Multiple Linear Regression Analysis
Chapter 12

Rationale

Multiple regression analysis is used to estimate models to describe the distribution of a response variable with the help of a number of predictors. A function of the analysis is to search for predictor variables that help to explain significant variation in the response variable. If a number of significant predictors can be identified, then a decision-maker can manage risks and maximize the odds of favorable outcomes. The basic concepts of the simple linear regression analysis generalize in a straightforward manner to the multiple regression analysis.

Objectives

To understand:
- Multiple linear regression model
- Significance testing of predictors
- Multiple coefficient of determination
- Multicollinearity among predictor variables

Key terms

Multiple regression model
Multiple coefficient of determination
Multicollinearity
Indicator variables

Outline

- Multiple regression model
- Estimation of regression coefficients and significance testing
- Testing the model significance
- Multicollinearity and selection of predictors
- Diagnostic plots
Multiple Regression
A methodology for model building

Introduction
The simple regression analysis seeks a relationship of a response variable with only one predictor. A methodology for modeling a response variable Y with the help of more than one predictor is called multiple regression. All the basic steps and concepts from simple regression analysis extend to multiple regression.

The Multiple Regression Model
Let $X_1, X_2, \ldots, X_k$, denote k predictors to be investigated for their relationship with a response variable Y to study its distribution. A multiple linear regression model is a hypothetical relationship such as described below.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

In the equation $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, are called regression coefficients of predictors. The regression coefficient of a predictor quantifies the amount of linear trend in Y. It is the amount of change in Y corresponding to one unit change in a predictor while all other predictors are held fixed at some specified levels. If the scatter plot of Y with a predictor suggests a non-linear trend, then the predictor suitably transformed may preserve the linearity. The expression

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

is hypothesized to be the mean of Y. As in the simple regression, $\epsilon$ captures the sampling error as well as the variation in Y values from its mean. It is assumed normally distributed with mean 0 and standard deviation $\sigma$.

The regression coefficients are estimated from a sample observed through designed experiments on a random sample of n units. For each unit, a vector of (k+1) observations is recorded on Y, and $X_1, X_2, \ldots, X_k$ resulting in a multivariate sample of size n.
Multivariate sample of size $n$

<table>
<thead>
<tr>
<th>Unit</th>
<th>Response $Y$</th>
<th>Predictor $X_1$</th>
<th>Predictor $X_2$</th>
<th>Predictor $X_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_1$</td>
<td>$X_{11}$</td>
<td>$X_{12}$</td>
<td>$X_{1k}$</td>
</tr>
<tr>
<td>1</td>
<td>$Y_2$</td>
<td>$X_{21}$</td>
<td>$X_{22}$</td>
<td>$X_{2k}$</td>
</tr>
<tr>
<td>1</td>
<td>$Y_n$</td>
<td>$X_{n1}$</td>
<td>$X_{n2}$</td>
<td>$X_{nk}$</td>
</tr>
</tbody>
</table>

Examples in Section 12.1

**Estimated model and testing significance of a predictor**

All the sample statistics referred to here are part of the computer print outs following the regression methodology. Therefore, no emphasis is placed here on formulas and their computations.

Let $b_i$ denote an estimate of $\beta_i$ obtained from observed sample with its standard error $s_{bi}$.

The estimated relationship can be expressed by

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

The standard error of $\hat{Y}$ for specified values of predictors is denoted by $s_{\hat{Y}}$. An estimate of $\sigma$ is $s$ which given by

$$s = \sqrt{\sum (Y - \hat{Y})^2 / (n - k - 1)}$$

For comparing regression coefficient $\beta_i$ with a reference number $\beta_{i0}$ use the t-statistic defined below.

$$t = (b_i - \beta_{i0}) / s_{bi}$$

An appropriate decision rule corresponding one or two sided alternative hypothesis is used. If the reference number $\beta_{i0} = 0$, then the testing amounts to deciding if $X_i$ is a significant predictor of $Y$.

For estimating $\beta_i$ by a confidence interval, use
\[ b_i \pm t_{\frac{\alpha}{2}} (s_{bi}), \]

where \( t_{\frac{\alpha}{2}} \) is \((1 \mp \frac{\alpha}{2})\) 100 percentile of the t-distribution with \((n - k - 1)\) degrees of freedom.

\textit{Examples in Section 12.1}

\textbf{Testing model significance}

The coefficient of determination, \( R^2 \), indicates the percentage of variation in \( Y \) that is explained by all the predictors in the equation. The coefficient of determination \( R^2 = 80\% \), say, indicates that 20\% of the variation in \( Y \) is due to all causes other than the predictors as they appear in the expression \( \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \). Equivalently, it is stated that 20\% variation in \( Y \) remains unexplained. It is possible to increase \( R^2 \) by modeling additional predictors at the expense of having a bit more complicated model. In a model building process, a decision is made if an additional predictor contributes significantly to \( R^2 \) to earn its inclusion in the model.

Does the hypothesized model describe a significant amount of variation in the distribution of \( Y \) for different combinations of predictor variables? The null hypothesis that predictors in the relationship have no predictive power to explain the distribution of \( Y \) may be stated as:

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_k \text{ against } H_1 : \text{At least one of } \beta_i \neq 0. \]

The easiest way to reach a decision is by means of p-values. A p-value less than 5\% (1\%) suggests that the estimated model is significant (highly).

\textit{Examples in Section 12.1}

\textbf{Multicollinearity and selection of predictors}

The Multicollinearity occurs when predictors are highly correlated among themselves. In that situation, the model may have high \( R^2 \) value but individual coefficients will be less
reliable having large standard errors. In a model building task, a screening process (such as stepwise regression) is employed to weed out highly correlated variables. While weeding out some variables it is important to also keep in mind which predictors are important to keep from business considerations besides statistical.

Examples in Section 12.2

Diagnostic plots
For the $i^{th}$ unit, there is a $Y_i$ observation corresponding to predictors $X_{i1}, X_{i2}, \ldots, X_{ik}$. For these values of predictors, we can also calculate $\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_k X_{ik}$.

The difference $e_i = Y_i - \hat{Y}_i$ are called residuals; $i = 1, 2, \ldots, n$. These residuals are used to diagnose the validity of the model. The diagnostic plot for multiple regression is a scatter plot of the residuals $e_i$ against the predicted values $\hat{Y}_i$. Such a plot may be used to see if the predictions can be improved by identifying outliers, transformation of predictors to achieve linearity, and unequal variability etc.

Examples in Section 12.2 and 12.3

Indicator predictor variables
If there are predictor variables which are qualitative indicating categories, then they must be represented by indicator variables (also called dummy variables) using numerical codes. For example, a gender variable is a qualitative predictor. For using it in regression analysis, it must be represented by a quantitative indicator variable. For this purpose, define an indicator variable to be 0 for men and 1 for women (or the other way around).

If a qualitative predictor has $c$ categories, then we need to define $(c - 1)$ qualitative indicator variables. First, a baseline category is selected. No indicator variable is defined for it. For each of the remaining categories there is an indicator variable taking value 1 if
the unit belongs to this category and 0 otherwise. When all the \((c - 1)\) indicator variables are 0, it means that the observation comes from the baseline category.

*Examples in Section 12.4*