LINEAR REGRESSION ANALYSIS

MODULE – VIII

Lecture - 27

Indicator Variables

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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 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

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

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 qualitative variable increases the complexity of the model.

The degrees of freedom for error are also reduced. This can effect the inferences if data set is small.

In large data sets, such 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 usually used in analyzing the data from the designed experiments. There is a connection between the statistical tools used in 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 population differ only in their means but they have 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

$$H_0 : \tau_1 = \tau_2 = ... = \tau_k = 0$$

$H_1 : \tau_i \neq 0$ for atleast one i.

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

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

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

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

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

Based on this, the corresponding test statistic is

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

which follows *F*-distribution with k - 1 and k (n - 1) degrees of freedom when 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 connection 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 regression model are $\beta_0, \beta_1, \beta_2$.
- parameters in 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 $D_{1j} = 1$, $D_{1j} = 0$ and

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

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

$$y_{1j} = \mu + \tau_1 + \varepsilon_{1j}$$
$$= \mu_1 + \varepsilon_{1j} \quad \text{where } \mu_1 = \mu + \tau_1$$
$$\Rightarrow \beta_0 + \beta_1 = \mu_1.$$

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

- in regression model set up,

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

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

- in analysis of variance model set up,

$$y_{2j} = \mu + \tau_2 + \varepsilon_{2j}$$
$$= \mu_2 + \varepsilon_{2j} \quad \text{where } \mu_2 = \mu + \tau_2$$
$$\Rightarrow \beta_0 + \beta_2 = \mu_2.$$

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

- in 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 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

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

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

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.