Chapter 6: Multiple Regression

6.1 Multiple Regression Models

Need for Several Predictor Variables

A single predictor variable in the model would have provided an inadequate description since a number of key variables affect the response variable in important and distinctive ways. A more complex model, containing additional predictor variables. typically is more helpful in providing sufficiently precise predictions of the response variable.

First-Order Model with Two Predictor Variables

When there are two predictor variables X_l and X_2 , the regression model:

A first-order model-

Y_i -

X_{i1} and X_{i2}-

β_0, β_1 and β_2 -

ϵ_i -

Assuming that $E\{\epsilon_i\} = 0$, the regression function for model is:

Analogous to simple linear regression, where the regression function $E\{Y\} = \beta_0 + \beta_1 X_1$ is a line, regression function $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is a plane.

Please refer the Credit data example in the Ch 6 RScript.

Meaning of Regression Coefficients

β_0 represents the mean response $E\{Y\}$ at $X_1 = 0, X_2 = 0$. Otherwise, β_0 does not have any particular meaning as a separate term in the regression model.

β_1 indicates the change in the mean response $E\{Y\}$ per unit increase in X_1 . when X_2 is held constant.

Likewise, β_2 indicates the change in the mean response $E\{Y\}$ per unit increase in X_2 . when X_1 is held constant.

Example 1. Consider the Credit data example in R

1. What is the regression model?
2. Interpret all the regression coefficients

First-Order Model with More than Two Predictor Variables

We consider now the case where there are $p-1$ predictor variables $X_1, ..., X_{p-l}$. The regression model:

is called a first-order model with p - 1 predictor variables. It can also be written:

or, if We let $X_{i0} = 1$, it can be written as:

Assuming that $\{\epsilon_i\} = 0$, the response function for regression model is:

This response function is a _____, which is a plane in more than two dimensions. It is no longer possible to picture this response surface, as we were able to do for the case of two predictor variables. Nevertheless, the meaning of the parameters is analogous to the case of two predictor variables.

General linear Regression Model

In general, the variables $X_1, ..., X_{p_1}$ in a regression model do not need to represent different predictor variables, as we shall shortly See. We define the general linear regression model, with normal error terms, simply in terms of X variables:

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

This general linear model encompasses a vast variety of situations. We consider a few of these now. Namely,

- $p-1$ Predictor Variables.
- Qualitative Predictor Variables.
- Polynomial Regression.
- Transformed Variables.
- Interaction Effects.
- Combination of Cases.

$p-1$ **Predictor Variables.** When _____ represent _____ different predictor variables, general linear regression model is, as we have seen, a first-order model in which there are no interaction effects between the predictor variables.

Qualitative Predictor Variables.

The general linear regression model encompasses not only _____ predictor variables but also _____ ones.

Example 2. -

1.

2.

We Use _____ variables that take on the values _____ and _____ to identify the classes of a qualitative variable.

Example 3. -

1.

2.

Polynomial Regression.

Polynomial regression models are special cases of the general linear regression model. They contain _____ terms of the predictor variable(s), making the response function _____.

The following is a **polynomial regression model** with one predictor variable:

Despite the curvilinear nature of the response function for the above regression model, it is a special Case of general linear regression model.

If We let _____

Transformed Variables.

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

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:

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:

6.2 General Linear Regression Model in Matrix Terms

To express general linear regression model:

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

In matrix terms, the general linear regression model is:

$E\{\epsilon\} = 0$ and variance-covariance matrix:

Consequently, the random vector Y has expectation:

and the variance-covariance matrix of Y is the same as that of ϵ:

6.3 Estimation of Regression Coefficients

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

The least squares estimators are those values of _____ that minimize Q.

Let us denote the vector of the least squares estimated regression coefficients _____ as _____:

The least squares normal equations for the general linear regression model are:

and the least squares estimators are:

6.4 Fitted Values and Residuals

Let the vector of the fitted values _____ be denoted by _____ and the vector of the residual terms _____ be denoted by _____:

The fitted values are represented by:

and the residual terms by: The vector of the fitted values _____ can be expressed in terms of the hat matrix H as follows:

Similarly, the vector of residuals can be expressed as follows

The variance-covariance matrix of the residuals is:

which is estimated by:

6.5 Diagnostics and Remedial Measures

Diagnostics play an important role in the development and evaluation of multiple regression models. Most of the diagnostic procedures for simple linear regression that we described in Chapter 3 carry Over directly to multiple regression. We review these diagnostic procedures now, as well as the remedial measures for simple linear regression that carryover directly to multiple regression.

Following are some of diagnostics and remedial procedures for multiple regression.

1. **Scatter Plot Matrix**

   Scatter plots of the _____ variable against each _____ variable can aid in determining the nature and strength of the bivariate relationships between each of the predictor variables and the response variable and in identifying gaps in the data points as well as outlying data points.

   Scatter plots of each _____ variable against each of the other _____ variables are helpful for studying the bivariate relationships among the predictor variables and for finding gaps and detecting outliers.

   Example 4. **Dwaine Studios example:**

   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) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2). Sales are expressed in thousands of dollars and are labeled Y or $SALES$; the number of persons aged 16 or younger is expressed in thousands of persons and is labeled X_l or $TARGTPOP$ for target population; and per capita disposable personal income is expressed in thousands of dollars and labeled X_2 or $DISPOINC$ for disposable income.

2. **Correlation matrix:**

   A complement to the scatter plot matrix that may be useful at times is the correlation matrix. This matrix contains the coefficients of simple correlation between Y and each of the predictor variables, as well as all of the coefficients of simple correlation among the predictor variables.

3. **Residual Plots:**

   A plot of the residuals against the fitted values: assessing the appropriateness of the multiple regression function and the constancy of the variance of the error terms, as well as for providing information about outliers, just as for simple linear regression.

   In addition, residuals should be plotted against each of the predictor variables. Each of these plots can provide further information about the adequacy of the regression function with respect to that predictor variable (e.g., whether a curvature effect is required for that variable) and about possible variation in the magnitude of the error variance in relation to that predictor variable.

4. **Correlation Test for Normality:** *Described in Chapter 3*

   (a) Find the correlation between residuals and their expected values under normality using R

   (b) **Critical value:** Use the table below to find the critical value depending on given α value and the sample size n

   (c) **Conclusion:** If the observed coefficient exceeds this **Critical value**, we have support for our earlier conclusion that the distribution of the error terms does not depart substantially from a normal distribution.

5. **Breusch-Pagan Test for Constancy of Error Variance:** *Same test described in Chapter 3*

   This test, a large-sample test, assumes that the error terms are independent and normally distributed and that the variance of the error term _____ denoted by _____ , is related to the level of X in the following way:

   Here _____ either _____ or _____ with the level of X, depending on the sign of _____.

   Constancy of error variance corresponds to _____.

   The test:

   (a) H_0 : H_a :

   (b) Test statistic:

   (c) P-value:

   (d) Conclusion:

6. **F Test for Lack of Fit:**

Example 5. Consider Problem 5 from exercises:

Brand preference. In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1I) and sweetness (X_2) of the product, data were obtained from the experiment based on a completely randomized design.

Refer to Ch 6 RScript

Testing the hypothesis:

H_0 : Reduced model is good H_a : Full model is good

Test statistics =

$P - value =$

Conclusion:

6.6 Analysis of Variance Results

- SSTO =
- SSE =
- SSR =

Source	SS	df	MS

Table 1. ANOVA table for general linear regression

F Test for Regression Relation

To test whether there is a regression relation between the response variable _____ and the set of _____ variables _____:

H_0 : H_a :

Test statistics =

$P - value =$

Conclusion:

Example 6. Refer to the RScript for Ch 6: *F Test for Regression Relation, Problem 6.6 in the exercises*

Coefficient of Multiple Determination

The coefficient of multiple determination, denoted by R^2, is defined as follows:

It measures the proportionate reduction of total variation in Y associated with the use of the set of X variables _____. The coefficient of multiple determination R^2 reduces to the coefficient of simple determination for simple linear regression when _____, i.e., when one X variable is in regression model. Just as before, we have:

- Adding more X variables to the regression model can only increase R^2 and never reduce it, because SSE can never become larger with more X variables and SSTO is always the same for a given set of responses.
- Since R^2 usually can be made larger by including a larger number of predictor variables, it is sometimes suggested that a modified measure be used that adjusts for the number of X variables in the model.
- The adjusted coefficient of multiple determination, denoted by _____, _____ by dividing each sum of squares by its associated degrees of freedom:

6.7 Inferences about Regression Parameters

Test for individual β's

H_0 : H_a :

Test statistics =

$P - value =$

Conclusion:

Interval Estimation of β_k

For both Test and the CI we use R. Please refer to Ch 6 RScript.

6.8 Estimation of Mean Response and Prediction of New Observation

Prediction of New Observation $Y_{h(new)}$

Here we get the prediction limits for a new observation $Y_{h(new)}$ corresponding to X_h, the specified values of the X variables using R

Prediction of Mean of m New Observations at X_h

Here we get a prediction interval when m new observations are to be selected at the same levels X_h and their mean $\bar{Y}_{h(new)}$ is to be predicted using R.