# LINEAR REGRESSION ANALYSIS

## MODULE – VIII

## Lecture - 26

# Indicator Variables

**Dr. Shalabh**
**Department of Mathematics and Statistics**
**Indian Institute of Technology Kanpur**

In general, the explanatory variables in any regression analysis are assumed to be quantitative in nature.

For example, the variables like temperature, distance, age etc. are quantitative in the sense that they are recorded on a well defined scale.

In many applications, the variables can not be defined on a well defined scale and they are qualitative in nature.

For example, the variables like sex (male or female), colour (black, white), nationality, employment status (employed, unemployed) are defined on a nominal scale. Such variables do not have any natural scale of measurement. Such variables usually indicate the presence or absence of a "quality" or an attribute like employed or unemployed, graduate or non-graduate, smokers or non- smokers, yes or no, acceptance or rejection, so they are defined on a nominal scale. Such variables can be quantified by artificially constructing the variables that take the values, e.g., 1 and 0 where "1" indicates usually the presence of attribute and "0" indicates usually the absence of attribute. For example, "1" indicates that the person is male and "0" indicates that the person is female. Similarly, "1" may indicate that the person is employed and then "0" indicates that the person is unemployed.

Such variables classify the data into mutually exclusive categories. These variables are called **indicator variables** or **dummy variables.**

Usually, the indicator variables take on the values 0 and 1 to identify the mutually exclusive classes of the explanatory variables. For example,

$$D = \begin{cases} 1 & \text{if person is male} \\ 0 & \text{if person is female,} \end{cases}$$

$$D = \begin{cases} 1 & \text{if person is employed} \\ 0 & \text{if person is unemployed.} \end{cases}$$

Here we use the notation $D$ in place of $X$ to denote the dummy variable. The choice of 1 and 0 to identify a category is arbitrary. For example, one can also define the dummy variable in above examples as

$$D = \begin{cases} 1 & \text{if person is female} \\ 0 & \text{if person is male,} \end{cases}$$

$$D = \begin{cases} 1 & \text{if person is unemployed} \\ 0 & \text{if person is employed.} \end{cases}$$

It is also not necessary to choose only 1 and 0 to denote the category. In fact, any distinct value of $D$ will serve the purpose. The choices of 1 and 0 are preferred as they make the calculations simple, help in easy interpretation of the values and usually turn out to be a satisfactory choice.

In a given regression model, the qualitative and quantitative variables may also occur together, i.e., some variables may be qualitative and others quantitative.

When all explanatory variables are

- **quantitative**, then the model is called a **regression model,**
- **qualitative**, then the model is called an **analysis of variance model** and
- **quantitative and qualitative both**, then the model is called a **analysis of covariance model**.

Such models can be dealt within the framework of regression analysis. The usual tools of regression analysis can be used in case of dummy variables.

## Example:

Consider the following model with $x_1$ as quantitative and $D_2$ as indicator variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_2 + \varepsilon,\ E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$$

$$D_2 = \begin{cases} 0 & \text{if an observation belongs to group } A \\ 1 & \text{if an observation belongs to group } B. \end{cases}$$

The interpretation of result is important. We proceed as follows:

If $D_2 = 0$ then

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 .0 + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \varepsilon \end{aligned}$$

$$E(y \mid D_2 = 0) = \beta_0 + \beta_1 x_1$$

which is a straight line relationship with intercept $\beta_0$ and slope $\beta_1$.

If $D_2 = 1$ then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 .1 + \varepsilon$$
$$= (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$
$$E(y \mid D_2 = 1) = (\beta_0 + \beta_2) + \beta_1 x_1$$

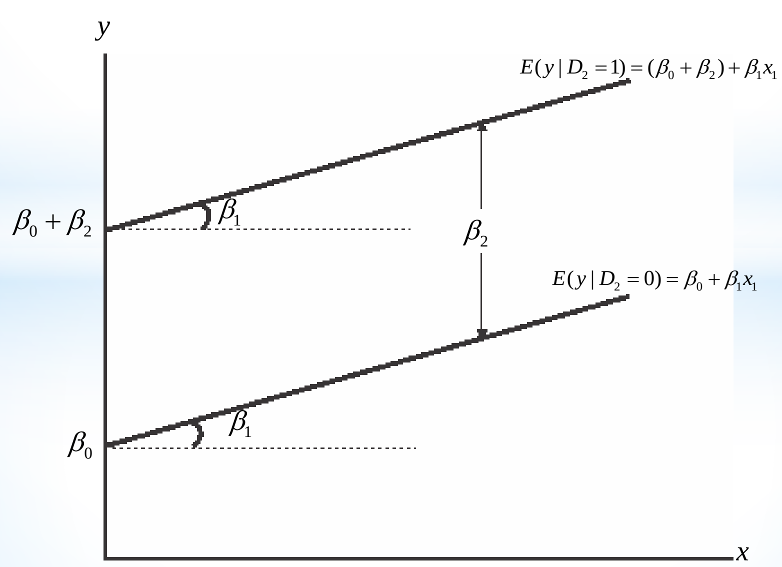
which is a straight line relationship with intercept $(\beta_0 + \beta_2)$ and slope $\beta_1$.

The quantities $E(y|D_2 = 0)$ and $E(y|D_2 = 1)$ are the average responses when an observation belongs to group *A* and group *B*, respectively. Thus

$$\beta_2 = E(y \mid D_2 = 1) - E(y \mid D_2 = 0)$$

which has an interpretation as the difference between the average values of $y$ with $D_2 = 0$ and $D_2 = 1$.

Graphically, it looks like as in the following figure. It describes two parallel regression lines with same variances .



If there are three explanatory variables in the model with two indicator variables $D_2$ and $D_3$ then they will describe three levels, e.g., groups *A, B* and The levels of indicator variables are as follows:

1. $D_2 = 0,\ D_3 = 0$ if the observation is from group $A$.

2. $D_2 = 1,\ D_3 = 0$ if the observation is from group $B$.

3. $D_2 = 0,\ D_3 = 1$ if the observation is from group $C$.

The concerned regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon,\ E(\varepsilon) = 0,\ Var(\varepsilon) = \sigma^2.$$

In general, if a qualitative variable has *m* levels, then (*m* - 1) indicator variables are required and each of them takes value 0 and 1.

Consider the following examples to understand how to define such indicator variables and how they can be handled.

**Example:**

Suppose $y$ denotes the monthly salary of a person and $D$ denotes whether the person is graduate or non-graduate. The model is

$$y = \beta_0 + \beta_1 D + \varepsilon,\ E(\varepsilon) = 0,\ \text{var}(\varepsilon) = \sigma^2.$$

With $n$ observations, the model is

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i,\ i = 1, 2, ..., n$$

$$E(y_i \mid D_i = 0) = \beta_0$$

$$E(y_i \mid D_i = 1) = \beta_0 + \beta_1$$

$$\beta_1 = E(y_i \mid D_i = 1) - E(y_i \mid D_i = 0).$$

Thus

- $\beta_0$ measures the mean salary of a non-graduate
- $\beta_1$ measures the difference in the mean salaries of a graduate and non-graduate person.

Now consider the same model with two indicator variables defined in the following way:

$$D_{i1} = \begin{cases} 1 & \text{if person is graduate} \\ 0 & \text{if person is nongraduate,} \end{cases}$$

$$D_{i2} = \begin{cases} 1 & \text{if person is nongraduate} \\ 0 & \text{if person is graduate.} \end{cases}$$

The model with *n* observations is

$$y_i = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \varepsilon_i,\ E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2, i = 1, 2, ..., n.$$

Then we have

1. $E[y_i \mid D_{i1} = 0, D_{i2} = 1] = \beta_0 + \beta_2 :$ Average salary of non-graduate
2. $E[y_i \mid D_{i1} = 1, D_{i2} = 0] = \beta_0 + \beta_1 :$ Average salary of graduate
3. $E[y_i \mid D_{i1} = 0, D_{i2} = 0] = \beta_0$ **:** cannot exist
4. $E[y_i \mid D_{i1} = 1, D_{i2} = 1] = \beta_0 + \beta_1 + \beta_2 :$ cannot exist.

Notice that in this case

$D_{i1} + D_{i2} = 1$ for all *i*

which is an exact constraint and indicates the contradiction as follows:

$D_{i1} + D_{i2} = 1 \Rightarrow$ person is graduate

$D_{i1} + D_{i2} = 1 \Rightarrow$ person is non-graduate.

So multicollinearity is present in such cases. Hence the rank of matrix of explanatory variables falls short by 1. So $\beta_0, \beta_1$ and $\beta_2$ are indeterminate and least squares method breaks down. So the proposition of introducing two indicator variables is useful but they lead to serious consequences. This is known as **dummy variable trap.**

If the intercept term is ignored, then the model becomes

$$y_i = \beta_1 D_{i1} + \beta_2 D_{i2} + \varepsilon_i,\ E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2,\ i = 1, 2, ..., n$$

then

$$E(y_i \mid D_{i1} = 1, D_{i2} = 0) = \beta_1 \Rightarrow \text{Average salary of a graduate.}$$

$$E(y_i \mid D_{i1} = 0, D_{i2} = 1) = \beta_2 \Rightarrow \text{Average salary of a non-graduate.}$$

So when intercept term is dropped, then $\beta_1$ and $\beta_2$ have proper interpretations as the average salaries of a graduate and non-graduate persons, respectively.

Now the parameters can be estimated using ordinary least squares principle and standard procedures for drawing inferences can be used.

**Rule:** When the explanatory variable leads to *m* mutually exclusive categories classification, then use (*m* - 1) indicator variables for its representation. Alternatively, use indicator variables but drop the intercept term.

## Interaction term

Suppose a model has two explanatory variables – one quantitative variable and other an indicator variable. Suppose both interact and an explanatory variable as the interaction of them is added to the model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 D_{i2} + \beta_3 x_{i1} D_{i2} + \varepsilon_i,\ E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2, i = 1, 2, ..., n.$$

To interpret the model parameters, we proceed as follows:

Suppose the indicator variables are given by

$$D_{i2} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ person belongs to group } A \\ 0 & \text{if } i^{\text{th}} \text{ person belongs to group } B \end{cases}$$

$y_i =$ salary of $i^{th}$ person.

Then

$$\begin{aligned} E\left(y_i \mid D_{i2} = 0\right) &= \beta_0 + \beta_1 x_{i1} + \beta_2 .0 + \beta_3 x_{i1} .0 \\ &= \beta_0 + \beta_1 x_{i1}. \end{aligned}$$

This is a straight line with intercept $\beta_1$ and slope $\beta_2$ . Next

$$\begin{aligned} E\left(y_i \mid D_{i2} = 1\right) &= \beta_0 + \beta_1 x_{i1} + \beta_2 .1 + \beta_3 x_{i1} .1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1}. \end{aligned}$$

This is a straight line with intercept term $(\beta_0 + \beta_2)$ and slope $(\beta_1 + \beta_3)$.

The model

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 D_{i2} + \beta_3 x_{i1} D_{i2}$$

has different slopes and different intercept terms.

Thus

$\beta_2$ reflects the change in intercept term associated with the change in the group of person i.e., when group changes from *A* to *B*.

$\beta_3$ reflects the change in slope associated with the change in the group of person, i.e., when group changes from *A* to *B*.

Fitting of the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 D_{i2} + \beta_3 x_{i1} D_{i2} + \varepsilon_i$

is equivalent to fitting two separate regression models corresponding to $D_{i2}$=1 and $D_{i2}$=0, i.e.,

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 .1 + \beta_3 x_{i1} .1 + \varepsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1} D_{i2} + \varepsilon_i \end{aligned}$$

and

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 .0 + \beta_3 x_{i1} .0 + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \varepsilon_i , \end{aligned}$$

respectively.

The test of hypothesis becomes convenient by using an indicator variable. For example, if we want to test whether the two regression models are identical, the test of hypothesis involves testing

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0.$$

Acceptance of $H_0$ indicates that only single model is necessary to explain the relationship.

In another example, if the objective is to test that the two models differ with respect to intercepts only and they have same slopes, then the test of hypothesis involves testing $H_0 : \beta_3 = 0$

$$H_1 : \beta_3 \neq 0.$$