that should affect level of a person's wage are Schooling ( $S$ ) and Experience ( $EX$ ). We may specify the following statistical model for the determination of wages:

$$E(W|S, EX) = \beta_1 + \beta_2 S + \beta_3 EX \quad (2.1)$$

where we expect that both  $\beta_2$  and  $\beta_3$  are positive, i.e. a person's wage increases on average with both  $S$  and  $EX$ . Note that here  $\frac{\partial E(Y|X)}{\partial EX} = \beta_3$  and  $\frac{\partial E(Y|X)}{\partial S} = \beta_2$  are both constant, i.e. if we increase either schooling or experience by one unit, on average a person's wage will change by  $\beta_3$  or  $\beta_2$  units, respectively, no matter what is the level of  $EX$  or  $S$ .

However, because a person's human capital generally depreciates with age and hence with experience (since as a person grows older his (her) productivity falls), we expect a decreasing effect of experience, that is, we expect  $E(W|S, EX)$  to be a concave function of  $EX$ , in which case we will want to specify the model as:

$$E(W|S, EX) = \beta_1 + \beta_2 S + \beta_3 EX + \beta_4 EX^2 \quad (2.2)$$

where we expect  $\beta_4$  to be negative.

Note that now,

$$\frac{\partial E(W|S, EX)}{\partial EX} = \beta_3 + 2\beta_4 EX$$

i.e. the effect of  $EX$  on  $E(W|S, EX)$  is changing with the level of  $EX$ .

Furthermore, we may think that the more schooling one has, the more on-the-job training one receives, so that a person's productivity increases with schooling. This means that the effect of  $EX$  is greater for higher values of  $S$ . Then we may write the model as:

$$E(W|S, EX) = \beta_1 + \beta_2 S + \beta_3 EX + \beta_4 EX^2 + \beta_5 (S \cdot EX) \quad (2.3)$$

Here we introduced an **interaction term**, that is, the product between  $S$  and  $EX$ . Because of the previous argument (on-the-job training), we expect that  $\beta_5$  is positive. This means that

$$\frac{\partial E(W|S, EX)}{\partial EX} = \beta_3 + 2\beta_4 EX + \beta_5 S$$

i.e. the effect of  $EX$  on  $E(W|S, EX)$  is higher (because  $\beta_5$  is positive) for higher values of  $S$ .

Note that the population regression functions (2.2) and (2.3) are no longer linear in  $S$  and  $EX$ . However, the introduction of  $EX^2$  and  $S \cdot EX$  as new variables in the regression functions does *not* destroy the full rank of the regressor matrix required by Assumption 1.4 of the CLR model. This is because  $EX^2$  and  $S \cdot EX$  are *nonlinear* functions of  $EX$  and  $S$ , so that their introduction into the model does not create perfect collinearity among the columns of the regressor matrix.

Model (2.1) above assumes that, no matter what the schooling level (or experience) of an individual is, an additional year will increase wages on average by the same *amount*. Researchers however have found more evidence of the hypothesis that an additional year schooling (or experience) will increase wages on average by the same *percentage*, i.e. that the returns to schooling (and experience) are constant. This implies that the basic specification should be

$$E(\ln W|S, EX) = \beta_1 + \beta_2 S + \beta_3 EX$$

that is, the dependent variable should be expressed in natural logarithm form. The model in the latter equation is sometimes referred to as the **semilog** model and the coefficients on  $S$  and  $EX$  as **semielasticities** or **rates of return**, or, if the independent variable is time and its coefficient is positive, as **rates of growth**.

Another popular model is the so called **loglinear** model:

$$E(\ln Y|X) = \beta_1 + \beta_2 \ln X$$

Here  $\beta_2$  is the **elasticity** of  $Y$  with respect to  $X$  and is assumed constant at all levels of  $X$ .

Another model of flexible functional form used in studies of demand and production is the **translog model**

$$E(\ln Y|X_1, \dots, X_K) = \beta_1 + \sum_{j=1}^K \beta_j \ln X_j + \frac{1}{2} \sum_{j=1}^K \sum_{l=1}^K \beta_{jl} \ln X_j \ln X_l$$

which may be interpreted as second order approximation to an unknown functional form. This model allows analysts to study second order effects such as elasticities of substitution which are functions of the second derivatives of production, cost and utility functions. These are restricted to be zero in the linear model and to -1 or +1 in the loglinear model.

Summarizing, by introducing quadratic, cubic, quartic, etc. terms, or other nonlinear transformations of the independent variables, in a regression and/or interaction terms between the independent variables we are able to capture nonlinearities in the relationship between the (mean of the) dependent variable and the independent variables. In addition, nonlinearities may be also captured by taking appropriate nonlinear transformations, e.g. the natural logarithm, of the dependent and/or the independent variables.

## 2.2. Non-Constancy of Parameters in the Population Regression Function

The assumption of the CLR model that  $E(Y_i|X) = X_i\beta$ , imposes that the unknown coefficient vector  $\beta$  is the same for all units  $i$  in the sample. At first glance, this assumption may seem too restrictive. For example, there is evidence using various data sets that on average wages are higher for men than for women for the same levels of schooling and experience. To allow for parameters that are not constant for all groups in the sample, we may use indicator or dummy variables. An **indicator or dummy** variable takes the value 1 if the individual belongs to a certain group and is zero otherwise.

### Example:

Suppose that we think that gender may matter in determining wages. Define the dummy variables

$$M_i = \begin{cases} 1 & \text{if individual } i \text{ is male} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad F_i = \begin{cases} 1 & \text{if individual } i \text{ is female} \\ 0 & \text{otherwise} \end{cases}$$

We may then specify the wage determination model as

$$E(W|S, EX, M) = \beta_1 + \beta_2 S + \beta_3 EX + \beta_4 M \tag{2.4}$$

Thus, for a male ( $M = 1$ )

$$E(W|S, EX, M = 1) = (\beta_1 + \beta_4) + \beta_2 S + \beta_3 EX$$

while for a female ( $M = 0$ )

$$E(W|S, EX, M = 0) = \beta_1 + \beta_2 S + \beta_3 EX$$

By writing the model as in (2.4), we are allowing the intercept of the population regression function for men,  $(\beta_1 + \beta_4)$ , to be different from the intercept of the population regression function for women, which is equal to  $\beta_1$ . If  $\beta_4$  is non-zero then on average men and women do not receive the same wages for the same levels of schooling and experience. If it is positive, then on average men receive higher wages on average than women.

Note that we could have specified the model as

$$E(W|S, EX, F) = \gamma_1 + \gamma_2 S + \gamma_3 EX + \gamma_4 F \tag{2.5}$$

Thus, for a male ( $F = 0$ )

$$E(W|S, EX, F = 0) = \gamma_1 + \gamma_2 S + \gamma_3 EX$$

while for a female ( $F = 1$ )

$$E(W|S, EX, M = 0) = (\gamma_1 + \gamma_4) + \gamma_2 S + \gamma_3 EX$$

Now evidence in favor of men would be indicated by  $\gamma_4$  being negative.

We may even specify the model using both dummy variables:

$$E(W|S, EX, F, M) = \delta_1 M + \delta_2 S + \delta_3 EX + \delta_4 F$$

Thus, for a male ( $M = 1, F = 0$ )

$$E(W|S, EX, F = 0, M = 1) = \delta_1 + \delta_2 S + \delta_3 EX$$

while for a female ( $M = 0, F = 1$ )

$$E(W|S, EX, F = 1, M = 0) = \delta_4 + \delta_2 S + \delta_3 EX$$

Note however that in this last specification we do not include a constant in the regression since that would create perfect collinearity among the constant column of 1's and the columns that contain the values for  $M$  and  $F$ . In particular, for any individual  $i$ ,  $M_i + F_i = 1$ .

Naturally the parameters  $(\beta_1, \dots, \beta_4)$ ,  $(\gamma_1, \dots, \gamma_4)$  and  $(\delta_1, \dots, \delta_4)$  of the three models above are related. In particular,  $\beta_2 = \gamma_2 = \delta_2$ ,  $\beta_3 = \gamma_3 = \delta_3$ ,  $\beta_1 = \gamma_1 + \gamma_4 = \delta_4$  and  $\gamma_1 = \beta_1 + \beta_4 = \delta_1$ .

Using dummy variables, we may allow not only for different intercepts across the different groups in the sample, but also for different slope coefficients. This is accomplished by including **interaction terms**, that is, by including as an additional variable(s) in the model the product of the dummy with one (or more) other explanatory variables. For example, we may specify the model to allow for differential effects of schooling by gender as follows,

$$E(W|S, EX, M) = \beta_1 + \beta_2 S + \beta_3 EX + \beta_4 M + \beta_5 (M \cdot S)$$

in which case for a male,

$$E(W|S, EX, M = 1) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) S + \beta_3 EX$$

while for a female ( $M = 0$ )

$$E(W|S, EX, M = 0) = \beta_1 + \beta_2 S + \beta_3 EX$$

Thus if  $\beta_5$  is non-zero, the effect of schooling on wages is different across genders, and is equal to  $(\beta_2 + \beta_5)$  for men while it is equal to  $\beta_2$  for women. As before we might have written the model using only the female dummy  $F$ , or using both the male and the female dummies. In the last case however, one should be careful not to include terms that may create perfect collinearities among the columns of the regressor matrix.

Summarizing, by including indicator or dummy variables and their interactions with other explanatory variables we may allow for parameters that are non-constant across the different groups in the sample.

### 2.3. Implications of Random Sampling

In cross sectional settings it is often the case that we are willing to assume that the sample  $\{(Y_i, X_i)\}_{i=1}^n$  is random, i.e. that the  $(K + 1)$ -tuples  $(Y_i, X_i)$  are independent and identically distributed (i.i.d.) across  $i$ . A question of interest that arises then is *whether the assumption of random sampling is consistent with the CLR model*, or, in other words, *does a random sample satisfy the assumptions of the CLR model*. The answer is yes, provided that the (common) joint distribution of  $(Y_i, X_i)$  satisfies the following assumptions:

$$E(Y_i|X_i) = X_i\beta \quad (1.1'')$$

$$Var(Y_i|X_i) = \sigma^2 \quad (1.2'')$$

$E(Y_i|X_i)$  is called the **conditional expectation function (CEF) or Population Regression Function (PRF) of  $Y_i$  given  $X_i$** . In general, a CEF is a nonlinear function of the conditioning variable(s). Here, however, we are assuming that for each unit  $i$  in the sample, this conditional expectation function is a *linear* combination of the conditioning vector  $X_i$ , with coefficients  $(\beta_1, \beta_2, \dots, \beta_K)$  that are the *same* for all  $i$ . Note that, although random sampling implies that the unconditional mean of  $Y_i$ ,  $E(Y_i)$ , is the same across  $i$ , and here equal to  $E(X_i)\beta$ , the conditional mean of each  $Y_i$  varies with  $i$  because  $X_i$  is potentially different for each  $i$ . Furthermore, random sampling implies that the unconditional variance of  $Y_i$  is the same across  $i$  (unconditional homoskedasticity). However, in general the conditional variance will be a function of  $X_i$  and thus non-constant across  $i$ . However, here we are imposing that the conditional variance is constant across all units  $i$  in the sample and equal to  $\sigma^2$ .

It is easy to verify that assumptions 1.1'' – 1.2'' along with the assumption of random sampling imply assumptions 1.1-1.3 of the CLR model. Indeed, under random sampling,

$$E(Y_i|X) = E(Y_i|X_1, \dots, X_{i-1}, X_i, X_{i+1}, \dots, X_n) = E(Y_i|X_i) = X_i\beta$$

where the last equality comes from (1.1'') above. Thus, assumption 1.1' of the CLR model holds. Similarly, by random sampling,

$$V(Y_i|X) = V(Y_i|X_1, \dots, X_{i-1}, X_i, X_{i+1}, \dots, X_n) = V(Y_i|X_i) = \sigma^2$$

where again the last equality comes from 1.2'' above. Thus assumption 1.2' of the CLR model holds. Finally, random sampling implies that

$$\begin{aligned} Cov(Y_i, Y_j|X) &= Cov(Y_i, Y_j|X_i, X_j) \\ &= E(Y_i Y_j|X_i, X_j) - E(Y_i|X_i) E(Y_j|X_j) \\ &= E(Y_i|X_i) E(Y_j|X_j) - E(Y_i|X_i) E(Y_j|X_j) \\ &= 0 \end{aligned}$$

and hence assumption 1.3' of the CLR model holds.

By defining  $\varepsilon_i \equiv Y_i - E(Y_i|X_i) = Y_i - X_i\beta$ , it is trivial to verify that  $E(\varepsilon_i|X_i) = 0$  and that  $Var(\varepsilon_i|X_i) = E(\varepsilon_i^2|X_i) = \sigma^2$ . Furthermore, the error terms  $\varepsilon_i$  will be independent and identically distributed as functions of the i.i.d. RV's  $(Y_i, X_i)$ . The assumption that  $E(\varepsilon_i|X_i) = 0$  is often referred to as **mean-independence** and it implies that  $\varepsilon_i$  is uncorrelated with *any* function of  $X_i$ .

## 2.4. Fixed Regressors

Most textbooks treat  $X$  as fixed instead of random, in which case assumptions 1.2 and 1.3 are stated as:

**Assumption 1.2' (Zero Mean)**  $E(\varepsilon_i) = 0$

**Assumption 1.3' (Spherical Errors)**

- $E(\varepsilon_i^2) = \sigma^2 > 0$
- $E(\varepsilon_i\varepsilon_j) = 0$

However the assumption of nonstochastic regressors is clearly inappropriate for non-experimental data. It should be also noted here that in this case it *cannot* be assumed that the data come from a random sample. For one thing, Assumption 1.2' above implies that  $E(Y_i) = X_i\beta$ , that is, each  $Y_i$  has a mean that depends on  $X_i$ . Hence the  $Y_i$ 's cannot be identically distributed.