Chapter 8
Indicator Variables

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" usually indicates the presence of attribute and "0" usually indicates the absence of the attribute. For example, "1" indicator 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 variable** 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 the 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 the easy interpretation of the values and usually turn out to be a satisfactory choice.

In a given regression model, the qualitative and quantitative can also occur together, i.e., some variables are qualitative, and others are 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 an **analysis of covariance model**.

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

Example:

Consider the following model with x_1 as quantitative and D_2 as an 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 the result is essential. We proceed as follows:

If $D_2 = 0,$ then

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 .0 + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \varepsilon \\ E(y / D_2 = 0) &= \beta_0 + \beta_1 x_1 \end{aligned}$$

which is a straight line relationship with intercept β_0 and slope β_1.

If $D_2 = 1,$ then

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

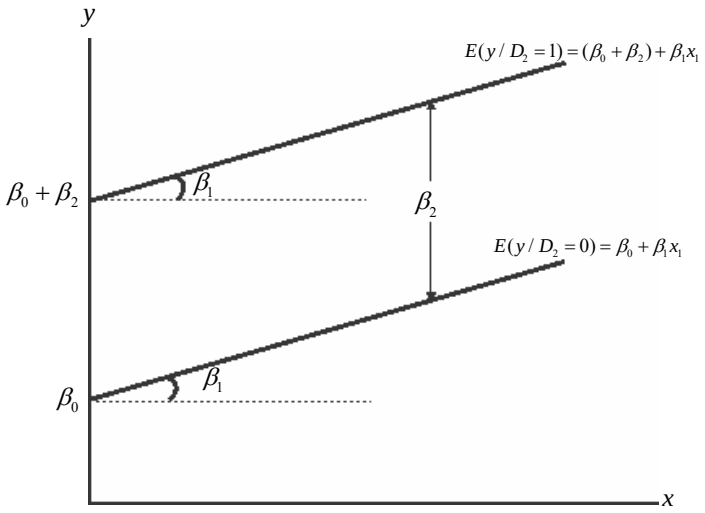
which is a straight-line relationship with intercept $(\beta_0 + \beta_2)$ and slope β_1.

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

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

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 the same variances σ^2.



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 C. 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

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

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

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i,\ i = 1,2,...,n$$
$$E(y_i / D_i = 0) = \beta_0$$
$$E(y_i / D_i = 1) = \beta_0 + \beta_1$$
$$\beta_1 = E(y_i / D_i = 1) - E(y_i / D_i = 0)$$

Thus

- β_0 measures the mean salary of a non-graduate.
- β_1 measures the difference in the mean salaries of a graduate and a 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. $E\left[y_i / D_{i1} = 0, D_{i2} = 1\right] = \beta_0 + \beta_2$: Average salary of a non-graduate
2. $E\left[y_i / D_{i1} = 1, D_{i2} = 0\right] = \beta_0 + \beta_1$: Average salary of a graduate
3. $E\left[y_i / D_{i1} = 0, D_{i2} = 0\right] = \beta_0$: cannot exist
4. $E\left[y_i / D_{i1} = 1, D_{i2} = 1\right] = \beta_0 + \beta_1 + \beta_2$: cannot exist.

Notice that in this case

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

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 the matrix of explanatory variables falls short by 1. So β_0, β_1 and β_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 the **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

$$E(y_i / D_{i1} = 1, D_{i2} = 0) = \beta_1 \Rightarrow \text{Average salary of a graduate.}$$
$$E(y_i / D_{i1} = 0, D_{i2} = 1) = \beta_2 \Rightarrow \text{Average salary of a non} - \text{graduate.}$$

So when intercept term is dropped, then β_1 and β_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 m 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

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

This is a straight line with intercept β_0 and slope β_1. Next

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

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

β_2 reflects the change in intercept term associated with the change in the group of person i.e., when the group changes from A to B.

β_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.

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

and

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 .0 + \beta_3 x_{i1} .0 + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

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 a 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 the same slopes, then the test of hypothesis involves testing

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0.$$

Indicator variables versus quantitative explanatory variable

The quantitative explanatory variables can be converted into indicator variables. For example, if the ages of persons are grouped as follows:

Group 1: 1 day to 3 years

Group 2: 3 years to 8 years

Group 3: 8 years to 12 years

Group 4: 12 years to 17 years

Group 5: 17 years to 25 years

then the variable “age” can be represented by four different indicator variables.

Since it is difficult to collect the data on individual ages, so this will help in an easy collection of data. A disadvantage is that some loss of information occurs. For example, if the ages in years are 2, 3, 4, 5, 6, 7 and suppose the indicator variable is defined as

$$D_i = \begin{cases} 1 & \text{if age of } i^{th} \text{ person is} > 5 \text{ years} \\ 0 & \text{if age of } i^{th} \text{ person is} \le 5 \text{ years.} \end{cases}$$

Then these values become 0, 0, 0, 1, 1, 1. Now looking at the value 1, one can not determine if it corresponds to age 5, 6 or 7 years.

Moreover, if a quantitative explanatory variable is grouped into m categories, then $(m-1)$ parameters are required whereas if the original variable is used as such, then only one parameter is required.

Treating a quantitative variable as a qualitative variable increases the complexity of the model. The degrees of freedom for error is also reduced. This can affect the inferences if the data set is small. In large data sets, such an effect may be small.

The use of indicator variables does not require any assumption about the functional form of the relationship between study and explanatory variables.

Regression analysis and analysis of variance

The analysis of variance is often used in analyzing the data from the designed experiments. There is a connection between the statistical tools used in the analysis of variance and regression analysis.

We consider the case of analysis of variance in one way classification and establish its relation with regression analysis.

One way classification:

Let there are k samples each of size n from k normally distributed populations $N(\mu_i, \sigma^2)$, $i = 1,2,...,k$. The populations differ only in their means, but they have the same variance σ^2. This can be expressed as

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij},\ i = 1,2,...,k;\ j = 1,2,...,n \\ &= \mu + (\mu_i - \mu) + \varepsilon_{ij} \\ &= \mu + \tau_i + \varepsilon_{ij} \end{aligned}$$

where y_{ij} is the j^{th} observation for the i^{th} fixed treatment effect $\tau_i = \mu_i - \mu$ or factor level, μ is the general mean effect, ε_{ij} are identically and independently distributed random errors following $N(0, \sigma^2)$.

Note that

$$\tau_i = \mu_i - \mu, \qquad \sum_{i=1}^{k} \tau_i = 0.$$

The null hypothesis is

$$H_0 : \tau_1 = \tau_2 = ... = \tau_k = 0$$
$$H_1 : \tau_i \neq 0 \text{ for atleast one } i.$$

Employing method of least squares, we obtain the estimator of μ and τ_i as follows

$$S = \sum_{i=1}^{k}\sum_{j=1}^{n} \varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} \left(y_{ij} - \mu - \tau_i\right)^2$$

$$\frac{\partial S}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \bar{y}$$

$$\frac{\partial S}{\partial \tau_i} = 0 \Rightarrow \hat{\tau}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij} - \hat{\mu} = \bar{y}_i - \bar{y}$$

where $\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$.

Based on this, the corresponding test statistic is

$$F_0 = \frac{\left(\frac{n}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2\right)}{\left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2}{k(n-1)}\right)}$$

which follows F-distribution with $k-1$ and $k(n-1)$ degrees of freedom when the null hypothesis is true. The decision rule is to reject H_0 whenever $F_0 \geq F_\alpha(k-1, k(n-1))$ and it is concluded that the k treatment means are not identical.

Connection with regression:

To illustrate the relationship between fixed effect one-way analysis of variance and regression, suppose there are 3 treatments so that the model becomes

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ i = 1,2,...,3,\ j = 1,2,...,n.$$

There are 3 treatments which are the three levels of a qualitative factor. For example, the temperature can have three possible levels – low, medium and high. They can be represented by two indicator variables as

$$D_1 = \begin{cases} 1 & \text{if the observation is from treatment 1} \\ 0 & \text{otherwise,} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{if the observation is from treatment 2} \\ 0 & \text{otherwise.} \end{cases}.$$

The regression model can be rewritten as

$$y_{ij} = \beta_0 + \beta_1 D_{1j} + \beta_2 D_{2j} + \varepsilon_{ij},\ i = 1,2,3;\ j = 1,2,...,n$$

where

D_{1j} : value of D_1 for j^{th} observation with 1^{st} treatment

D_{2j} : value of D_2 for j^{th} observation with 2^{nd} treatment.

Note that

- parameters in the regression model are $\beta_0, \beta_1, \beta_2$.
- parameters in the analysis of variance model are $\mu, \tau_1, \tau_2, \tau_3$.

We establish a relationship between the two sets of parameters.

Suppose treatment 1 is used on j^{th} observation, so $D_{1j} = 1, D_{2j} = 0$ and

$$\begin{aligned} y_{1j} &= \beta_0 + \beta_1.1 + \beta_2.0 + \varepsilon_{1j} \\ &= \beta_0 + \beta_1 + \varepsilon_{1j}. \end{aligned}$$

In case of analysis of variance model, this is represented as

$$\begin{aligned} y_{1j} &= \mu + \tau_1 + \varepsilon_{1j} \\ &= \mu_1 + \varepsilon_{1j} \quad \text{where } \mu_1 = \mu + \tau_1 \end{aligned}$$

$$\Rightarrow \beta_0 + \beta_1 = \mu_1.$$

If treatment 2 is applied on j^{th} observation, then

- in the regression model set up,

$D_{1j} = 0,\ D_{2j} = 1$ and

$$y_{2j} = \beta_0 + \beta_1.0 + \beta_2.1 + \varepsilon_{2j}$$
$$= \beta_0 + \beta_2 + \varepsilon_{2j}$$

- in the analysis of variance model set up,

$$y_{2j} = \mu + \tau_2 + \varepsilon_{2j}$$
$$= \mu_2 + \varepsilon_{2j} \text{ where } \mu_2 = \mu + \tau_2$$

$$\Rightarrow \beta_0 + \beta_2 = \mu_2.$$

When treatment 3 is used on j^{th} observation, then

- in the regression model set up,

$$D_{1j} = D_{2j} = 0$$
$$y_{3j} = \beta_0 + \beta_1.0 + \beta_2.0 + \varepsilon_{3j}$$
$$= \beta_0 + \varepsilon_{3j}$$

- in the analysis of variance model set up

$$y_{3j} = \mu + \tau_3 + \varepsilon_{3j}$$
$$= \mu_3 + \varepsilon_{3j} \text{ where } \mu_3 = \mu + \tau_3$$

$$\Rightarrow \beta_0 = \mu_3.$$

So finally, there are following three relationships

$$\beta_0 + \beta_1 = \mu_1$$
$$\beta_0 + \beta_2 = \mu_2$$
$$\beta_0 = \mu_3$$

$$\Rightarrow \beta_0 = \mu_3$$
$$\beta_1 = \mu_1 - \mu_2$$
$$\beta_2 = \mu_2 - \mu_3.$$

In general, if there are k treatments, then $(k-1)$ indicator variables are needed. The regression model is given by

$$y_{ij} = \beta_0 + \beta_1 D_{1j} + \beta_2 D_{2j} + ... + \beta_{k-1} D_{k-1,j} + \varepsilon_{ij},\ i = 1,2,...,k;\ j = 1,2,...,n$$

where

$$D_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ observation gets } i^{th} \text{ treatment} \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the relationship is

$$\beta_0 = \mu_k$$
$$\beta_i = \mu_i - \mu_k,\ i = 1,2,...,k-1.$$

So β_0 always estimates the mean of k^{th} treatment and β_i estimates the differences between the means of i^{th} treatment and k^{th} treatment.