Multiple Linear Regression

Many applications of regression analysis involve situations in which there are more than one independent variables. A regression model that contains more than one independent variable is called a multiple regression model.

Multiple regression analysis is the study of how a dependent variable Y is related to two or more independent variables. Therefore, multiple regression analysis describes and interprets the relationship between several independent variables and a dependent variable.

Multiple linear regression analysis attempts to model the relationship between two or more independent variables and a dependent variable by fitting a linear equation to the observed data.

Every value of the independent variable x is associated with a value of the dependent variable y.

The concepts of a regression model and a regression equation introduced in simple linear regression are applicable in the multiple linear regression and the concepts are extended in this case.

Multiple Linear Regression Model

The equation that describes how the dependent variable *y* is related to the independent variables $x_1, x_2, \ldots, x_p$ and an error term ε is called the multiple regression model. Here p denotes the number of independent variables.

The general **multiple linear regression model** is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

In the multiple regression model, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters and the error term ε is a random variable. This model reveals that *y* is a linear function of $x_1, x_2, \ldots, x_p$ plus the error term ε. The error term accounts for the variability in *y* that cannot be explained by the linear effect of the *p* independent variables.

Multiple Linear Regression Equation

The equation that describes how the mean value of *y* is related to $x_1, x_2, \ldots, x_p$ is called the multiple regression equation. The **multiple regression equation** is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

If the values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are known, the above equation can be used to compute the mean value of *y* at given values of $x_1, x_2, \ldots, x_p$. Unfortunately, these parameter values are not known and need to be estimated from sample data. A simple random sample is used to compute sample statistics $b_0, b_1, b_2, \ldots, b_p$ that are used as the point estimators of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

Estimated Multiple Linear Regression Equation

The estimated multiple regression equation is as follows

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

where,

$b_0, b_1, b_2, \ldots, b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ and

$\hat{y}$ is the estimated value of the dependent variable (*y*).

Assumptions of Multiple Linear Regression

- The relationship between the dependent variable and independent variables is linear.
- The error term ε is a random variable with a mean or expected value of zero; that is, E(ε) = 0
- The variance of ε, denoted by σ^2 is the same for all values of the independent variables $x_1, x_2, \ldots, x_p$.
- The values of ε are independent.
- The error term ε is a normally distributed random variable.

Estimation Method

Ordinary Least Square method is used to estimate the parameters of multiple linear regression model.

In simple linear regression analysis, we have used the ordinary least square method to develop the estimated regression equation that best approximated the straight-line relationship between the dependent and independent variable. This same approach is used to develop the estimated multiple regression equation.

Interpretation

Interpretation of $b_j (j = 1, 2, \cdots, p)$: The expected change in Y for one unit change in X_j while the other $(p-1)$ number of explanatory variables remain constant.

Interpretation of b_0: It is the value of the dependent variable (Y) when the value of $X_j = 0$. It is of little significance.

Coefficient of Multiple Determination, R^2

The same concept of the coefficient of determination can be extended for a multivariate model.

Coefficient of multiple determination, R^2 is a number that determines the proportion of total variation in Y which is explained by independent variable of the regression line.

It is defined as

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y}_i)^2}$$

The value of R^2 lies between 0 and 1. The greater value of R^2 implies better fitting of the model to the data.

Adjusted R^2

In general R^2 increases when an independent variable is added to the model, regardless of the value of contribution of that variable. Therefore, it is difficult to judge whether an increase in R^2 is really telling us anything important. So, it is preferred to use adjusted R^2.

Adjusted R^2 is the coefficient of multiple determination adjusted for degrees of freedom.

$$R_a{}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

$$= 1 - \frac{(n-1)}{(n-p)}\frac{SSE}{SST}$$

$$= 1 - \frac{(n-1)}{(n-p)}(1 - R^2)$$

Since, $\frac{SSE}{(n-p)}$ is the residual mean square and $\frac{SST}{(n-1)}$ is constant regardless of how many variables are in the model, ${R_a}^2$ will only increase on adding a variable to the model if the addition of variable reduces the residual mean square.

Adjusted R^2 increases only for potential independent variable but R^2 increases for all variables. So, sometimes R^2 may give us misleading conclusion.

For an important or potential independent variable, both R^2 and adjusted R^2 will increase. But for an unimportant independent variable, R^2 will increase but adjusted R^2 will decrease or increase negatively.

Degrees of freedom plays an important role in small samples. So, when sample size is small and when the difference between R^2 and adjusted R^2 is large, we will use adjusted R^2.

${R_a}^2$ is always smaller than R^2. As after adding an independent variable, SSE will decrease, hence R^2 will increase. Besides, ${R_a}^2$ will increase if the independent variable reduces the full term $\frac{SSE}{(n-p)}$.

When $R^2 = 0$ then ${R_a}^2 = \frac{1-p}{n-p}$. When $R^2 = 1$ then ${R_a}^2 = 1$. So $\frac{1-p}{n-p} < {R_a}^2 < 1$

Merits and Demerits of Multiple Linear Regression

Merits

- Multiple linear regression helps us to predict trends and future values.
- It is used to forecast the effects or impacts of changes.
- It helps to understand how much the dependent variable changes when we change the independent variables.
- It can be used to identify the effect that the independent variables have on a dependent variable.

Demerits

- Multiple linear regression assumes that each independent variable has a linear relationship with the dependent variable.
- Multiple linear regression assumes that the independent variables are uncorrelated with each other.
- Violation of assumptions produce biased or inefficient estimates in multiple linear regression.

Example

The following data shows the number of miles traveled, the number of deliveries and the total travel time (in hours) of ten driving assignments of a trucking company.

Miles Traveled, x_1	No. of Deliveries, x_2	Travel Times (y)
100	4	9.3
50	3	4.8
100	4	8.9
100	2	6.5
50	2	4.2
80	2	6.2
75	3	7.4
65	4	6.0
90	3	7.6
90	2	6.1

a) Determine the estimated regression equation on travel times given the number of miles traveled and number of deliveries.

b) What is your interpretation for the model?

c) Comment on the goodness of fit of the model.

d) Predict the travel time for 85 miles traveled and 5 deliveries.

Output from R

lm(formula = y ~ x1 + x2)

Residuals:

Min	1Q	Median	3Q	Max
-0.79875	-0.32477	0.06333	0.29739	0.91333

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**-0.868701**	0.951548	-0.913	0.391634
x1	**0.061135**	0.009888	6.182	0.000453 ***
x2	**0.923425**	0.221113	4.176	0.004157 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5731 on 7 degrees of freedom

Multiple **R-squared: 0.9038**, **Adjusted R-squared: 0.8763**

F-statistic: 32.88 on 2 and 7 DF, p-value: 0.0002762

Interpretation

a) The estimated multiple regression equation is

$$\hat{y} = -0.869 + 0.061x_1 + 0.923x_2$$

b) Interpretation:

$b_0 = -0.869$ means that travel time will be -0.869, when number of miles traveled and the number of deliveries are zero.

$b_1 = 0.061$ means that the average travel time will increase by 0.061 unit when the number of miles traveled will increase by one unit keeping the number of deliveries fixed.

$b_2 = 0.923$ means that the average travel time will increase by 0.923 unit when the number of deliveries will increase by one unit keeping the number of miles traveled fixed.

c) Comment on goodness of fit of the model:

$$R^2 = 0.9038$$

Interpretation: 90.38% variation in total travel times can be explained by the variables number of miles traveled and the number of deliveries.

$$Adjusted\ R^2 = 0.8763$$

Interpretation: 87.63% variation in total travel times can be explained by the variables number of miles traveled and the number of deliveries.

d) The travel time for 85 miles traveled and 5 deliveries is

$$\hat{y} = -0.869 + 0.061 * 85 + 0.923 * 5 = 8.931$$

So, the predicted travel time is approximately 8.9 minutes.

Practice Problems:

Textbook: Probability and Statistics for Engineering and the Sciences (Devore)

CHAPTER 13 Nonlinear and Multiple Regression

Page 568-573: 37(a, b), 39, 47(a), 52(a).

Textbook: Statistical Techniques in Business & Economics (LIND MARCHAL WATHEN)

Chapter 14: MULTIPLE REGRESSION ANALYSIS

Page 530: 17(a), Page 544: 1(b, c, d)