Multiple Linear Regression

In This chapter, we generalized the simple linear regression model as

General Linear Regression Model

In general, the variables $X_1, \ldots, X_{p-1}$ in a regression model do not need to represent different predictor variables, as we shall shortly see. We therefore define the general linear regression model, with normal error terms, simply in terms of X variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

where:

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
$X_{i1}, \ldots, X_{i,p-1}$ are known constants
ε_i are independent $N(0, \sigma^2)$
$i = 1, \ldots, n$

If we let $X_{i0} \equiv 1$, regression model can be written as follows:

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

where $X_{i0} \equiv 1$, or:

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i \qquad \text{where } X_{i0} \equiv 1$$

$k=0$

The response function for regression model is, since $E\{\varepsilon_i\} = 0$:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

This model can be specialized for different cases as follows

1- Simple linear model when p=2

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon,$$

2- Model with some Qualitative Predictor Variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

This model cab be used in different applications such as:

The first-order regression model then is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

where:

X_{i1} = patient's age

$$X_{i2} = \begin{cases} 1 & \text{if patient female} \\ 0 & \text{if patient male} \end{cases}$$

3- Polynomial regression

Polynomial Regression. Polynomial regression models are special cases of the general linear regression model. They contain squared and higher-order terms of the predictor variable(s), making the response function curvilinear. The following is a polynomial regression model with one predictor variable:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

4- Transformed Variables

Transformed Variables. Models with transformed variables involve complex, curvilinear response functions, yet still are special cases of the general linear regression model. Consider the following model with a transformed Y variable:

$$\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

Here, the response surface is complex, yet model can still be treated as a general linear regression model. If we let $Y_i' = \log Y_i$, we can write regression model

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

which is in the form of general linear regression model The response variable just happens to be the logarithm of Y.

Many models can be transformed into the general linear regression model. For instance, the model:

$$Y_i = \frac{1}{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i}$$

can be transformed to the general linear regression model by letting $Y_i' = 1/Y_i$. We then have:

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

5- Interaction Effects

Interaction Effects. When the effects of the predictor variables on the response variable are not additive, the effect of one predictor variable depends on the levels of the other predictor variables. The general linear regression model encompasses regression models with nonadditive or interacting effects. An example of a nonadditive regression model with two predictor variables X_1 and X_2 is the following:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

Here, the response function is complex because of the interaction term $\beta_3 X_{i1} X_{i2}$. Yet regression model is a special case of the general linear regression model. Let $X_{i3} = X_{i1} X_{i2}$ and then write

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

6-Combination of Cases

Combination of Cases. A regression model may combine several of the elements we have just noted and still be treated as a general linear regression model. Consider the following regression model containing linear and quadratic terms for each of two predictor variables and an interaction term represented by the cross-product term:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \varepsilon_i$$

Let us define:

$$Z_{i1} = X_{i1} \qquad Z_{i2} = X_{i1}^2 \qquad Z_{i3} = X_{i2} \qquad Z_{i4} = X_{i2}^2 \qquad Z_{i5} = X_{i1} X_{i2}$$

We can then write regression model (6.16) as follows:

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \beta_4 Z_{i4} + \beta_5 Z_{i5} + \varepsilon_i$$

General Linear Regression Model in Matrix Terms

To express general linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

in matrix terms, we need to define the following matrices:

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Note that the $\mathbf{Y}$ and $\boldsymbol{\varepsilon}$ vectors are the same as for simple linear regression. The $\boldsymbol{\beta}$ vector contains additional regression parameters, and the $\mathbf{X}$ matrix contains a column of 1s as well as a column of the n observations for each of the $p-1$ X variables in the regression model. The row subscript for each element X_{ik} in the $\mathbf{X}$ matrix identifies the trial or case, and the column subscript identifies the $\mathbf{X}$ variable.

In matrix terms, the general linear regression model

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{n\times p}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

where:

$\mathbf{Y}$ is a vector of responses

$\boldsymbol{\beta}$ is a vector of parameters

$\mathbf{X}$ is a matrix of constants

$\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with expectation

$\mathbf{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$ and variance-covariance matrix:

$$\underset{n\times n}{\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

Consequently, the random vector $\mathbf{Y}$ has expectation:

$$\underset{n\times 1}{\mathbf{E}\{\mathbf{Y}\}} = \mathbf{X}\boldsymbol{\beta}$$

and the variance-covariance matrix of $\mathbf{Y}$ is the same as that of $\boldsymbol{\varepsilon}$:

$$\underset{n\times n}{\boldsymbol{\sigma}^2\{\mathbf{Y}\}} = \sigma^2\mathbf{I}$$

Estimation of Regression Coefficients

The least squares criterion is generalized as follows for general linear regression model

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2$$

The least squares estimators are those values of $\beta_0, \beta_1, \ldots, \beta_{p-1}$ that minimize Q. Let us denote the vector of the least squares estimated regression coefficients $b_0, b_1, \ldots, b_{p-1}$ as **b**:

$$\underset{p\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

The least squares normal equations for the general linear regression model

$$\mathbf{X'Xb = X'Y}$$

Hence

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = \hat{\beta} = (X'X)^{-1} X'Y$$

$$\hat{Y} = Xb = X\hat{\beta} = HY, \quad H = X(X'X)^{-1}X'$$

The prove is similar to the simple linear model in the matrix form.

Example: Multiple Regression with Two Predictor Variables (Dwaine Studios)

Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size. These studios specialize in portraits of children. The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y)- thousands- in a community can be predicted from the number of persons aged 16 or younger in the community (X1)- thousands- and the per capita disposable personal income in the community *(*X2)- *thousands*. Data on these variables for the most recent year for the 21 cities in which Dwaine Studios is now operating are shown below:

city	X1	X2	Y
1	68.5	16.7	174.4
2	45.2	16.8	164.4
3	91.3	18.2	244.2
4	47.8	16.3	154.6
5	46.9	17.3	181.6
6	66.1	18.2	207.5
7	49.5	15.9	152.8
8	52	17.2	163.2
9	48.9	16.6	145.4
10	38.4	16	137.2
11	87.9	18.3	241.9
12	72.8	17.1	191.1
13	88.4	17.4	232
14	42.9	15.8	145.3
15	52.5	17.8	161.1
16	85.7	18.4	209.7
17	41.3	16.5	146.4
18	51.7	16.3	144
19	89.6	18.1	232.6
20	82.7	19.1	224.1
21	52.3	16	166.5

```
# How to read txt file in R
mat <- scan('DSD.txt')
mat <- matrix(mat, ncol = 3, byrow = TRUE)
X1=mat[,1]
X2=mat[,2]
Y=mat[,3]
n=length(mat[,1])
one=as.vector(rep(1, n))
X=cbind(one,X1,X2)
b=solve(t(X)%*%X)%*%t(X)%*%Y
model=lm(Y~X1+X2)
summary(model)
```

then we obtain

$$Y = -68.9 + 1.46X1 + 9.37X2.$$

The model coefficients can be interpreted as:

1- There is -68.9 of the sales when X1 and X2 are zeros
2- When in population size increases by one unit (thousand), the sales increases by 1.46 thousand with fixed income.
3- When in income increases by one unit (thousand), the sales increases by 9.37 thousand with fixed population zise.