

Multiple Linear Regression

Nihan Acar-Denizli

20 March 2020

Outline

- 1 SIMPLE vs. MULTIPLE LINEAR REGRESSION
- 2 MULTIPLE LINEAR REGRESSION
- 3 Example: Multiple Regression
- 4 Exercise: Simple Linear Regression Computation by Hand

Simple vs. Multiple Linear Regression

Multiple linear regression is used to analyze the relationship between more than two numerical variables.

In multiple regression there are two or more number of predictors.

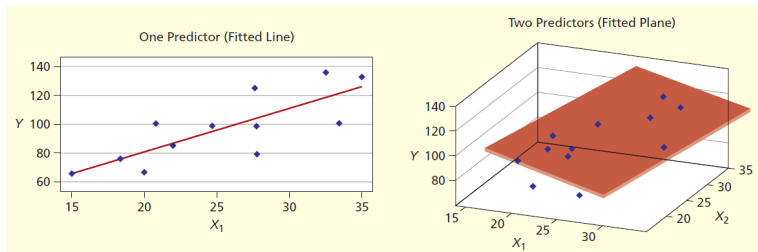


Figure: Simple linear regression vs. multiple linear regression with two predictors

Outline

1 SIMPLE vs. MULTIPLE LINEAR REGRESSION

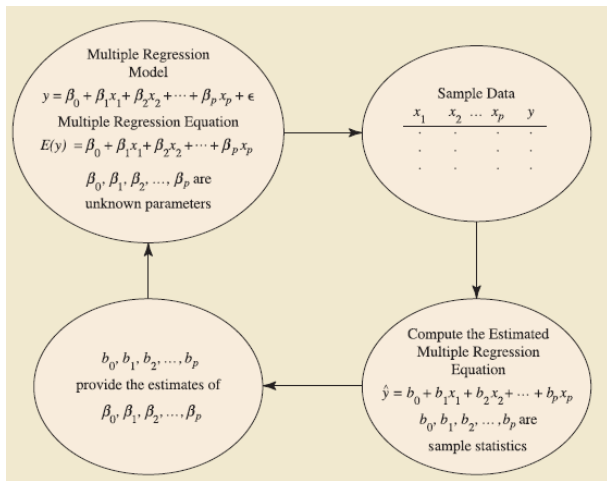
2 MULTIPLE LINEAR REGRESSION

- Multiple Linear Regression
- Matrix Approach to Multivariate Regression
- Measures of Variation
- Multiple Coefficient of Determination
- F Test for Significance in Multiple Linear Regression
- t Test for Significance of the slopes

3 Example: Multiple Regression

4 Exercise: Simple Linear Regression Computation by Hand

Multiple Linear Regression



Multiple Linear Regression

In multiple linear regression model we have more than one independent variable.

Example: The size of a house effects its price. Houses of the same size could have different prices due to the number of rooms, location, presence of a swimming pool or backyard, etc...

Multiple Linear Regression

The basic form of a multiple linear regression is written by,

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon.$$

The intercept β_0 is the expected value of the response when all predictors are equal to zero.

Each regression slope β_j is the expected change of the response Y when the corresponding predictor X_j changes by one unit while all other predictors remain constant ($j = 1, 2, \dots, k$).

Matrix Approach to Multivariate Regression

Assume that we collect a sample of n units (say, houses) and measure all k predictors on each unit (area, number of rooms, etc.).

Then we can write the regression model in the matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where \mathbf{y} is a $n \times 1$ dimensional response vector, \mathbf{X} is $n \times (k + 1)$ dimensional predictor vector, $\boldsymbol{\beta}$ is a $(k + 1) \times 1$ dimensional vector and $\boldsymbol{\epsilon}$ is a $n \times 1$ dimensional vector .

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The Least Squares Method in Matrix Form

For Multivariate regression analysis the least square method minimizes

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})^t (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^t (\mathbf{y} - \mathbf{X}\mathbf{b})$$

Hence, the estimated slopes of the regression model is found from,

$$\mathbf{b} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$$

ANOVA Table

There are three types of measures of variation in the multiple linear regression as in the case of simple linear regression. These measures are summarized in an ANOVA table as follows:

Source	Sum of Squares	Degrees of freedom	Mean Squares	F
Model	$SSR = (\hat{\mathbf{y}} - \bar{\mathbf{y}})^t(\hat{\mathbf{y}} - \bar{\mathbf{y}})$	k	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
Error	$SSE = (\mathbf{y} - \hat{\mathbf{y}})^t(\mathbf{y} - \hat{\mathbf{y}})$	$n - k - 1$	$MSE = \frac{SSE}{n - k - 1}$	
Total	$SST = (\mathbf{y} - \bar{\mathbf{y}})^t(\mathbf{y} - \bar{\mathbf{y}})$	$n - 1$		

Multiple Coefficient of Determination

The coefficient of determination (R^2) is a measure of goodness of fit of the model. For multiple linear regression model it is computed in the same way as in the case of simple linear regression.

$$R^2 = \frac{SSR}{SST}.$$

It is interpreted as the proportion of the variability in the dependent variable that can be explained by the independent variables in the regression model.

Multiple Coefficient of Determination

The value of R^2 increases as independent variables are added to the model. Therefore, in multiple linear regression it is advised to use adjusted multiple coefficient of determination (R_a^2) which is computed as follows:

Adjusted R^2 :

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

where n denotes the number of observations and p denotes the number of independent variables in the model.

F Test for Significance in Multiple Linear Regression

Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0.$$

H_a : At least one of the parameters is not equal to zero.

Test Statistic:

$$F = \frac{SSR/k}{SSE/n - 2} = \frac{MSR}{MSE}$$

Rejection Rule:

- p-value approach: Reject H_0 if $p \leq \alpha$.
- Critical value approach: Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distribution with k degrees of freedom in the numerator and $n - k - 1$ degrees of freedom in the denominator.

F Test is used to test overall significance of the regression model.

t Test for Significance of Model Coefficients

For testing individual slopes we compute the variance covariance matrix of \mathbf{b} .

$$\text{VAR}(\mathbf{b}) = \begin{pmatrix} \text{Var}(b_1) & \text{Cov}(b_1, b_2) & \cdots & \text{Cov}(b_1, b_k) \\ \text{Cov}(b_2, b_1) & \text{Var}(b_2) & \cdots & \text{Cov}(b_2, b_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(b_k, b_1) & \text{Cov}(b_k, b_2) & \cdots & \text{Var}(b_k) \end{pmatrix}$$

t Test for Significance of Model Coefficients

By substituting the equation of \mathbf{b} into the variance function we obtain

$$Var(\mathbf{b}) = \sigma^2(\mathbf{X}^t \mathbf{X})^{-1}.$$

The diagonal elements of this matrix are variances of individual slopes and can be estimated by sample variances:

$$s^2(b_1) = s^2(\mathbf{X}^t \mathbf{X})_{11}^{-1}, \dots, s^2(b_k) = s^2(\mathbf{X}^t \mathbf{X})_{kk}^{-1}$$

t Test for Significance of Model Coefficients

Hypotheses: For the i th variable, the hypotheses are

$$H_0 : \beta_i = 0.$$

$$H_a : \beta_i \neq 0.$$

Test Statistic:

$$t = \frac{b_i}{s_{b_i}}$$

Rejection Rule:

- Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$
- p-value approach: Reject H_0 if $p \leq \alpha$.

where $t_{\alpha/2}$ is based on a t distribution with $n - k - 1$ degrees of freedom.

Outline

- 1 SIMPLE vs. MULTIPLE LINEAR REGRESSION
- 2 MULTIPLE LINEAR REGRESSION
- 3 Example: Multiple Regression
- 4 Exercise: Simple Linear Regression Computation by Hand

An Example: Database Structure

A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. She would like to predict the processed request from the data size (in gigabytes) and the number of tables that used to arrange each data set.

Processed requests	Data size in gigabytes (x_1)	Number of tables (x_2)
40	6	4
55	7	20
50	7	20
41	8	10
17	10	10
26	10	2
16	15	1

The predictor matrix \mathbf{X} and the response vector \mathbf{Y} are written in the form of,

$$\mathbf{X} = \begin{pmatrix} 1 & 6 & 4 \\ 1 & 7 & 20 \\ 1 & 7 & 20 \\ 1 & 8 & 10 \\ 1 & 10 & 10 \\ 1 & 10 & 2 \\ 1 & 15 & 1 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 40 \\ 55 \\ 50 \\ 41 \\ 17 \\ 26 \\ 16 \end{pmatrix}.$$

Then to estimate the vector of slopes we calculate,

$$X^T X = \begin{pmatrix} 7 & 63 & 67 \\ 63 & 623 & 519 \\ 67 & 519 & 1021 \end{pmatrix} \quad \text{and} \quad X^T Y = \begin{pmatrix} 245 \\ 1973 \\ 2908 \end{pmatrix},$$

Hence, we obtain

$$b = (X^T X)^{-1}(X^T Y) = \begin{pmatrix} 52.7 \\ -2.87 \\ 0.85 \end{pmatrix}.$$

Then the regression equation is written as,

$$\hat{y} = 52.7 - 2.87 x_1 + 0.85 x_2,$$

or

number of requests = $52.7 - 2.87$ size of data + 0.85 number of tables.

The predicted values of the response are computed as,

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{pmatrix} 38.9 \\ 49.6 \\ 49.6 \\ 38.2 \\ 32.5 \\ 25.7 \\ 10.5 \end{pmatrix},$$

By using this vector we can compute SSR and SST as,

$$\text{SSR} = (\hat{\mathbf{y}} - \bar{\mathbf{y}})^t (\hat{\mathbf{y}} - \bar{\mathbf{y}}) = 1143.3,$$

and

$$\text{SST} = (\mathbf{y} - \bar{\mathbf{y}})^t (\mathbf{y} - \bar{\mathbf{y}}) = 308.7.$$

Then the ANOVA table is constructed as,

Source	SS	Df	MS	F
Model	1143.3	2	571.7	7.41
Error	308.7	4	77.2	
Total	1452	6		

The F statistic of statistic of 7.41 with 2 and 4 d.f. shows that the model is significant at the level of 0.05.

The coefficient of determination (R^2) value is equal to the ratio,

$$R^2 = \frac{SSR}{SST} = \frac{1143.3}{1452} = 0.787.$$

So, this model with two predictors (data size and the number of tables) explained almost 79% of the variation of the response (processed requests).

The regression variance σ^2 is estimated by s^2 which is equal to the value of MSE in the ANOVA table ($s^2 = 77.2$).

From here, we can compute the estimated variance-covariance matrix,

$$\widehat{\text{VAR}}(b) = s^2(X'X)^{-1} = \begin{pmatrix} 284.7 & -22.9 & -7.02 \\ -22.9 & 2.06 & 0.46 \\ -7.02 & 0.46 & 0.30 \end{pmatrix}.$$

For instance, the t statistic for the second coefficient is computed by,

$$t = \frac{b_2}{s(b_2)} = \frac{0.85}{0.55} = 1.54.$$

Since this value is less than $t_{\alpha/2, n-3} = t_{0.025, 4} = 2.776$, we accept the null hypothesis ($H_0 : \beta_2 = 0$). Therefore, the variable number of tables is not significant in the model at 0.05 significance level.

On Variable Selection

A model with a larger set of predictors is called a full model. Including only a subset of predictors, we obtain a reduced model. We can compare these models by using a partial F statistic which is calculated based on the difference between SSE values of full and the reduced models.

In multiple linear regression, we can construct a model that includes only the most important variables by applying some selection algorithms such as stepwise (forward) selection, backward elimination, etc...

Categorical Predictors and Dummy Variables

Suppose that each sampled computer has one of three operating systems: Unix, Windows, or DOS. In order to use this information for the regression modeling and more accurate forecasting, we create two dummy variables,

$$Z_1 = \begin{cases} 1, & \text{if computer } i \text{ has Unix;} \\ 0, & \text{otherwise.} \end{cases}$$

$$Z_2 = \begin{cases} 1, & \text{if computer } i \text{ has Windows;} \\ 0, & \text{otherwise.} \end{cases}$$

Then the regression model takes the form of,

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + v_1 Z_1 + v_2 Z_2.$$

The slope v_j for a dummy variable is the expected change in the response caused by incrementing Z_j by one unit while keeping all other predictors constant.

In other words, the slope v_j is the difference in the expected response comparing the reference category C with the category j .

Multicollinearity Problem

In the case of multivariate predictors in linear regression, there is another assumption that should be checked. The independent variables should be uncorrelated.

To check this assumption the correlation matrix and Variance Inflation Factor (VIF) values could be used.

If there is a correlation between some of the variables, either one of the correlated variables should be removed from the analysis by using some variable selection procedures or a dimension reduction method of Principal Component Analysis is applied and new uncorrelated variables are created and used in regression analysis.

Outline

- 1 SIMPLE vs. MULTIPLE LINEAR REGRESSION
- 2 MULTIPLE LINEAR REGRESSION
- 3 Example: Multiple Regression
- 4 Exercise: Simple Linear Regression Computation by Hand

Exercise: Simple Linear Regression Computation by Hand






For practice try to construct a simple linear regression model between the processed requests and data size.

Find the regression line and interpret the coefficients. Is the model significant at 0.05 significance level?

Construct the ANOVA table.

Find and interpret the R^2 value.

REFERENCES

-  Anderson, D.R., Sweeney, D.J. and Williams, T.A. (2011). Essentials of modern business statistics with Microsoft Excel. Cengage Learning.
-  Baron, Michael (2014). Probability and Statistics for Computer Scientists. CRC Press.
-  Berenson, M., Levine, D., Szabat, K.A. and Krehbiel, T.C. (2012). Basic business statistics: Concepts and applications, Pearson Higher Education AU.
-  Doane, D.P. and Seward, L.W. (2011). Applied statistics in business and economics. New York, NY: McGraw-Hill/Irwin.
-  Montgomery, D.C. and Runger, G.C. (2003). Applied Statistics and Probability for Engineers. John Wiley & Sons.