Regression Analysis

Jianing Yao
Department of MSIS-RUTCOR
Rutgers University, the State University of New Jersey
Piscataway, NJ 08854 USA

June 9, 2015

Abstract

The notes is used for preparing the qualify exam of multivariate statistics.

Contents

Chapter 1

Single Variable Linear Regression

Assume that we have two variables X and Y, we want to investigate the statistical relationship between them. That is to say: given $X = x$, what is distribution of $Y(x)$? Assume that $Y(x)$ has mean $\mu(x)$ and variance $\sigma^2(x)$, then we define:

$$\epsilon(x) = Y(x) - \mu(x)$$

has mean 0 and variance $\sigma^2(x)$, and we can write the following relationship between Y and X,

$$Y(x) = \mu(x) + \epsilon(x)$$

If we assume further $Y(x)$, or equivalently, $\epsilon(x)$, has normal distribution, then we can visualize it in *Figure* (1.1) In this course, we will focus on when $\mu(\cdot)$ is a linear function of x.

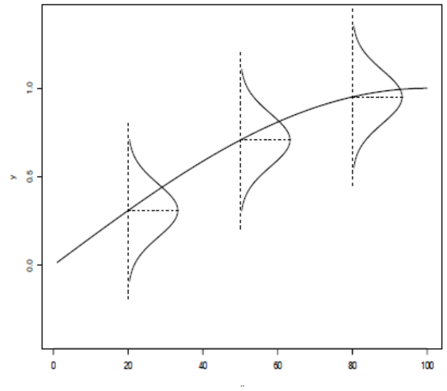


Figure 1.1: Visualization

1.1 Model Analysis

1.1.1 Model Formulation

Let's first list the assumptions for *simple linear regression*:

1. *Linearity:* $\mu(x) = \beta_0 + \beta_1 x$, where $\beta_0, \beta_1 \in \mathbb{R}$;
2. *Constant Variance:* $\mathrm{Var}(\epsilon(x)) = \sigma^2$, for all x, the variance is a constant;
3. *Independence:* ϵ_i's are independent,so observations are;
4. *Normal distribution (optional):* $\epsilon \sim N(0, \sigma^2)$.

Under above assumptions, we can state our linear regression model:

$$Y(x) = \beta_0 + \beta_1(x) + \epsilon$$

For paired variables (X_i, Y_i), where Y_i is random and X_i can be a given constant or random, we have

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.1}$$

where ϵ_i are i.i.d. with mean 0 and variance σ^2.

The model (1.1) says the random variable Y_i follows a probability distribution with $\beta_0 + \beta_1 X_i$ and variance σ^2, but if we take expectation on both sides, then $\mathbb{E}[Y]$ is a linear function of X, we call it *regression functional*, i.e.,

$$\mathbb{E}[Y] = \beta_0 + \beta_1 X \tag{1.2}$$

Here, β_0 is the intercept meaning the mean of Y at $X = 0$. The slope β_1 reflects the amount of increase in expectation of Y when X is increased by one unit. Of course, the observation pair (X_i, Y_i)'s are independently drawn. Also, we shall keep in mind, in the discussion above, we are concerned about the population, which is characterized by random variables.

To end up with this section, we give an alternative version of regression model. Assume $X_0 = 1$,

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i \tag{1.3}$$

where $\beta_0^* = \beta_0 + \beta_1 \bar{X}$.

1.1.2 Model Fitting – Method of Least Square

In model (1.1), X_i, Y_i are random variables (X_i usually not), so it is the relationship between random variables. In practice, we will obtain a sample, a large set of data, (x_i, y_i), where y_i is independently drawn from the population with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. Our objective is to estimate the parameters β_0, β_1 and σ^2. The natural idea will be to find a straight line that fits the data the best !

One of the commonly used criteria is *Least Square Criterion*: to find b_0 and b_1 such that

$$Q(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{1.4}$$

is minimized. Essentially, it minimize the vertical distance between the point (x_i, y_i) and the line. But why we choose a quadratic function instead of

$$\sum_{i=1}^{n} (y_1 - b_0 - b_1 x_i), \tag{1.5}$$

$$\sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|, \tag{1.6}$$

Remember the goal is to reduce the deviation of (x_i, y_i) to the line, however, if the first equation is used, it may happen that deviation is still large but they offset each other. For the second one, it is called *least absolute deviation*, which is also in use, but the non-differentiability of absolute value make it harder than quadratic function.

Setting $\nabla Q(b_0, b_1) = 0$,

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{1.7}$$

$$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - b_1 \bar{x}. \tag{1.8}$$

If we define

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y},$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2,$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Then,

$$
\begin{aligned}
b_1 &= \frac{S_{xy}}{S_{xx}}, \\
b_0 &= \bar{y} - b_1 \bar{x}.
\end{aligned}
$$

We denote

$$
\hat{y}_i = b_0 + b_1 x_i \tag{1.9}
$$

the *fitted value*, and *residual*: $y_i - \hat{y}_i$. There is a number of properties worth noting:

1. the sum of the residuals , i.e., $\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n e_i = 0$;
2. the sum of the observed values y_i equals the sum of the fitted value $\hat{y}_i$, i.e., $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$;
3. the sum of the weighted residuals is 0, i.e., $\sum_{i=1}^n x_i e_i = 0$, and, $\sum_{i=1}^n \hat{y}_i e_i = 0$;
4. the point $(\bar{x}, \bar{y})$ is always on the line $f(t) = b_0 + b_1 t$.

1.1.3 Estimation of Error Terms Varinace σ^2

The variance σ^2 of the error term ϵ_i in regression model needs to be estimated to obtain an indication of the variability of the probability distribution of Y. In addition, as we shall see in the next section, a variety of inferences concerning the regression function and the prediction of Y require an estimate of σ^2.

We know that the variance σ^2 of a single population is estimated by the sample variance s^2. In obtaining the sample variance s^2, we consider the deviation of an observation Y_i from the estimated mean $\bar{Y}$, square it, and then sum all such squared deviations:

$$
\sum_{i=1}^n (Y_i - \bar{Y})^2
$$

Such a sum is called a *sum of squares*. The sum of square is then divided by the degree of freedom associated with it. This number is $n-1$ here, because one degree of freedom is lost by using $\bar{Y}$ as an estimate of the unknown population mean μ. The resulting estimator is the usual sample variance:

$$
s^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}
$$

which is an unbiased estimator of the variance σ^2 of an infinite population. The sample variance is often called a *mean square*, because a sum of square has been divided by the appropriate number of degree of freedom.

Let's now consider our regression model, the variance of each observation Y_i for regression model is σ^2, the same as that of each error term ϵ_i. We need to calculate a sum of squared deviations, but must recognized that the Y_i now come from different probability distributions with different means that depend upon the level X_i. Thus, the deviation of an observation Y_i must be calculated around its own estimated mean $\hat{Y}_i$. Hence, the deviation are the residuals:

$$Y_i - \hat{Y} = e_i$$

and the appropriate sum of squares, denoted by *SSE*, is

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \sum_{i=1}^{n} e_i^2$$

where *SSE* stands for *error sum of squares or residual sum of squares.* The sum of squares *SSE* has $n-2$ degrees of freedom associated with it. Two degrees of freedom are lost because both β_0 and β_1 had to be estimated in obtaining the estimated means $\hat{Y}_i$ and the estimation comes from Y_i's. Hence, the appropriate mean square, denoted by *MSE* or s^2, is:

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

where *MSE* stands for the *error mean square or* or *residual mean square.* It can be shown that *MSE* is an unbiased estimator of σ^2 for regression model, i.e.,

$$\mathbb{E}[MSE] = \sigma^2$$

An estimator of the standard deviation σ is simply $s = \sqrt{MSE}$, the positive square root of *MSE*.

1.1.4 Normality & Maximum Likelihood

We didn't specify the distribution of error term ϵ_i (and hence Y_i), the least square method provides unbiased point estimators of β_0 and β_1 that have minimum variance among all unbiased linear estimators (will be explained in next section). However, to set up interval estimates and make test, we need to make an assumption about the form of the distribution of ϵ_i. The standard assumption is that the error term ϵ_i are normally distributed, and we will adopt it here.

A normal error term greatly simplifies the theory of regression analysis, actually, it is justifiable in many real-world situation where regression analysis is applied. The reason is that error terms frequently represent the effect of factors omitted from the model that affect the response to some extent and that vary at random without reference to the variable X. Thus, as these random effects have a degree of mutual independence, the composite error term ϵ_i representing all these factors would tend to comply with central limit theorem and the error term distribution would approach normality as the number of factor effects becomes large.

When the functional form of the probability distribution of the error term is specified, estimators of the parameters β_0, β_1, σ^2 can be obtained by the method of *maximum likelihood*. The method of *MLE* chooses as estimates those values of the parameters that are most consistent with sample data.

In regression model, each observation Y_i is normally distributed with mean $\beta_0 + \beta_1 X_i$ and standard deviation σ. Suppose we know σ, then we wish to determine the likelihood value for the parameter values β_0, β_1. Fix an X_i, we can get the $\mathbb{E}[Y_i]$, if Y_i is far from $\mathbb{E}[Y_i]$, the density there is relatively small, i.e.,

$$f_i(Y_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\big[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2 \big]$$

The likelihood function for n observations $Y_1, Y_2, ..., Y_n$ is the product of the individual densities above. Since σ^2 of the error terms is usually unknown, the likelihood function is a function of three parameters, β_0, β_1 and σ^2,

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\big[-\frac{1}{2\sigma^2}\left(Y_i - \beta_0 - \beta_1 X_i\right)^2 \big]$$

We shall solve

$$\max_{\beta_0, \beta_1, \sigma^2} L(\beta_0, \beta_1, \sigma^2)$$

The result is the following

$$\hat{\beta_0} = b_0, \ \hat{\beta_1} = b_1, \ \hat{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2}{n}$$

Thus, the estimates $(\hat{\beta_0}, \hat{\beta_1})$ obtained from *MLE* is the same as the one derived from *LSE*. The only difference occurs on σ^2, but only a little, as n becomes large, they are almost equal to each other.

1.2 Statistical Inference

In the previous section, (x_i, y_i)'s are real observations. From the point view of sampling, we can abstract (X_i, Y_i)'s as a set of random variables, where Y_i is independent having a distribution with mean $\beta_0 + \beta_1 X_i$ and variance σ^2.

Theorem 1.2.1 (*Gauss-Markov Theorem*) Under the assumption of simple linear regression model, the least squares estimator b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

Proof. (*we only prove for b_1, b_0 can be verified in a similar fashion*) Firstly, notice

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n}(X_i - \bar{X})Y_i - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X}) \\ &= \sum_{i=1}^{n}(X_i - \bar{X})Y_i\end{aligned}$$

Then

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i$$

where

$$k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

This shows that b_1 is a linear combination of the Y_i's with coefficients k_i. Additionally, k_i's have the following properties:

$$\begin{aligned}&\sum_{i=1}^{n} k_i = \sum_{i=1}^{n}\big[\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\big] = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sum_{i=1}^{n}(X_i - \bar{X}) = 0 \\ &\sum_{i=1}^{n} k_i^2 = \sum_{i=1}^{n}\big[\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\big]^2 = \frac{1}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &\sum_{i=1}^{n} k_i X_i = 1\end{aligned}$$

Given above identity, the unbiasedness of the point estimator b_1 follows easily,

$$\mathbb{E}[b_1] = \mathbb{E}[\sum_{i=1}^{n} k_i Y_i] = \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i X_i = \beta_1$$

The variance of b_1 is also easy to derive, because Y_i's are independent random variables with variance σ^2,

$$\sigma^2(b_1) = \sigma^2\left(\sum_{i=1}^{n} k_i Y_i\right) = \sum_{i=1}^{n} k_i^2 \sigma^2(Y_i) = \sum_{i=1}^{n} k_i^2 \sigma^2 = \sigma^2 \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Moreover, if we replace σ^2 by *MSE* (the unbiased estimator),

$$s^2(b_1) = \frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

is the unbiased estimator of $\sigma^2(b_1)$. Taking the positive square root, we obtain $s(b_1)$, the point estimator $\sigma(b_1)$.

We now prove that b_1 has minimum variance among all unbiased linear estimators of the form:

$$\tilde{b}_1 = \sum_{i=1}^{n} c_i Y_i$$

where c_i's are arbitrary constants. since $\tilde{b}_1$ is required to be unbiased, the following must hold:

$$\mathbb{E}[\tilde{b}_1] = \mathbb{E}[\sum_{i=1}^{n} c_i Y_i] = \beta_1$$

In terms of population, it satisfies $\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i$, thus

$$\mathbb{E}[\tilde{b}_1] = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i = \beta_1$$

To have this hold:

$$\sum_{i=1}^{n} c_i = 0, \;\; \sum c_i X_i = 1.$$

The variance of $\tilde{b}_1$ is,

$$\sigma^2(\tilde{b}_1) = \sum_{i=1}^{n} c_i^2 \sigma^2(Y_i) = \sigma^2 \sum c_i^2$$

Let us define $c_i = k_i + d_i$, where k_i are the least squares constants and d_i are arbitrary constant. We can then write

$$\sigma^2(\tilde{b}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} (k_i + d_i)^2 = \sigma^2 (\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 + 2 \sum_{i=1}^{n} k_i d_i)$$

We know that $\sigma^2 \sum_{i=1}^{n} k_i = \sigma^2(b_1)$ and $\sum_{i=1}^{n} k_i d_i = 0$. Thus,

$$\sigma^2(\tilde{b}_1) = \sigma^2(b_1) + \sigma^2 \sum_{i=1}^{n} d_i^2$$

Note that the smallest value of $\sum_{i=1}^{n} d_i^2$ is zero. Hence, the variance of $\tilde{b}_1$ is at a minimum when $\sum_{i=1}^{n} d_i^2 = 0$. But this can only occur if all $d_i = 0$, which implies $c_i = k_i$. Thus, the least square estimator b_1 has minimum variance among all unbiased linear estimators. □

Remark 1.2.2 Since b_0 and b_1 are both best linear unbiased estimator ($BLUE$), its linear combination is also $BLUE$, i.e., for any given x,

$$(b_0 + b_1 x), \text{ is a } BLUE \text{ of, } (\beta_0 + \beta_1 x).$$

1.2.1 Inference Concerning β_1 and β_0

For statistical inference, let's make normality assumption of ϵ_i, i.e., $\epsilon_i \sim N(0, \sigma^2)$. Then, having above statistics, let's carry out the test below:

$$H_0: \ \beta_1 = 0, \ H_\alpha: \ \beta_1 \neq 0$$

Since b_1 is linear combination of normal random variables, thus it is also normal. We know

$$\frac{b_1 - \beta_1}{\sigma(b_1)}$$

is a standard normal variable. Of course, we need to estimate $\sigma(b_1)$ by $s(b_1)$, and hence are interested in the distribution of statistic

$$\frac{b_1 - \beta_1}{s(b_1)}$$

When a statistics is standardized but the denominator is an estimated standard deviation rather than the true standard deviation, it is called *studentized statistic*. An important theorem in statistics states the following about the studentized statistic

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t(n-2)$$

Since $\frac{b_1-\beta_1}{s(b_1)}$ follows a t distribution, we can make the following probability statement:

$$\mathbb{P}\{t(\alpha/2; n-2) \leq \frac{b_1 - \beta_1}{s(b_1)} \leq t(1-\alpha/2; n-2)\} = 1 - \alpha$$

Here, $t(\alpha/2; n-2)$ denotes the $(\alpha/2)100$ percentile of the t distribution with $n-2$ degree of freedom. Because of the symmetry of the t distribution around its mean 0,

$$\mathbb{P}\{b_1 - t(1-\alpha/2; n-2)s(b_1) \leq \beta_1 \leq b_1 + t(1-\alpha/2; n-2)s(b_1)\} = 1 - \alpha$$

Thus, we have the $1-\alpha$ confidence limits for β_1 are:

$$[b_1 - t(1-\alpha/2; n-2)s(b_1), b_1 + t(1-\alpha/2; n-2)s(b_1)] \tag{1.10}$$

Let's understand what it means. Before drawing the sample, the interval above is random. Thus, the interval containing true β_1 with probability $(1-\alpha)100$ means, among all those intervals, around $(1-\alpha)100$ contains true β_1.

To finish the test, denote

$$t^* = \frac{b_1}{s(b_1)}$$

We reject H_0 if $|t^*| > t(1-\alpha/2; n-2)$ or p-value

$$\mathbb{P}(|t(n-2)| > |t^*|) < \alpha/2$$

Follow the same procedure, we can carry out the test for β_0. Before that, we want to know the sampling distribution of b_0. Notice that

$$b_0 = \bar{Y} - b_1\bar{X}$$

Obviously, b_0 is normal, the mean and variance can be calculated in a similar way as for b_1. They are

$$\begin{aligned}
\mathbb{E}[b_0] &= \beta_0, \\
\sigma^2(b_0) &= \sigma^2\Big[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\Big]
\end{aligned}$$

An estimator of $\sigma^2(b_0)$ is obtained again by replacing σ^2 by its point estimator *MSE*,

$$s^2(b_0) = MSE\Big[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\Big]$$

The positive square root, $s(b_0)$, is an estimator of $\sigma(b_0)$.

Then,

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t(n-2)$$

The $1-\alpha$ confidence limits for β_0 are obtained in the same manner as those for β_1 derived earlier. They are

$$[b_0 - t(1-\alpha/2; n-2)s(b_0), b_0 + t(1-\alpha/2; n-2)s(b_0)]$$

1.2.2 Considerations on Making Inferences

1. If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal, and the use of t distribution will provide approximately the specified confidence coefficient or level of significance. Even if the distributions of Y are far from normal, the estimators b_0 and b_1 generally have the property of *asymptotic normality* – their distributions approach normality under very general condition as the sample size increases;

2. Since the regression model assumes that the X_i are known constants, the confidence coefficient and risks of errors are interpreted with respect to taking repeated samples in which the X observations are kept at the same levels as in the observed sample. The coefficient means that if many independent samples are taken where the levels of X are the same as in the data set and a $1-\alpha$ confidence interval is constructed for each sample, 95 percent of the intervals will contain true value of β_1;

3. From the formula of calculating the variance of b_0 and b_1, we observe that for given n and σ^2, it is affected by the spacing in the X levels, the larger is the quantity $\sum(X_i - \bar{X})$, the smaller is the variance of b_1.

1.2.3 Interval Estimation of Y_h (Mean Response)

A common objective in regression analysis is to estimate the mean for one or more probability distribution of Y. Let X_h denote the level of which we wish to estimate the mean response (Note X_h may be a value which occurred in the sample, or it may be some other value of the predictor variable within the scope of the model). The mean response when $X = X_h$ is denoted by $\mathbb{E}[Y_h]$. We have the point estimator,

$$\hat{Y}_h = b_0 + b_1 X_h$$

for $\mathbb{E}[Y_h]$. We now consider the sampling distribution of $\hat{Y}_h$.

The sampling distribution of $\hat{Y}_h$ refers to the different values of $\hat{Y}_h$ that would be obtained if repeated samples were selected, each holding the levels of the predictor variable X constant, and calculating $\hat{Y}_h$ for each sample.

Theorem 1.2.3 For normal error regression model, the sampling distribution of $\hat{Y}_h$ is normal, with mean and variance

$$\mathbb{E}[\hat{Y}_h] = \mathbb{E}[Y_h],$$
$$\sigma^2(\hat{Y}_h) = \sigma^2\big[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\big].$$

The interesting part if not the proof but the following comments:

Remark 1.2.4 The first identity says that $\hat{Y}_h$ is an unbiased estimator of $\mathbb{E}[Y_h]$. Speaking of variance, we notice the variability o the sampling distribution of $\hat{Y}_h$ is affected by how far X_h is from $\bar{X}$ through the term $(X_h - \bar{X})^2$. The further from $\bar{X}$ is X_h, the greater is the quantity $(X_h - \bar{X})^2$ and the larger is the variance $\hat{Y}_h$. Actually, we find that the variance of $\hat{Y}_h$ is smallest when $X_h = \bar{X}$. Thus, in an experiment to estimate the mean response at a particular level X_h of the predictor variable, the precision of the estimate will be greatest if the observations on X are spaced so that $\bar{X} = X_h$.

When MSE is substituted for σ^2, we obtain $s^2(Y_h)$, the estimated variance of $\hat{Y}_h$:

$$s^2(Y_h) = MSE\big[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\big]$$

The estimated standard deviation of $\hat{Y}_h$ is then $s(\hat{Y}_h)$, the positive square root of $s^2(\hat{Y}_h)$.

Having these statistics, we can also do the test. It should not be surprising

$$\frac{\hat{Y}_h - \mathbb{E}[Y_h]}{s(\hat{Y}_h)} \sim t(n-2)$$

And a confidence interval for $\mathbb{E}[Y_h]$ is constructed in the standard fashion, making use of t distribution,

$$[\hat{Y}_h - t(1-\alpha/2; n-2)s(\hat{Y}_h), \hat{Y}_h + t(1-\alpha/2; n-2)s(\hat{Y}_h)]$$

1.2.4 Prediction of New Observation (Prediction)

Now, we consider the prediction of a new observation Y corresponding to a given level X of the predictor variable. The new observation on Y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based. We denote the level of X for the new trial as X_h and the new observation on Y as $Y_{h(\text{new})}$. Of course, we assume that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation.

The distinction between estimation of the mean response $\mathbb{E}[Y_h]$ and prediction of a new response $Y_h(new)$ is basic. In the former case, we estimate the mean of the distribution of Y. In the present case, we predict an individual outcome drawn from the distribution of Y. Of course, the great majority of the individual outcomes deviate from the mean response, and this must be taken into account by the procedure for predicting $Y_{h(\text{new})}$.

Let's first assume all regression parameters are known. The basic idea of a prediction interval is thus to choose a range in the distribution of Y wherein most of the observations will fall, and then to declare that the next observation will fall in this range. The usefulness of the prediction interval depends, as always, on the width of the interval and the needs for precision by the user. In general, when the regression parameters of normal error regression model are known, the $1-\alpha$ prediction limits for $Y_{h(\text{new})}$ are:

$$\mathbb{E}[Y_h] \pm z(1-\alpha/2)\sigma$$

In centering the limits around $\mathbb{E}[Y_h]$, we obtain the narrowest interval consistent with the specified probability of a correct prediction.

When the regression parameters are unknown, they must be estimated. Prediction limits for a new observation $Y_{h(\text{new})}$ at a given level X_h are obtained by means of the following theorem:

$$\frac{Y_{h(\text{new})} - \hat{Y}_h}{s(\text{pred})} \sim t(n-2)$$

Note that the studentized statistic use the point estimator $\hat{Y}_h$ in the numerator rather than the true mean $\mathbb{E}[Y_h]$ because the true mean is unknown and cannot be used in making a prediction. The estimated standard deviation of the prediction, $s(\text{pred})$, in the denominator of the studentized statistic will be defined shortly. Thus, we can now give the $1-\alpha$ prediction limits for a new observation $Y_{h(\text{new})}$,

$$[\hat{Y}_h - t(1-\alpha; n-2)s(\text{pred}), \hat{Y}_h + t(1-\alpha/2; n-2)s(\text{pred})]$$

Note that the numerator of the studentized statistic represents how far the new observation $Y_{h(\text{new})}$ will deviate from the estimated mean $\hat{Y}_h$ based on the original n case in the study. This difference may be viewed as the prediction error, with $\hat{Y}_h$ serving as the best point estimate of the value of the new observation $Y_{h(\text{new})}$. The variance of this prediction error can be readily obtained by utilizing the independence of the new observation $Y_{h(\text{new})}$ and the original n sample case on which $\hat{Y}_h$ is based. We denote the variance of the prediction error by $\sigma^2(\text{pred})$ and obtain:

$$\sigma^2(\text{pred}) = \sigma^2(Y_{h(\text{new})} - \hat{Y}_h) = \sigma^2(Y_{h(\text{new})}) + \sigma^2(\hat{Y}_h) = \sigma^2 + \sigma(\hat{Y}_h)$$

Note that $\sigma(\text{pred})$ has two components:

1. the variance of the distribution of Y at $X = X_h$, namely σ^2;
2. the variance of the sampling distribution of $\hat{Y}_h$, namely $\sigma^2(\hat{Y}_h)$.

An unbiased estimator of $\sigma(\text{pred})$ is

$$s^2(\text{pred}) = \text{MSE} + s^2(\hat{Y}_h)$$

which can be expressed as follows,

$$s^2(\text{pred}) = \text{MSE}\big[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\big]$$

Remark 1.2.5 The $(1-\alpha)100$ percent prediction interval for $Y_{h(\text{new})}$ is wider than $(1-\alpha)100$ percent confidence interval for $\mathbb{E}[Y_h]$. The reason is that when predicting, we encounter both variability in $\hat{Y}_h$ from sample to sample as well as the variation within the probability distribution of Y. In the meanwhile, as in the case of mean response, the prediction interval is narrow when X_h is located around $\bar{X}$. Although prediction intervals resemble confidence intervals. However, they differ conceptually. A confidence interval represents an inference on a parameter and is an interval that is intended to cover the value of the parameter. A prediction interval, on the other hand, is a statement about the value to be taken by a random variable, the new observation $Y_{h(\text{new})}$.

1.2.5 Confidence Band for Regression Line

At times, we would like to obtain a confidence band for the entire regression line $\mathbb{E}[Y] = \beta_0 + \beta_1 X$. This band enables us to see the region in which the entire regression line lies. It is particularly useful for determining the appropriateness of a fitted regression function.

The *Working-Hotelling* $(1-\alpha)$ confidence band for the regression line for regression model has the following two boundary values at any level X_h:

$$\hat{Y}_h \pm W s(\hat{Y}_h)$$

where

$$W^2 = 2F(1-\alpha; 2, n-2)$$

Note that the formula for the boundary values is of exactly the same form as in the mean response case except for that the t multiple has been replaced by the W multiple. Consequently, the boundary points of the confidence band for the regression line are wider apart the further X_h is from the mean $\bar{X}$ of the X observations. The W multiple will be larger than the t multiple because the confidence band must encompass the entire regression line, whereas the confidence limits for $\mathbb{E}[Y_h]$ at X_h apply only at the single level X_h.

1.3 Analysis of Variance Approach to Regression Analysis (ANOVA)

The analysis of variance approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y. First, let's define the measure of total variation by sum of the squared deviations:

$$\text{SST} = S_{yy} = \sum_{i=1}^{n} (Y_i - \bar{Y})$$

Here SST stands for total sum of squares. If all Y_i observations are the same, SSTO$= 0$. The greater the variation among the Y_i observations, the larger is SSTO. Notice, in this measurement, we disregard variable X.

When utilize the predictor variable X, the variation reflecting the uncertainty concerning the variable Y is that of the Y_i observations around the fitted regression line $Y_i - \hat{Y}_i$. The measure of variation in the Y_i observation that is present when the predictor variable X is taken into account is the sum of the squared deviations, which is the familiar SSE,

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Again, SSE denotes error sum of squares. If all Y_i observations fall on the fitted regression line, SSE= 0. The great the variation of the Y_i observations around the fitted regression line, the larger is SSE.

What accounts for the substantial difference between theses two sums of square is

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

where SSR stands for *regression sum of squares*. Note that SSR is a sum of square deviations, the deviations being $\hat{Y}_i - \bar{Y}$. Each deviation implies the difference between the fitted values on the regression line and the mean of the fitted value $\bar{Y}$. If the regression line is horizontal so that $\hat{Y}_i - \bar{Y} = 0$, then SSR $= 0$. SSR may be considered a measure of that part of the variability of the Y_i which is associated with regression line. The larger SSR is in relation to SST, the greater is the effect of the regression relation in accounting for the total variation in the Y_i observations.

The total deviation $Y_i - \bar{Y}_i$, used in the measure of the total variation of the observations Y_i without taking the predictor variable into account, can be decomposed into two components:

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i$$

The two components are:

1. the deviation of the fitted value $\hat{Y}_i$ around the mean $\bar{Y}$;
2. the deviation of the observation Y_i around the fitted regression line.

It is remarkable property that

$$\text{SST} = \text{SSR} + \text{SSE}$$

Corresponding to the partitioning of the total sum of squares SST, there is a partitioning of the associated degrees of freedom (d.f.). We have $n-1$ d.f. associated with SST. One d.f. is lost because the deviations $Y_i - \bar{Y}$ are subject to one constraint: they mus sum to zero. SSE, as noted earlier, has $n-2$ d.f. because the two parameters β_0 and β_1 are estimated in obtaining the fitted value $\hat{Y}_i$. SSR has one d.f. associated with it. Although there are n deviations $\hat{Y}_i - \bar{Y}$, all fitted values $\hat{Y}_i$ are calculated from the same estimated regression line. Two d.f. are associated with a regression line, corresponding to the intercept and the slope of the line. One of the two d.f. is lost since the deviations $\hat{Y}_i - \bar{Y}$ are subject to a constraint: they must sum up to zero. Thus, the d.f. are additive:

$$n - 1 = 1 + (n - 2)$$

A sum of squares divided by its associated d.f. is called a mean square (MS). We are interested here in the regression mean square, denoted by MSR:

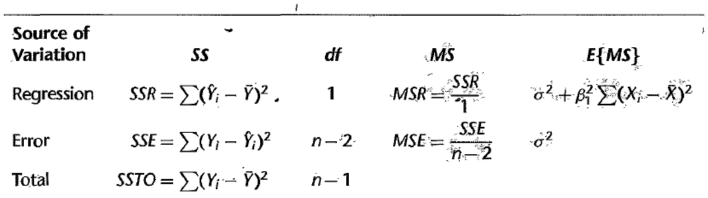
$$\text{MSR} = \frac{\text{SSR}}{1} = \text{SSR},$$

and in the error mean square, MSE, defined earlier,

$$\text{MSE} = \frac{\text{SSE}}{n-2},$$

Remark 1.3.1 Note that two mean squares MSR and MSE are not additive.

Usually, all those informations are summarized in the ANOVA table (*figure* 1.2): In order to make inferences based on the analysis of variance approach, we need to know the expected value of each of the mean squares. The expected value of a mean square is the mean of its sampling distribution and tells us what is being estimated by the mean square. Statistical theory provides the following results:



Source of Variation	*SS*	*df*	*MS*	*E{MS}*
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

Figure 1.2: ANOVA Table

$$\mathbb{E}[\text{MSE}] = \sigma^2,$$
$$\mathbb{E}[\text{MSR}] = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Two important implications of the expected mean squares are the following: (1) the mean of the sampling distribution of MSE is σ^2 whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$; (2) the mean of the sampling distribution of MSR is also σ^2 when $\beta_1 = 0$. Hence, when $\beta_1 = 0$, the sampling distribution of MSR and MSE are located identically and MSR and MSE will tend to be of the same order of magnitude. On the other hand, when $\beta_1 \neq 0$, the mean of the sampling distribution of MSR is greater than σ^2. Thus, when $\beta_1 \neq 0$, the mean of the sampling distribution of MSR is located to the right of that of MSE and, hence MSR will tend to be larger than MSE. Consequently, it suggests that a comparison of MSR and MSE is useful for testing whether or not $\beta_1 = 0$. If MSR and MSE are of the same order of magnitude, this would suggest $\beta_1 = 0$.

The analysis of variance approach provides us with a battery of highly useful test for regression models. For the simple linear regression case considered here, the analysis of variance provides us with a test for:

$$H_0 : \ \beta_1 = 0, \ H_\alpha : \ \beta_1 \neq 0$$

The test statistic for the analysis of variance approach is denoted by F^*. It compares MSR and MSE in the follow fashion:

$$F^* = \frac{\text{MSR}}{\text{MSE}}$$

By *Cochran's theorem*,

$$F^* \sim F(1, n-2)$$

If $F^* \leq F(1-\alpha; 1, n-2)$, we conclude H_0, otherwise, H_α.

1.4 General Linear Test Approach

The analysis of variance test $\beta_1 = 0$ versus $\beta_1 \neq 0$ is an example of the general test for a linear statistical model. We now explain this general test approach in terms of the simple linear regression model.

We begin with the model considered to be appropriate for the data, which in this context is called *full* or *unrestricted model.* For simple linear regression case, the full model is the normal error regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

We fit this full model, either by the method of least squares or by the method of maximum likelihood, and obtain the error sum of squares,

$$\text{SSE(F)} = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \text{SSE}$$

Next, we consider H_0. In this instance, we have

$$H_0 : \ \beta_1 = 0, \ H_\alpha : \ \beta_1 \neq 0$$

The model when H_0 holds is called the *reduced* or *restricted model.* When $\beta_1 = 0$, model reduces to:

$$Y_i = \beta_0 + \epsilon_i$$

We fit this reduced model, by either the method of LSE or MLE and obtain the error sum of squares for this reduced model, denoted by SSE(R). When we fit the particular reduced model, it can be shown that the least squares and maximum likelihood estimator of β_0 is $\bar{Y}$. Hence, the estimated expected value for each observation is $b_0 = \bar{Y}$, and the error sum of squares for this reduced model is:

$$\text{SSE(R)} = \sum_{i=1}^{n}(Y_i - b_0)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{SST}$$

The logic now is to compare the two error sums of squares SSE(F) and SSE(R). It can be shown that

$$\text{SSE(F)} \leq \text{SSE(R)}$$

The reason is that the more parameters are in the model, the better one can fit the data and the smaller are the deviations around the fitted regression function. When SSE(F) is close to SSE(R), the variation of the observations around the fitted regression function for the full model is almost as great as the variation around the fitted regression function for the reduced model. In this case, the added parameters in the full model really do not help to reduce the variation in the Y_i about the fitted regression function. Thus, a small difference SSE(R) − SSE(F) suggests that H_0 holds.

The actual test statistic is a function of SSE(R) − SSE(F), namely:

$$F^* = \frac{\text{SSE(R)} - \text{SSE(F)}}{df_R - df_F} / \frac{\text{SSE(F)}}{df_F}$$

which follows the F distribution when H_0 holds. The degrees of freedom df_R and df_F are those associated with the reduced and full model error sums of squares, respectively. Large value of F^* lead to H_α because a large difference SSE(R) − SSE(F) suggests that H_α holds. In simple linear regression,

$$F^* = \frac{\text{SST} - \text{SSE}}{(n-1)-(n-2)} / \frac{\text{SSE}}{n-2} = \frac{\text{MSR}}{\text{MSE}}$$

which is identical to the analysis of variance test statistic.

1.5 Descriptive Measures of Linear Association between X and Y

In this section, we want to measure ”degree of linear association”. We saw earlier that SST measures the variation in the observations Y_i, or the uncertainty in predicting Y, when no account of the predictor variable X is taken. Thus, SST is a measure of the uncertainty in predicting Y when X is not concerned. Similarly, SSE measures the variation in the Y_i when a regression model utilizing the predictor variable X is employed. A natural measure of the effect of X in reducing the variation in Y, i.e., in reducing the uncertainty in predicting Y, is to express reduction in the variation SST − SSE = SSR as proportion of the total variation:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

The measure R^2 is called *coefficient of determination.* Since $0 \le \text{SSE} \le \text{SST}$, it follows that

$$0 \le R^2 \le 1$$

We may interpret R^2 as the proportionate reduction of total variation associated with the use of the predictor variable X. Thus, the larger R^2 is, the more the total variation of Y is reduced by introducing the predictor variable X.

Remark 1.5.1 We shall realize that no single measure will be adequate for describing the usefulness of a regression model for different applications. The coefficient of determination is subjected to serious misunderstandings.

1. a high coefficient of determination indicates that useful predictions can be made;
2. a high coefficient of determination indicates that the estimated regression line is a good fit;
3. a coefficient of determination near 0 indicates that X and Y are note related.

A measure of linear association between Y and X when both Y and X are random is the *coefficient of correlation.* This measure is the signed square root of R^2: $r = \pm\sqrt{R^2}$. A plus or minus sign is attached to this measure according to whether the slope of the fitted line is positive or negative. Thus the range of r is $-1 \le r \le 1$.

1.6 Normal Correlation Models

In our regression model, we always assume that the X values are known constants. As a consequence of this, the confidence coefficients and risks of errors refer to repeated sampling when the X values are kept the same from sample to sample. Frequently, it may not be appropriate to consider the X values as known constants. The normal correlation model is for this sake.

Let us denote two variables Y_1 and Y_2, we say that Y_1 and Y_2 are jointly normal distributed if the density function of their joint distribution is that of the *bivariate normal distribution*:

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} e^{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right) + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]} \tag{1.11}$$

where μ_i, σ_i are the mean and standard deviation of Y_i respectively, for all i and ρ_{12} is the coefficient of correlation between Y_1 and Y_2. We an show that their marginal distribution have the following characteristics:

$$f_i(Y_i) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{\left[-\frac{1}{2}\left(\frac{Y_i-\mu_i}{\sigma_i}\right)^2\right]}, \text{ for all } i \tag{1.12}$$

Note the following relationship holds:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2} \tag{1.13}$$

where

$$\sigma_{12} = \mathbb{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)] \tag{1.14}$$

The coefficient of correlation ρ_{12} can take on any value between -1 and 1 inclusive. It assumes 1 if the linear relation between Y_1 and Y_2 is perfectly positive and -1 if it is perfectly negative.

As noted, one principal use of bivariate correlation model is to make conditional inference regarding one variable, given the other variable. We denote $f(Y_1|Y_2)$ denote the conditional probability distribution, it can be shown that the conditional probability distribution of Y_1 for any given value of Y_2 is normal with mean $\alpha_{1|2} + \beta_{12}Y_2$ and standard deviation $\sigma_{1|2}$ and its density function is:

$$f(Y_1|Y_2) = \frac{1}{\sqrt{2\pi}\sigma_{1|2}} e^{\left[-\frac{1}{2}\left(\frac{Y_1-\alpha_{1|2}-\beta_{12}Y_2}{\sigma_{1|2}}\right)^2\right]} \tag{1.15}$$

where

$$\alpha_{1|2} = \mu_1 - \mu_2\rho_{12}\frac{\sigma_1}{\sigma_2},\ \beta_{12} = \rho_{12}\frac{\sigma_1}{\sigma_2},\ \sigma_{1|2}^2 = \sigma_1^2(1 - \rho_{12}^2). \tag{1.16}$$

The conditional probability distribution of Y_1 for any given value of Y_2 is normal. Imagine that we slice a bivariate normal distribution vertically at a given value of Y_2, say, at Y_{2h}. That is, we slice it parallel to the Y_1 axis. The exposed cross section has the shape of a normal distribution. This property of normality holds no matter what the value Y_{h2} is. Thus, whenever we slice the bivariate normal distribution parallel to the Y_1 axis, we obtain a normal conditional probability distribution. The means of the conditional probability distributions of Y_1 fall on a straight line, and hence are a linear function of Y_2,

$$\mathbb{E}[Y_1|Y_2] = \alpha_{1|2} + \beta_{12}Y_2 \tag{1.17}$$

Here, $\alpha_{1|2}$ is the intercept parameter and β_{12} the slope parameter. Thus, the relation between the conditional means and Y_2 is given by a linear regression function. All conditional probability distributions of Y_1 have the same standard deviation $\sigma_{1|2}$ Thus, no matter where we slice the bivariate normal distribution parallel to the Y_1 axis, the resulting conditional probability distribution has the same standard deviation. Hence, constant variances characterize the conditional probability distributions of Y_1.

Suppose that we select a random sample of observations (Y_1, Y_2) from a bivariate normal population and wish to make conditional inferences about Y_1, given Y_2. The preceding discussion makes it clear that normal error regression model is entirely applicable because:

1. the Y_1 observations are independent;
2. the Y_1 observations when Y_2 is considered given or fixed are normally distributed with mean $\mathbb{E}[Y_1|Y_2] = \alpha_{1|2} + \beta_{12}Y_2$ and constant variance $\sigma^2_{1|2}$.

In view of the equivalence of each of the conditional bivariate normal correlation models and normal error regression model, all conditional inferences with these correlation models can be made by means of usual regression methods. If the data is generated from a bivariate normal distribution and one wishes to make inferences about Y_2, given a particular value of Y_1, the ordinary regression techniques will be applicable.

Can we still use regression model if Y_1 and Y_2 are not bivariate normal? It can be shown that all results on estimation, testing and prediction obtained from regression model apply if $Y_1 = Y$, $Y_2 = X$ are random variables and if the following conditions hold:

1. the conditional distributions of the Y_i, given X_i, are normal and independent, with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance σ^2;
2. the X_i are independent random variables whose probability distribution $g(X_i)$ does not involve parameters $\beta_0, \beta_1, \sigma^2$.

These conditions require only that regression model is appropriate for the conditional distribution of Y_i, and that the probability distribution of the X_i does not involve the regression parameters. If these conditions are met, all earlier estimation, testing and prediction still hold even though X_i are now random variables. The major modification occurs in the interpretation of the confidence coefficients and specified risks of error. When X is random, these refer to repeated sampling of pairs of (X_i, Y_i) values, where the X_i values as well as Y_i values change from sample to sample. Thus, a confidence coefficient would refer to the proportion of correct interval estimates if repeated samples were obtained and confidence interval calculated for each sample.

A principal use of the bivariate normal correlation model is to study the relationship between two variables. In a bivariate normal model, the parameter ρ_{12} provides information about the degree of the linear relationship between the two variables Y_1 and Y_2. The MLE estimator of ρ_{12}, denoted by r_{12}, is given by

$$r_{12} = \frac{\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sqrt{\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)^2 \sum_{i=1}^{n}(Y_{i2} - \bar{Y}_2)^2}} \tag{1.18}$$

This estimator is often called the *Pearson product-moment correlation coefficient.* It is a biased estimator of ρ_{12} (unless $\rho_{12} = 0, 1$), but the bias is small when n is large. It can be shown that the range of r_{12} is:

$$-1 \leq r_{12} \leq 1$$

Generally, values of r_{12} near 1 indicate a strong positive linear association between Y_1 and Y_2 whereas values of r_{12} near -1 indicate a strong negative linear association. Values of r_{12} near 0 indicate little or no linear association between Y_1 and Y_2.

When the population is bivariate normal, it is frequently desired to test whether the coefficient of correlation is 0,

$$H_0 : \ \rho_{12} = 0, \ H_\alpha : \ \rho_{12} \neq 0 \tag{1.19}$$

The reason for interest in this test is that in the case where Y_1 and Y_2 are jointly normal distributed, $\rho_{12} = 0$ implies that Y_1 and Y_2 are independent. We can use regression procedures for the test since it is equivalent to:

$$H_0 : \ \beta_{12} = 0, \ H_\alpha : \ \beta_{12} \neq 0 \tag{1.20}$$

It can be shown that the test statistics is the same as before and can be expressed directly in terms of r_{12},

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1 - r_{12}^2}} \tag{1.21}$$

If H_0 holds, t^* follows the $t(n-2)$ distribution.

Chapter 2

Simple Linear Regression – Miscellanea

2.1 Diagnostic – Residual Analysis

The residual term is defined as:

$$e_i = Y_i - \hat{Y}_i$$

which can be regarded as the observed error, in distinction to the unknown true error ϵ_i in the regression mode:

$$\epsilon_i = Y_i - \mathbb{E}[Y_i]$$

From the assumptions of linear regression, we assume the error terms ϵ_i are independent normal random variables with mean 0 and constant variance σ^2. If the model is appropriate for the data at hand, the observed residuals e_i should then reflect the properties assumed for the ϵ_i. This is the basic idea underlying residual analysis.

2.1.1 Properties of Residuals

The mean of the n residuals e_i for the simple linear regression model is

$$\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n} = 0$$

Thus, since $\bar{e}$ is always 0, it provides no information as to whether the true errors ϵ_i have expected value $\mathbb{E}[\epsilon_i] = 0$. While the variance of the n residuals e_i is defined as

$$s^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \text{MSE}$$

If the model is appropriate, MSE is, as noted earlier, an unbiased estimator of the variance of the error terms σ^2.

The residuals e_i are not independent random variables because they involve the fitted values $\hat{Y}_i$ which are based on the same fitted regression function. As a result, the residuals for regression model are subject to two constraints: the sum of e_i must be 0 and the sum of the product of $X_i e_i$ is 0 as well. When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals e_i is relatively unimportant and can be ignored for most purpose.

Often-times, it is helpful to standardize the residuals for residual analysis. Since the standard deviation of the error term ϵ_i is σ, which is estimated by $\sqrt{\text{MSE}}$, it is natural to consider the following form of standardization:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{\text{MSE}}}$$

We call the statistic e_i^* a *semi-studentized residual.*

2.1.2 Diagnostic of Residuals

We shall consider the use of residuals for examining six important types of departures from the simple linear regression model with normal errors:

1. the regression function is not linear;
2. the error terms do not have constant variance;
3. the error terms are not independent;
4. the model fits all but one or a few outliers observations;
5. the error terms are not normally distributed;
6. one or several important predictor variables have been omitted from the model.

We examine them one by one via plotting effective graphs.

(1) Non-linearity of Regression Function: this can be studied from a residual plot against the predictor variable (or equivalently, fitted values). Note, it can also be studied by the scatter plot, but this plot is not always as effective as a residual plot. Observation of systematic fashion, e.g., residuals are negative for smaller X values, positive for medium-size X values and negative again for large X values (see in *Figure* 2.1), shows a bigger chance of non-linearity. The advantages of residual plot against scatter plot is: first, the residual plot

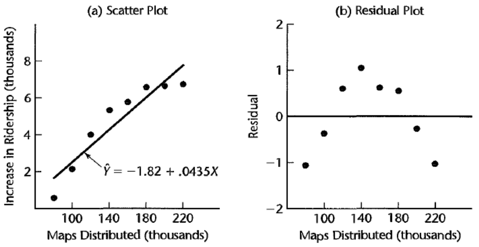


Figure 2.1: Scatter plot v.s. Residual against Predictors

can easily be used for examining other facets of the aptness of the model; second, there are occasions when the scaling of the scatter plot places the Y_i observations close to the fitted values $\hat{Y}_i$, for instance, when there is a steep slope, it then becomes more difficulty to study the appropriateness of a linear regression function from the scatter plot.

(2) Non-constancy of Error Variance: plots of the residuals against the predictor variable is also helpful to check whether the variance is constant or not. If, for different level of X, we observe different level of spreading of e_i, then the variance may not be constant (see in *Figure* 2.2). Plots of the absolute values of the residuals or of the squared residuals

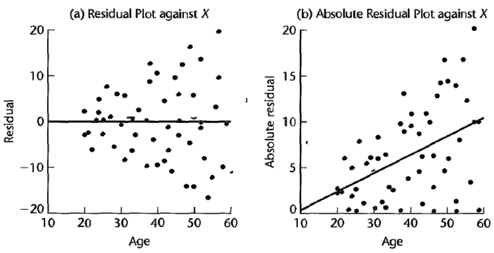


Figure 2.2: Residual against Predictor v.s. Absolute Residual against predictor

against the predictor variable X or against the fitted values $\hat{Y}$ are also useful for diagnosing non-constancy of the error variance since the signs of the residuals are not meaningful for examining the constancy of the error variance. These plots are especially useful when there are not many cases in the data set because plotting of either the absolute or squared residuals places all of the information on changing magnitudes of the residuals above the horizontal zero line so that one can more readily see whether the magnitude of the residuals is changing with the level of X or $\hat{Y}$.

(3) Presence of Outliers: outliers are extreme observations. Residual outliers can be identified from residual plots against X or $\hat{Y}$, as well as from box plots, stem-and-leaf lots, and dot plots of the residuals. Plotting of semi-studentized residuals is particularly helpful for distinguishing outlying observations, since it then becomes easy to identify residuals that

lie many standard deviations from 0. A rough rule of thumb when the number of cases is large is to consider semi-studentized residuals with absolute value of four or more to be outliers.

Outliers can create great difficulty. When we encounter one, our first suspicion is that the observation resulted from a mistake or other extraneous effect, and hence should be discarded. A major reason for discarding it is that under the LSE method, a fitted line may be pulled disproportionately toward an outlying observation because the sum of the squared deviation is minimized. This could cause a misleading fit if indeed the outlying observation resulted from a mistake or other extraneous cause. On the other hand, outliers may convey significant information, as when the outlier occurs because of an interaction with another predictor variable omitted from the model. A safe rule frequently suggested is to discard an outlier only if there is a direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance.

(4) Non-independence of Error Terms: whenever data are obtained in a time sequence or some other type of sequence, such as for adjacent geographic areas, it is a good idea to prepare a sequence plot of the residuals. The purpose of plotting the residuals against time or in some other type of sequence is to see if there is any correlation between error terms that are near each other in the sequence (see in *Figure* 2.3). Another type of non-independence of the error terms is illustrated in *Figure* 2.3. Here the adjacent error terms are also related, but the resulting pattern is a cyclical one with no trend effect present. when the error terms are independent, we expect the residuals in a sequence plot to fluctuate in more or less random pattern around the base line 0. Lack of randomness can take form of too much or too little alternation of points around the zero line. In practice there is little concern with the former because it does not arise frequently.

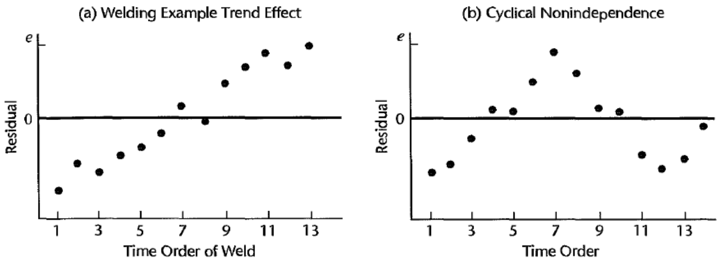


Figure 2.3: Non-independence Check

(5) Non-normality of Error Terms: a box plot, histogram, stem-and-leaf plot of the residuals can be helpful for detecting non-normality, but the requirement is to have sufficiently large sample. Another possibility when the number of case is reasonably large is to compare actual frequencies of the residuals against expected frequencies under normality. For example, one can determine whether, say, about 68 percent of the residuals e_i fall between $\pm\sqrt{\text{MSE}}$. When the sample size is moderately large, corresponding t values may be

used for the comparison. Still another possibility is to prepare a normal probability plot of the residuals. Here each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality whereas a plot that departs substantially from linearity suggest that the error distribution is nor normal (see in *Figure* 2.4).

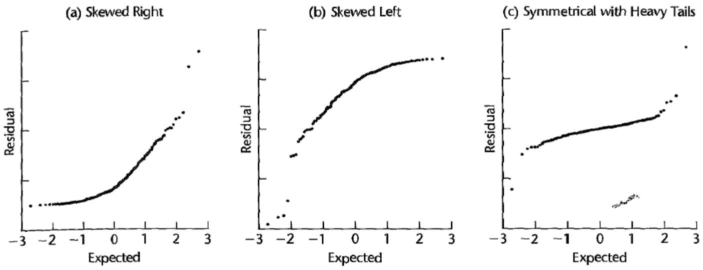


Figure 2.4: Normal Probability Plots when Error Term Distribution is not Normal

(6) Omission of Important Predictor Variables: residuals should also be plotted against variables omitted from the model that might have important effects on the response. The purpose of this additional analysis is to determine whether there are any other key variables that could provide important additional descriptive and predictive power to the model. The residuals are plotted against the additional predictor variable to see whether or not the residuals tend to vary systematically with the level of the additional predictor variable.

2.2 Tests Involving Residuals

Graphic analysis of residuals is inherently subjective. Nevertheless, subjective analysis of a variety of interrelated residual plots will frequently reveal difficulties with the model more clearly than particular formal tests. There are occasions, however, when one wishes to put specific questions to a test. We now briefly review some of the relevant tests.

1. *test for randomness:* the *Durbin-Watson test* is specifically designed for lack of randomness in least square residuals;
2. *test for constancy of variance:* a simple test is based on the rank correlation between the absolute values of the residuals and the corresponding values of the predictor variable. Two other simple tests for constancy of the error variance – the *Brown-Forsythe test* and the *Breusch-Pegan test*;
3. *test for outliers:* a simple test for identifying an outlier observation involves fitting a new regression line to the other $n - 1$ observations. The suspect observation, which was not used in fitting the new line, can now be regarded as a new observation. one

can calculate the probability that in n observations, a deviation from the fitted line as great as that of the outlier will be obtained by chance. If this probability is sufficiently small, the outlier can be rejected as not having come from the same population as the other $n-1$ observations;

4. *test for normality* goodness of fit tests can be used for examining the normality of the error terms. For instance, the chi-square test for the *Kolmogorov-smirnov test* and its modification, the *Lilliefors test*, can be employed for testing the normality of the error terms by analysing the residuals.

2.2.1 F Test for Lack of Fit

We next take up a formal test for determining whether a specific type of regression function adequately fits the data. We illustrate this test for ascertaining whether a linear regression function is a good fit for the data.

The lack of fit test assumes that the observations Y for given X are independent and normally distributed and that the distributions of Y have the same variance σ^2. The lack of fit test requires repeat observations at one or more X levels. Repeat trials for the same level of the predictor variable, of the type described, are called replications. The resulting observations are called replicates.

First, we need to modify our notation to recognize the existence of replications at some levels of X. We shall denote the different X levels in the study, whether or not replicated observations are present, as $X_1, ..., X_c$. Further, we shall denote the number of replicates for the j-th level of X as n_j, thus the total number of observations n is given by:

$$n = \sum_{j=1}^{c} n_j$$

We shall denote the observed value of the response variable for the i-th replicate for the j-th level of X by Y_{ij}, $i = 1, ..., n_j$, $j = 1, ..., c$.

The general linear test approach begins with the specification of the full model. The full model used for the lack of fit test makes the same assumptions as the simple linear regression model except for assuming a linear regression relation, the subjective of the test. This full model is:

$$Y_{ij} = \mu_j + \epsilon_{ij}, \text{ full model}$$

where μ_j are parameters $j = 1, ..., c$, ϵ_{ij} are independent $N(0, \sigma^2)$. Since the error term have expectation 0, it follows that

$$\mathbb{E}[Y_{ij}] = \mu_j$$

Thus, the parameter μ_j, $j = 1, ..., c$ is the mean response when $X = X_j$ and a random error term. The difference between the two models is that in the full model, there are no restrictions on the means μ_j, whereas in the regression model the mean response are linearly related to X, i.e., $\mathbb{E}[Y] = \beta_0 + \beta_1 X$. To fit the full model to the data, we require the least squares or maximum likelihood estimators for the parameters μ_j. It can be shown that these estimators of μ_j are simply $\bar{Y}_j$:

$$\hat{\mu}_j = \bar{Y}_j$$

Thus, the estimated expected value for observation Y_{ij} is $\bar{Y}_j$, and the error sum of squares for the full model therefore is

$$\text{SSE(F)} = \sum_{j=1}^{c} \sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)^2 = \text{SSPE}$$

In the context of the test for lack of fit, the full model error sum of squares is called the *pure error sum of squares* and is denoted by SSPE. The degrees of freedom associated with SSPE can be obtained by recognizing that the sum of squared deviations at a given level of X is like an ordinary total sum of squares based on n observations, which has $n - 1$ degrees of freedom associated with it. Here, there are n_j observations when $X = X_j$; hence the degree of freedom are $n_j - 1$. Just as SSPE is the sum of the sum of squares, so the number of degrees of freedom associated with SSPE is the sum of the component degrees of freedoms:

$$d.f_F = \sum_{j} (n_j - 1) = n - c$$

The general linear test approach next requires consideration of the reduced model under H_0. For testing the appropriateness of a linear regression relation, the alternatives are:

$$H_0: \; \mathbb{E}[Y] = \beta_0 + \beta_1 X, \; H_\alpha: \; \mathbb{E}[Y] \neq \beta_0 + \beta_1 X.$$

Thus, H_0 postulates that μ_j in the full model is linearly related to X_j,

$$\mu_j = \beta_0 + \beta_1 X_j$$

The reduced model under H_0 therefore is:

$$Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}, \text{ reduced model}$$

Note that the reduced model is the ordinary simple linear regression model, with subscripts modified to recognize the existence of replications. We know that the estimated expected value for observation Y_{ij} with regression model is the fitted value $\hat{Y}_{ij}$,

$$\hat{Y}_{ij} = b_0 + b_1 X_j$$

Hence, the error sum of squares for the reduced model is the usual error sum of squares SSE:

$$\text{SSE(R)} = \sum_{j=1}^{c} \sum_{i=1}^{n} [Y_{ij} - (b_0 + b_1 X_j)]^2 = \text{SSE}$$

We also know that the degree freedom associated with SSE(R) are:

$$d.f_R = n - 2$$

The general linear test statistic is defined as:

$$F^* = \frac{\text{SSE}-\text{SSPE}}{(n-2)-(n-c)} / \frac{\text{SSPE}}{n-c}$$

The difference between the two error sums of squares is called the *lack of fit sum of squares* here and is denoted by SSLF:

$$\text{SSLF} = \text{SSE} - \text{SSPE}$$

We can then express the test statistic as follows:

$$F^* = \frac{\text{MSLF}}{\text{MSPE}}$$

where MSLF denotes the *lack of fit mean square* and MSPE denotes the *pure error mean square*. We now that large values of F^* lead to conclusion H_α in the general liear test. Decision rule becomes, if $F^* \leq F(1-\alpha; n-2, n-c)$, conclude H_0, otherwise, conclude H_α.

The definition of the lack of fit sum of squares SSLF indicates that we have, in fact, decompose the error sum of squares SSE into two components:

$$Y_{ij} - \hat{Y}_{ij} = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \hat{Y}_{ij})$$

This identity shows that the error deviations in SSE are made up of a pure error component and a lack of fit component. Note that we can define the lack of fit sum of squares directly as follows:

$$\text{SSLF} = \sum_{j=1}^{n} \sum_{i=1}^{c} (\bar{Y}_j - \hat{Y}_{ij})^2 = \sum_{j=1}^{n} n_j (\bar{Y}_j - \hat{Y}_j)^2$$

Formula indicates clearly why SSLF measures lack of fit. If the linear regression function is appropriate, then the means $\bar{Y}_j$ will be near the fitted values $\hat{Y}_j$ and SSLF will be small. On the other hand, if the linear regression function is no appropriate, the means $\bar{Y}_j$ will not be near the fitted values calculated.

2.3 Overview of Remedial Measures

If the simple linear regression model is not appropriate for a data set, there are two basic choices:

1. abandon regression model and develop and use a more appropriate model;
2. employ some transformation on the data so that regression model is appropriate for the transformed data.

2.3.1 General Ideas

We will focus on the transformation case. Before introducing various of transformation, let's give a brief overview of remedial measures:

1. *non-linearity of regression function:* when the regression function is not linear, a direct approach is to modify regression model by altering the nature of regression function. The transformation approach employs a transformation to linearize, at least approximately, a non-linear regression function;

2. *non-constancy of error variance:* when the error is not constant but varies in a systematic fashion, a direct approach is to modify the model to allow for this and use the method of weighted least squares to obtain the estimators of the parameters;

3. *non-independence of error terms:* when the error terms are correlated, a direct remedial measure is to work with a model that calls for correlated error terms. A simple remedial transformation that is often helpful is to work with the first differences;

4. *non-normality of error terms:* lack of normality and non-constant error variance frequently go hand in hand. Fortunately, it is often the case that the same transformation that helps stablize the variance is also helpful in approximately normalizing the error terms;

5. *omission of important predictor variables:* when residual analysis indicates that an important predictor variable has been omitted from the model the solution is to modify the model;

6. *outlying observations:* when outlying observations are present, use of the least squares and maximum likelihood estimators for regression model may lead to serious distortions in the estimated regression function. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations.

2.3.2 Transformation

We first consider transformation for linearizing a nonlinear regression relation **when the distribution of the error term is reasonably close to a normal distribution and the error terms have approximately constant variance**. In this situation, transformations on X should be attempted. The reason why transformations on Y may not be desirable here is that a transformation on Y may materially change the shape of the distribution of the error terms from the normal distribution and may also lead to substantially differing error term variances.

When the error variance is unequal and non-normality of the error terms frequently appear together, we need a transformation on Y, since the shapes and spreads of the distribution of Y need to be changed. Such a transformation on Y may also at the same time help to linearize a curvilinear regression relation. At other times, a simultaneous transformation on X may be needed to obtain or maintain a linear regression.

It is often difficulty to determine from diagnostic plots which transformation of Y is most appropriate for correcting skewness of the distributions of error terms, unequal error variances, and non-linearity of the regression function. *Box-Cox procedure* automatically identifies a transformation from the family of power transformations of Y. The family transformation is of the form:

$$Y' = Y^{\lambda}$$

where λ is a parameter to be determined from the data. The normal error regression model with response variable a member of the family power transformations becomes:

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \epsilon_i$$

Note that regression model includes an additional parameter, λ, which needs to be estimated. The *Box-Cox* procedure uses the method of MLE to estimate λ, as well as the other parameters $\beta_0, \beta_1, \sigma^2$. In this way, the *Box-Cox* identifies $\hat{\lambda}$, the maximum likelihood estimate of λ to use in the power transformation.

2.4 Simultaneous Inferences

The *Bonferroni procedure* for developing joint confidence intervals for β_0 and β_1 with a specified family confidence coefficient is very simple: each statement confidence coefficient is adjusted to be higher than $1-\alpha$ so that the family confidence is at least $1-\alpha$. The procedure is a general one that can be applied in many cases.

We start with ordinary confidence limits for β_0 and β_1:

$$\begin{gathered} b_0 \pm t(1-\alpha/2; n-2)s(b_0), \\ b_1 \pm t(1-\alpha/2; n-2)s(b_1), \end{gathered}$$

By simple probability knowledge, we realize that if β_0 and β_1 are separately estimated with say, 95 percent confidence intervals, the *Bonferroni inequality* guarantees us a family confidence coefficient of at least 90 percent, that both intervals based on the same sample are correct. We can easily use this to obtain a family confidence coefficients of at least $1-\alpha$ for estimating β_0 and β_1. We do this by estimating them separately with statement confidence coefficients of $1-\alpha/2$ each. This yields the *Bonferroni bound* $1 - 1 - \alpha/2 - \alpha/2 = 10\alpha$. Thus, the $1-\alpha$ family confidence limits for β_0 and β_1 for regression model by *Bonferroni* *procedure* are:

$$b_i \pm BS(b_i),\ i = 1, 2;$$

where $B = t(1-\alpha/4; n-2)$. Note that a statement confidence coefficient of $1-\alpha/2$ requires the $(1-\alpha/4)100$ percentile of the t distribution for a two-sided confidence interval.

The *Working-Hotelling procedure* is based on the confidence band for the regression line. The confidence band contains the entire regression line and therefore contains the mean responses at all X levels. Hence, we can use the boundary values of the confidence band at selected X level as simultaneous estimates of the mean response at these X levels. The family confidence coefficient for these simultaneous estimates will be at least $1-\alpha$ because the confidence coefficient that entire confidence band for the regression line is correct is $1-\alpha$.

The *Working-Hotelling procedure* for obtaining simultaneous confidence intervals for the mean responses at selected X level is therefore simply to use the boundary values for the X levels of interest. The simultaneous confidence limits for g mean responses $\mathbb{E}[Y_h]$ for regression model with *Working-Hotelling procedure* therefore are:

$$\hat{Y}_h \pm Ws(\hat{Y}_h)$$

where $W^2 = 2F(1-\alpha; 2, n-2)$.

On the other hand, the *Bonferroni procedure*, discussed earlier for simultaneous estimation of β_0 and β_1, is a completely general procedure. To construct a family of confidence intervals for mean responses at different X levels with this procedure, we calculate in each instance the usual confidence limits for a single mean response $\mathbb{E}[Y_h]$, adjusting the statement confidence coefficient to yield the specified family confidence coefficient.

When $\mathbb{E}[Y_h]$ is to be estimated for g levels X_h with family confidence coefficient $1-\alpha$, the *Bonferroni confidence limits* for the regression model are:

$$\hat{Y}_h \pm Bs(\hat{Y}_h)$$

where

$$B = t(1-\alpha/2g; n-2)$$

and g is the number of the confidence intervals in the family.

Remark 2.4.1 Simultaneous prediction intervals for new observations is exactly the same idea by *Bonferroni*, we just give the results:

$$\hat{Y}_h \pm Bs(\text{pred})$$

where $B = t(1-\alpha/2g; n-2)$.

2.5 Regression Through the Origin

Sometimes the regression function is known to be linear and to go through the origin at $(0, 0)$. The normal error model for these case is:

$$Y_i = \beta_1 X_i + \epsilon_i$$

The regression function is:

$$\mathbb{E}[Y] = \beta_1 X$$

The least squares estimator of β_1 is

$$Q = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

with respect to β_i. The resulting normal equation is:

$$\sum_{i=1}^{n} X_i(Y_i - b_1 X_i) = 0$$

leading to the point estimator:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

The estimator b_1 is also the maximum likelihood estimator. The value $\hat{Y}i$ for the i-th case is:

$$\hat{Y}_i = b_1 X_i$$

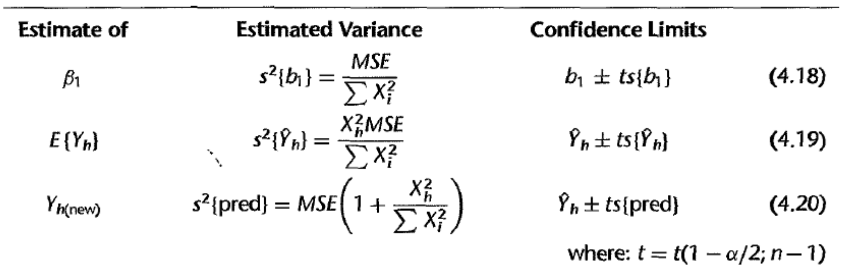
and the i-th residual is defined as usual, i.e.,

$$e_i = Y_i - b_1 X_i$$

An unbiased estimator of the error variance σ^2 for regression model is:

$$s^2 = \text{MSE} = \frac{\sum_{i=1}^{n} e_i^2}{n - 1}$$

The reason for the denominator $n - 1$ is that only one degree of freedom is lost in estimating the single parameter in the regression function. Confidence limits for β_1, $\mathbb{E}[Y_h]$, a new observation $Y_{h(\text{new})}$ for regression model is summarized below (in *Figure* 2.5) In using this type of model, the residuals must be interpreted with care because they do not sum to zero usually, the only constraint on the residual is of the form $\sum_{i=1}^{n} X_i e_i = 0$. thus, in a residual plot the residual will usually not be balanced around the zero line. Another important caution for regression of this type is that SSE may exceed the total sum of square SST. This can occur when the data form a curvilinear pattern or a linear pattern with an intercept away from the origin. hence, the coefficient of determination R^2 maybe negative. Consequently, the coefficient of determination R^2 has no clear meaning for regression model.



Estimate of	Estimated Variance	Confidence Limits	
β_1	$s^2\{b_1\} = \frac{MSE}{\sum X_i^2}$	$b_1 \pm ts\{b_1\}$	(4.18)
$E\{Y_h\}$	$s^2\{\hat{Y}_h\} = \frac{X_h^2 MSE}{\sum X_i^2}$	$\hat{Y}_h \pm ts\{\hat{Y}_h\}$	(4.19)
$Y_{h(\text{new})}$	$s^2\{\text{pred}\} = MSE\left(1 + \frac{X_h^2}{\sum X_i^2}\right)$	$\hat{Y}_h \pm ts\{\text{pred}\}$	(4.20)
		where: $t = t(1 - \alpha/2; n-1)$	

Figure 2.5: ANOVA for Regression Through the Origin

2.6 Effects of Measurements Errors

When random measurements errors are present in the observations on the response variable Y, no problems are created when these errors are uncorrelated and not biased. These measurement errors are simply absorbed in the model error term ϵ. Unfortunately, a different situation holds when the observations on the predictor variable X are subject to measurement errors. If X_i is the true value and X_i^* is the reported value, we define the measurement error δ_i

$$\delta_i = X_i^* - X_i$$

The regression model we would like to study is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

However, we observe only X_i^*, so we must replace the true X_i by X_i^*,

$$Y_i = \beta_0 + \beta_1(X_i^* - \delta_i) + \epsilon_i$$

We can now rewrite it as follows:

$$Y_i = \beta_0 + \beta_1 X_i^* + (\epsilon_i - \beta_1 \delta_i)$$

The predictor variable observation X_i^* is a random variable, which is correlated with the error terms $\epsilon_i - \beta_1 \delta_i$.

If we assume that the response Y and the random predictor variable X^* follow a bivariate normal distribution, then the conditional distribution of the Y_i, i=1,...,n, given X_i^*, $i = 1, ..., n$, are normal and independent, with conditional mean $\mathbb{E}[Y_i|X_i^*] = \beta_0^* + \beta_1^* X_i^*$ and conditional variance $\sigma^2_{Y|X^*}$. Furthermore, it can be shown that $\beta_1^* = \beta_1(\sigma_X^2/(\sigma_X^2 + \sigma_Y^2))$. Hence, the least squares slope estimate from fitting Y on X^* is not an estimate of β_1, but is an estimate of $\beta_1^* \leq \beta_1$. The resulting estimated regression coefficient of β_1^* will be too small on average, with magnitude of the bias dependent upon the relative sizes of σ_X^2 and σ_Y^2. If σ_Y^2 is relative to σ_X^2, then the bias would be small; otherwise, the bias is substantial.

Another approach is to use additional variables that are known to be related to the value of X but not the errors of measurement δ. Such variables are called instrumental variables because they are used as an instrument in studying the relation between X and Y.

Remark 2.6.1 There is one situation where measurement errors in X are no problem. This is discovered by *Berkson*, the situation is the predictor variable is set at a target value.

2.7 Choice of X Levels

Consider the variances of b_0, b_1, $\hat{Y}_h$ and for predicting Y_h(new), we know

$$\begin{aligned}
\sigma^2(b_0) &= \sigma^2\big[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\big], \\
\sigma^2(b_1) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}, \\
\sigma^2(\hat{Y}_h) &= \sigma^2\big[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\big] \\
\sigma^2(\text{pred}) &= \sigma^2\big[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\big]
\end{aligned}$$

If the main purpose of the regression analysis is to estimate the slope β_1, the variance of b_1 is minimized if $\sum_{i=1}^n (X_i - \bar{X})$ is maximized. This is accomplished by using two levels of X, at the two extremes for the scope of the model, and placing half of the observations at each of the two levels. Of course, if one were not sure of the linearity of the regression function, one would hesitant to use only two levels since they would provide no information about possible departures from linearity. If the main purpose is to estimate the intercept β_0, the number and placement of levels doe not affect the variance of b_0 as long as $\bar{X} = 0$. On the other hand, to estimate the mean response or to predict a new observation at the level X_h, the relevant variance is minimized by using X levels so that $\bar{X} = X_h$.

The general advice given by *D.R.Cox* is: use two levels when the object is primarily to examine whether or not the predictor has an effect and in which direction that effect is. Use three levels whenever a description of the response curve by its slope and curvature is likely to be adequate; this should cover most cases. Use four levels if further examination of the shape of the response curve is important. Use more than four levels when it is required to estimate the detailed shape of the response curve, or when the curve is expected to rise to an asymptotic value, or in general to show features not adequately described by slope and curvature. Except in theses last cases it is generally satisfactory to use equally spaced levels with equal number of observations per level.

Chapter 3

Multiple Regression

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i \tag{3.1}$$

where $\beta_0, \beta_1, ..., \beta_{p-1}$ are parameters, X_{i1}, ..., $X_{i,p-1}$ are parameters, $X_{i1}, ..., X_{i,p-1}$ are known constants, ϵ_i are independent $N(0, \sigma^2)$, for $i = 1, ..., n$. In general, the variables $X_1, ..., X_{p-1}$ do not need to represent different predictor variables. The *response function (response surface)* for regression model (3.1) is,

$$\mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \tag{3.2}$$

Thus the general linear regression model with normal error terms implies that the observations Y_i are independent normal variables, with mean mean $\mathbb{E}[Y_i]$ and constant variance σ^2. This general linear model encompasses a vast variety of situations. The predictor can be qualitative predictor variables, the regression model can be polynomial regression, or even log-transformation regression, they can all be turned into linear regression by setting non-linear term as a new variable or a simple transformation. It also includes the case when the predictors interact with each other. Of course, the combinations of them are also concerned.

It should be clear that general linear regression model is not restricted to linear response surfaces. The term linear model really refers to the fact that model is linear in the parameters. We say that a regression model is linear in the parameters when it can be written in the form:

$$Y_i = c_{i0}\beta_0 + c_{i1}\beta_1 + \cdots + c_{i,p-1}\beta_{p-1} + \epsilon_i \tag{3.3}$$

where the terms c_{i0}, c_{i1}, etc., are coefficients involving the predictor variables. As a counter-example of a non-linear regression model is the following:

$$Y_i = \beta_0 \exp\{\beta_1 X_i\} + \epsilon_i \tag{3.4}$$

3.1 General Linear Regression Model in Matrix Form

3.1.1 Notation and Formulation

In order to express the general linear model (3.1) in matrix format, we need to define the following matrices:

$$\mathbf{Y} = [Y_1 \; Y_2 \; \cdots \; Y_n]^\top,$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{pmatrix}$$

$$\beta = [\beta_0 \; \beta_1 \; \cdots \; \beta_{p-1}]^\top, \; \epsilon = [\epsilon_1 \; \epsilon_2 \; \cdots \; \epsilon_n]^\top$$

Then,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{Y}$ is a vector of responses, β is a vector of parameters, $\mathbf{X}$ is a matrix of constants, ϵ is a vector of independent normal random variables with expectation $\mathbb{E}[\epsilon] = 0$ and variance-covariance matrix $\sigma^2(\epsilon) = \sigma^2\mathbf{I}$. Consequently, the random vector $\mathbf{Y}$ has expectation $\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta$ and the variance-covariance matrix of $\mathbf{Y}$ is the same as that of ϵ, $\sigma^2(\mathbf{Y}) = \sigma^2\mathbf{I}$.

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 $\beta_0, ..., \beta_{p-1}$ that minimize Q. Let us denote the vector of the least squares estimated regression coefficients $b_0, ..., b_{p-1}$ as $\mathbf{b}$:

$$\mathbf{b} = [b_0 \; b_1 \; \cdots \; b_{p-1}]^\top$$

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

$$\mathbf{XX'b} = \mathbf{X'Y} \tag{3.5}$$

Thus, the least squares estimators are:

$$\mathbf{b} = (\mathbf{XX'})^{-1}\mathbf{X'Y} \tag{3.6}$$

The method of maximum likelihood leads to the same estimators for normal random error regression model as those obtained by the method of least squares. The likelihood function generalizes directly for multiple regression as follows:

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\big[- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2 \big]$$

Let the vector of the fitted values $\hat{Y}_i$ be denoted by $\hat{\mathbf{Y}}$ and the vector of the residual terms $e_i = Y_i - \hat{\mathbf{Y}}$ be denoted by $\mathbf{e}$:

$$\hat{\mathbf{Y}} = [\hat{Y}_1 \; \hat{Y}_2 \; \cdots \; \hat{Y}_n]^\top, \; \mathbf{e} = [e_1 \; e_2 \; \cdots \; e_n]^\top \tag{3.7}$$

The fitted values are represented by:

$$\hat{\mathbf{Y}} = \mathbf{Xb}$$

and the residual terms by:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb}$$

The vector of the fitted values $\hat{\mathbf{Y}}$ can be expressed in terms of the hat matrix $\mathbf{H}$ as follows:

$$\hat{\mathbf{Y}} = \mathbf{HY}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{XX}')^{-1}\mathbf{X}'$$

Similarly, the vector of the residuals can be expressed as follows:

$$\mathbf{e} = \mathbf{(I\text{-}H)Y}$$

and the variance-covariance matrix of the residuals is:

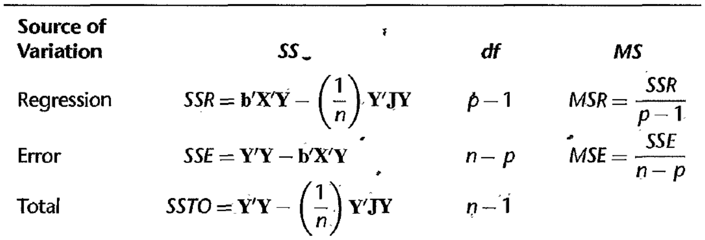
$$\sigma^2(\mathbf{e}) = \sigma^2\mathbf{(I\text{-}H)}$$

which is estimated by: $s^2(\mathbf{e}) = \text{MSE}\mathbf{(I\text{-}H)}$.

3.1.2 ANOVA

The sum of squares for the analysis of variance in matrix terms are:

$$\begin{aligned} \text{SST} &= \mathbf{Y}'\big[\mathbf{I} - \frac{1}{n}\mathbf{J}\big]\mathbf{Y} \\ \text{SSE} &= \mathbf{e}'\mathbf{e} = \mathbf{Y}'\mathbf{(I\text{-}H)Y} \\ \text{SSR} &= \mathbf{Y}'\big[\mathbf{H} - \frac{1}{n}\mathbf{J}\big]\mathbf{Y} \end{aligned}$$



Source of Variation	SS	df	MS
Regression	$SSR = \mathbf{b'X'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	$SSE = \mathbf{Y'Y} - \mathbf{b'X'Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	$SSTO = \mathbf{Y'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$	$n-1$	

Figure 3.1: ANOVA for Multiple Regression Model

where $\mathbf{J}$ is an $n \times n$ matrix of 1's. SST, as usual, has $n-1$ degrees of freedom associated with it. SSE has $n-p$ degrees of freedom since p parameters need to be estimated in the regression function. Finally, SSR has $p-1$ degrees of freedom associated with it, representing the number of X variables $X_1, ..., X_{p-1}$. We have ANOVA table below (in *Figure* 3.1): Notice the expectation of MSE is σ^2, as for simple linear regression. The expectation of MSR is σ^2 plus a quantity that is non-negative, for instance, when $p-1=2$,

$$\mathbb{E}[\text{MSR}] = \sigma^2 + \frac{1}{2}[\beta_1^2 \sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2 + \beta_2^2 \sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2 + 2\beta_1\beta_2 \sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)]$$

Note that if both β_1 and β_2 equal zero, $\mathbb{E}[\text{MSR}] = \sigma^2$, otherwise, it is bigger than σ^2.

3.1.3 F Test for Regression Relation & Coefficient of Multiple Determination

To test whether there is a regression relation between the response variable Y and the set of X variables $X_1, ..., X_{p-1}$, i.e., to choose between the alternatives:

$$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0, \ H_\alpha: \text{ not all } \beta_k \text{ equal to zero,}$$

We use the test statistic:

$$F^* = \frac{\text{MSR}}{\text{MSE}} \tag{3.8}$$

The decision rule to control the Type I error at α is to conclude H_0 if $F^* \leq F(1-\alpha; p-1, n-p)$, otherwise, we conclude H_α.

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

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

It measures the proportionate reduction of total variation in Y associated with the use of the set of X. The coefficient of multiple determination R^2 reduces to the coefficient of simple determination for simple linear regression when $p-1=1$. Just as before, we have $0 \leq R^2 \leq 1$ where R^2 assumes the value 0 when all $b_k = 0$ and the value 1 when all Y observations fall directly on the fitted regression surface.

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 SST 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 R^2_α, adjusts R^2 by dividing each sum of squares by its associated degrees of freedom:

$$R^2_a = 1 - (\frac{n-1}{n-p})\frac{\text{SSE}}{\text{SST}}$$

This adjusted coefficient of multiple determination may actually become smaller when another X variable is introduced into the model, because any decrease in SSE may be more than offset by the loss of a degree of freedom in the denominator $n-p$. The coefficient of multiple correlation R is the positive square root of R^2: $R = \sqrt{R^2}$.

3.2 Inferences about Regression Parameters

The least squares and maximum likelihood estimators $\mathbf{b}$ are unbiased:

$$\mathbb{E}[\mathbf{b}] = \beta$$

The variance-covariance matrix $\sigma^2(\mathbf{b})$ is given by $\sigma^2(\mathbf{b}) = \sigma^2(\mathbf{X'X})^{-1}$ and the estimated variance-covariance matrix $s^2(\mathbf{b}) = \text{MSE}(\mathbf{X'X})^{-1}$. For the normal error regression model, we have

$$\frac{b_k - \beta_k}{s(b_k)} \sim t(n-p),\ k = 0, 1, ..., p-1$$

Hence, the confidence limits for β_k with $1-\alpha$ confidence coefficients are:

$$b_k \pm t(1-\alpha/2; n-p)s(b_k)$$

The tests for β_k are set up in the usual fashion. To test:

$$H_0 :\ \beta_k = 0,\ H_\alpha :\ \beta_l \neq 0$$

We may use the test statistic:

$$t^* = \frac{b_k}{s(b_k)}$$

and the decision rule is that: if $|t^*| \leq t(1-\alpha/2; n-p)$, we conclude H_0.

Remark 3.2.1 The *Bonferroni joint confidence intervals* can be used to estimate several regression coefficients simultaneously as we discussed in last chapter.

For given values $X_1, ..., X_{p-1}$, denote by $X_{h1}, ..., X_{h,p-1}$, the mean response is denoted by $\mathbb{E}[Y_h]$, we define the vector

$$\mathbf{X}_h = [1 \ X_{h1} \ \cdots \ X_{h,p-1}]^\top$$

so that the mean response to be estimated is:

$$\mathbb{E}[Y_h] = \mathbf{X}_h'\beta$$

The estimated response corresponding to $\mathbf{X}_h$, denoted by $\bar{Y}_h$, is

$$\hat{Y}_h = \mathbf{X}_h'\mathbf{b}$$

This estimator is unbiased: $\mathbb{E}[\hat{Y}_h] = \mathbf{X}_h'\beta = \mathbb{E}[Y_h]$ and its variance is:

$$\sigma^2(\hat{Y}_h) = \sigma^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$$

This variance can be expressed as a function of the variance-covariance matrix of the estimated regression coefficients:

$$\sigma^2(\hat{Y}_h) = \mathbf{X}_h'\sigma^2(\mathbf{b})\mathbf{X}_h$$

Thus, the estimated variance is:

$$\sigma^2(\hat{Y}_h) = \mathbf{X}_h's^2(\mathbf{b})\mathbf{X}_h \tag{3.9}$$

The $1-\alpha$ confidence limits for $\mathbb{E}[Y_h]$ are:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s(\hat{Y}_h)$$

The $1-\alpha$ confidence region for the entire regression surface is an extension of the *Working-Hotelling* confidence band for the regression line when there is one predictor variable. Boundary points of the confidence region at $\hat{\mathbf{X}}_h$ are obtained from:

$$\hat{Y}_h \pm Ws(\hat{Y}_h)$$

where $W^2 = F(1-\alpha; p, n-p)$. The confidence coefficient $1-\alpha$ provides assurance that the region contains the entire regression surface over all combinations of values of the X variables.

Remark 3.2.2 The simultaneous confidence intervals for several mean response are along the same line as in the simple regression case.

The $1-\alpha$ prediction limits for a new observation $Y_{h(\text{new})}$ corresponding to X_h, the specified values of the X are

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s(\text{pred})$$

where:

$$s^2(\text{pred}) = \text{MSE} + s^2(\hat{Y}_h) = \text{MSE}(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

and $s^2(\hat{Y}_h)$ is given by (3.9). The prediction of m new observations at X_h will be skipped as its sameness as simple linear regression.

When estimating a mean response or predicting a new observation in multiple regression, one needs to be particularly careful that the estimate or prediction does not fall outside the scope of the model. The danger, of course, is that the model may not be appropriate when it is extended outside the region of the observations. In multiple regression, it is particularly easy to lose track of this region since the level of $X_1, ..., X_{p-1}$ jointly define the region. Thus one cannot merely look at the ranges of each predictor variable (*Figure* 3.2).

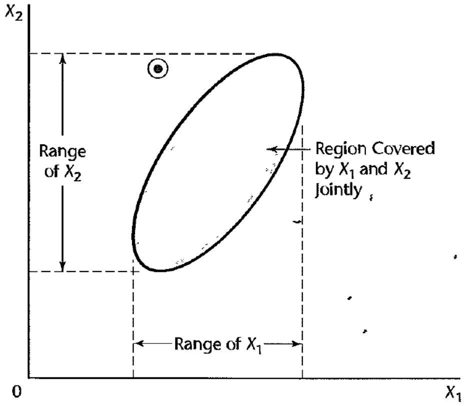


Figure 3.2: Extrapolation

3.3 Diagnostics and Remedial Measures

Box plots, sequence plots, stem-and-leaf plots and dot plots for each of the predictor variables and for the response variable can provide helpful, preliminary univariate information about theses variables. Scatter of the response variable against each predictor 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 predictor variable against each of the other predictor variables are helpful for studying the bivariate relationship among the predict variables and for finding gaps and detecting outliers.

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 $r_{Y1}, r_{Y2}, ..., r_{Y,p-1}$ between

Y and each of the predictor variables, as well as all of the coefficients of simple correlation among the predictor variables.

1. Residual Plots: A plot of the residuals against the fitted values is useful for 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. Similarly, a plot of the residual against time or against some other sequence can provide diagnostic information about possible correlations between the error terms in multiple regression. Box-plot and normal probability plots of the residuals are useful for examining whether the error terms are reasonably normally distributed. In addition, residuals should be plotted against each of the predictor variables. Each of these plots can provide further information about adequacy of the regression function with respect to that predictor variable and about possible variation in the magnitude of the error variance in relation to that predictor variable. Residuals should also be plotted against important predictor variables that were omitted from the model, to see if the omitted variables have substantial additional effects on the response variable that have not yet been recognized in the regression model. Also, residuals should be plotted against interaction terms for potential interaction effects not included in the regression model. A plot of the absolute residuals or the squared residuals against the fitted values is useful for examining the constancy of the variance of the error terms. If non-constancy is detected, a plot of the absolute residuals or the square residuals against each of the predictor variables may identify one or several of the predictor variables to which the magnitude of the error variability is related.

2. Correlation Test for Normality: The correlation test for normality carries forward directly to multiple regression. The expected values of the ordered residuals under normality are calculated and the coefficient of correlation between the residuals and the expected values under normality is then obtained.

Brown-Forsythe Test for Constancy of Error Variance: To conduct the test, we divide the data set into two groups, as for simple linear regression, where one group consists of cases where the level of the predictor variable is relatively low and the other group consists of cases where the level of the predictor variable is relatively high. The Brown-Forsythe test then proceeds as for simple linear regression.

F Test for Lack of Fit: Exactly the same as in the simple linear regression case except adjusting the degrees of freedom correspondingly. We have

$$H_0 : \mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1};$$
$$H_\alpha : \mathbb{E}[Y] \neq \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

the appropriate test statistic is:

$$F^* = \frac{MSLF}{MSPE}$$

and the appropriate decision rule is to conclude H_0 if $F^* \leq F(1-\alpha; c-p; n-c)$.

3.4 Extra Sums of Squares

An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model. Equivalently, one can view an extra sum of squares as measuring the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model. The reason for the equivalence of the marginal reduction in the error sum of squares and the marginal increase in the regression sum of squares is the basic analysis of variance identity:

$$\text{SST} = \text{SSR} + \text{SSE}$$

Since SST measures the variability of the Y_i observations and hence does not depend on the regression model fitted, any reduction in SSE implies an identical increase SSR.

3.4.1 Definition and Decomposition

We define:

$$\text{SSR}(X_1|X_2) = \text{SSE}(X_2) - \text{SSE}(X_1, X_2)$$

or equivalently,

$$\text{SSR}(X_1|X_2) = \text{SSR}(X_1, X_2) - \text{SSR}(X_2)$$

If X_2 is the extra variable, we define

$$\text{SSR}(X_2|X_1) = \text{SSE}(X_1) - \text{SSE}(X_1, X_2)$$

or equivalently,

$$\text{SSR}(X_2|X_1) = \text{SSR}(X_1, X_2) - \text{SSR}(X_1)$$

Extension for three or more variables are straightforward. For example,

$$\text{SSR}(X_3|X_1, X_2) = \text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3)$$

In multiple regression, unlike simple linear regression, we can obtain a variety of decomposition of the regression sum of squares SSR into extra sums of squares. Let us consider the case of two X variables. We begin with identity for variable X_1,

$$\text{SST} = \text{SSR}(X_1) + \text{SSE}(X_1)$$

Replacing $\text{SSE}(X_1)$ by above identity:

$$\text{SST} = \text{SSR}(X_1) + \text{SSR}(X_2|X_1) + \text{SSE}(X_1, X_2)$$

We now make use of the same identity for multiple regression with two X variables, namely,

$$\text{SST} = \text{SSR}(X_1, X_2) + \text{SSE}(X_1, X_2)$$

Solving for $\text{SSE}(X_1, X_2)$, this lead to:

$$\text{SSR}(X_1, X_2) = \text{SSR}(X_1) + \text{SSR}(X_2|X_1)$$

Thus, we have decomposed the regression sum of squares $\text{SSR}(X_1, X_2)$ into two marginal components: (1) $\text{SSR}(X_1)$, measuring the contribution by including X_1 alone in the model and (2) $\text{SSR}(X_2|X_1)$, measuring the additional contribution when X_2 is included, given that X_1 is already in the model. When the regression model contains three X variables, a variety of decomposition of $\text{SSR}(X_1, X_2, X_3)$ can be obtained:

$$\begin{aligned}
\text{SSR}(X_1, X_2, X_3) &= \text{SSR}(X_1) + \text{SSR}(X_2|X_1) + \text{SSR}(X_3|X_1, X_2)\\
\text{SSR}(X_1, X_2, X_3) &= \text{SSR}(X_2) + \text{SSR}(X_3|X_2) + \text{SSR}(X_1|X_2, X_3)\\
\text{SSR}(X_1, X_2, X_3) &= \text{SSR}(X_1) + \text{SSR}(X_2, X_3|X_1)
\end{aligned}$$

ANOVA tables can be constructed containing decompositions of the regression sum of squares into extra sums of squares. Note that each extra sum of squares involving a single extra X variables has associated with it one degree of freedom. The resulting mean squares are constructed as usual. For example, $\text{MSR}(X_2|X_1)$ is obtained as follows:

$$\text{MSR}(X_2|X_1) = \frac{\text{SSR}(X_2|X_1)}{1}$$

Extra sums of squares involving two extra X variables, such as $\text{SSR}(X_2, X_3|X_1)$, have two degree of freedom associated with them. This follows because we can express such an extra sum of squares as a sum of two extra sums of squares, each associated with one degree of freedom. For example, by the definition of the extra sum of squares, we have:

$$\text{SSR}(X_2, X_3|X_1) = \text{SSR}(X_2|X_1) + \text{SSR}(X_3|X_1, X_2)$$

The mean square $\text{MSR}(X_2, X_3|X_1)$ is therefore obtained as follows:

$$\text{MSR}(X_2, X_3|X_1) = \frac{\text{SSR}(X_2, X_3|X_1)}{2}$$

The reason why extra sums of squares are of interest is that they occur in a variety of tests about regression coefficients where the question of concern is whether certain X variables can be dropped from the regression model.

3.4.2 Use of Extra Sums of Squares in Tests for Regression Coefficients

1. Test whether a single $\beta_k = 0$.

When we wish to test whether the term $\beta_k X_k$ can be dropped from a multiple regression model, we are interested in the alternatives:

$$H_0: \ \beta_k = 0, \ H_\alpha: \ \beta_k \neq 0$$

We already know the test statistic: $t^* = \frac{b_k}{s(b_k)}$ is appropriate for this test. Equivalently, we can use the general linear test approach. We now show that this approach involves an extra sum of squares. Let us consider the first-order regression model with three predictor variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \text{ full model}$$

To test alternatives:

$$H_0: \ \beta_3 = 0, \ H_\alpha: \ \beta_3 \neq 0$$

We fit full model and obtain the error of sum of squares SSE(F). We now explicitly show the variables in the full model, as follows:

$$\text{SSE}(F) = \text{SSE}(X_1, X_2, X_3)$$

The degrees of freedom associated with SSE(F) are $d.f._F = n - 4$ since there are four parameters in the regression function for the full model. The reduced model when H_0 holds is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon \text{ reduced model}$$

We next fit this reduced model and obtain:

$$\text{SSE}(R) = \text{SSE}(X_1, X_2)$$

There are $d.f._R = n - 3$ degrees of freedom associated with the reduced model. The general linear test statistic:

$$F^* = \frac{\text{SSE}(R) - \text{SSE}(F)}{d.f._R - d.f._F} / \frac{\text{SSE}}{d.f._F}$$

Here becomes:

$$F^* = \frac{\text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3)}{(n-3) - (n-4)} / \frac{\text{SSE}(X_1, X_2, X_3)}{n-4}$$

Note that the difference between the two error sums of squares in the numerator term is the extra sum of squares:

$$\text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3) = \text{SSR}(X_3 | X_1, X_2)$$

Hence, the general linear test statistic here is:

$$F^* = \frac{\text{SSR}(X_3|X_1, X_2)}{1} / \frac{\text{SSE}(X_1, X_2, X_3)}{n-4} = \frac{\text{MSR}(X_3|X_1, X_2)}{\text{MSE}(X_1, X_2, X_3)}$$

We thus see that the test whether or not $\beta_3 = 0$ is a marginal test, given that X_1, X_2 are already in the model. We also note that the extra sum of squares $\text{SSR}(X_3|X_1, X_2)$ has one degree of freedom associated with it, just as we noted earlier. Test statistic of above shows that we do not need to fit both the full model and the reduced model to use the general linear test approach here. A single computer run can provide a fit of the full model and the appropriate extra sum of squares.

Remark 3.4.1 The F^* test statistic to test whether or not $\beta_3 = 0$ is called a *partial F test statistic* to distinguish it from the F^* statistic for testing whether all $\beta_k = 0$, which is called the *overall F test.*

2. Test whether several $\beta_k = 0$.

In multiple regression we are frequently interested in whether several terms in the regression model can be dropped. For example we may wish to know whether both $\beta_2 X_2$ and $\beta_3 X_3$ can be dropped from the full model. The alternatives are:

$$H_0: \ \beta_2 = \beta_3 = 0, \ H_\alpha: \text{ not both } \beta_2 \text{ and } \beta_3 \text{ equal zero.}$$

With the general linear test approach, the reduced model under H_0 is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \text{ reduced model}$$

and the error sum of squares for the reduced model is:

$$\text{SSE}(R) = \text{SSE}(X_1)$$

This error sum of square has $d.f._R = n-2$ degrees of freedom associated with it. The general linear test statistic thus becomes here:

$$F^* = \frac{\text{SSE}(X_1) - \text{SSE}(X_1, X_2, X_3)}{(n-2) - (n-4)} / \frac{\text{SSE}(X_1, X_2, X_3)}{n-4}$$

Again the difference between the two error sums of squares in the numerator term is an extra sum of squares, namely,

$$\text{SSE}(X_1) - \text{SSE}(X_1, X_2, X_3) = \text{SSR}(X_2, X_3|X_1)$$

Hence, the test statistic becomes:

$$F^* = \frac{\text{SSR}(X_2, X_3|X_1)}{2} / \frac{\text{SSE}(X_1, X_2, X_3)}{n-4} = \frac{\text{MSR}(X_2, X_3|X_1)}{\text{MSE}(X_1, X_2, X_3)}$$

Note that $\text{SSR}(X_2, X_3|X_1)$ has two degrees of freedom associated with it, as we pointed out earlier.

Remark 3.4.2 General linear test statistic for testing whether several X variables can be dropped from the general linear regression model can be expressed in terms of the coefficients of multiple determination for the full and reduced models. Denoting these by R_F^2 and R_R^2, respectively, we have

$$F^* = \frac{R_F^2 - R_R^2}{d.f._R - d.f._F} / \frac{1 - R_F^2}{d.f._F}$$

3. Other Test.

When test about regression coefficients are desired that do not involve testing whether one or several $\beta_k j$ equal zero, extra sums of squares cannot be used and the general linear test approach requires separate fitting of the full and reduced models. For instance, fo the full model containing three X variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon, \text{ full model}$$

we might wish to test:

$$H_0: \ \beta_1 = \beta_2, \ H_\alpha: \ \beta_1 \neq \beta_2$$

The procedure would be to fit the full model and then the reduced model:

$$Y_i = \beta_0 + \beta_c(X_{i1} + X_{i2}) + \beta_3 X_{i3} + \epsilon_i, \text{ reduced model}$$

where β_c denotes the common coefficient for β_1 and β_2 under H_0 and $X_{i1} + X_{i2}$ is the corresponding new X variable. We then use the general F^* test statistic with 1 and $n-4$ degrees of freedom.

Another example where extra sums of squares cannot be used is in the following test for regression model:

$$H_0: \ \beta_1 = 3, \beta_3 = 5, \ H_\alpha: \text{ not both equalities in } H_0 \text{ hold}$$

Here, the reduced model would be:

$$Y_i - 3X_{i1} - 5X_{i3} = \beta_0 + \beta_2 X_{i2} + \epsilon_i, \text{ reduced model}$$

Note the new response variables $Y - 3X_1 - 5X_3$ in the reduced model, since $\beta_1 X_1$ and $\beta_3 X_3$ are known constants under H_0. We the use the general linear test statistic F^* with 2 and $n-4$ degrees of freedom.

3.4.3 Coefficients of Partial Determination

Extra sums of squares are not only useful for test on the regression coefficients of a multiple regression model, but they are also encountered in descriptive measures of relationship called coefficients of partial determination. Recall that the coefficient of multiple determination R^2, measure the proportionate reduction in the variation of Y achieved by the introduction of the entire set of X variables considered in the model. A *coefficient of partial determination*, in contrast, measures the marginal contribution of one X variable when all others are already included in this model.

Let's use two predictor variables to illustration. The model is:

$$Y_0 = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

SSE(X_2) measures the variation in Y when X_2 is included in the model, SSE(X_1, X_2) measures the variation in Y when both X_1 and X_2 are included in the model. Hence, the relative marginal reduction in the variation in Y associated with X_1 when X_2 is already in the model is:

$$\frac{\text{SSE}(X_2) - \text{SSE}(X_1, X_2)}{\text{SSE}(X_2)} = \frac{\text{SSR}(X_1|X_2)}{\text{SSE}(X_2)}$$

This measure is the coefficient of partial determination between Y and X_1, given that X_2 is in the model. We denote this measure by $R^2_{Y1|2}$:

$$R^2_{Y1|2} = \frac{\text{SSR}(X_1|X_2)}{\text{SSE}(X_2)}$$

Thus, $R^2_{Y1|2}$ measures the proportionate reduction in the variation in Y remaining after X_2 is included in the model that is gained by also including X_1 in the model. The square root of a coefficient of partial determination is called a *coefficient of partial correlation.* It is given the same sign as that of the corresponding regression coefficient in the fitted regression function. Coefficients of partial correlation are frequently used in practice, although they do not have as clear a meaning as coefficients of partial determination. One use of partial correlation coefficients is in computer routines for finding the best predictor variable to be selected next for inclusion in the regression model.

3.5 Multicollinearity and its effects

When the predictor variables are correlated among themselves, *inter-correlation or multi-collinearity* among them is said to exist. We shall explore a variety of inter-related problems created by multi-collinearity among the predictor variables. First, however, we examine the situation when the predictor variables are not correlated.

Assume we have $Y \sim (X_1, X_2)$ and the predictor variables X_1 and X_2 are uncorrelated, i.e., $r^2_{12} = 0$, where r^2_{12} denotes the coefficient of simple determination between X_1 and X_2. An important feature of this model is that the regression coefficient for X_1, i.e., b_1, is the same whether only X_1 is included in the model or both predictor variables are included. The same holds for b_2. Thus, when the predictor variables are uncorrelated, the effects ascribed to them by a first order-regression model are the same no matter which of these predictor variables are included in the model.

Another important feature is related to the error sums of squares. In general, when two or more predictor variables are uncorrelated, the marginal contribution of one predictor variable in reducing the error sum of squares when the other predictor variables are in the model is exactly the same as when this predictor variable is in the model alone

When the two predictor variables are perfectly correlated, we assume the following first-order multiple regression function:

$$\mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

And different people may have different answer, e.g.,

$$\begin{aligned}\hat{Y} &= -87 + X_1 + 18X_2 \\ \hat{Y} &= -7 + 9X_1 + 2X_2\end{aligned}$$

Indeed, it can be shown that infinitely many response functions will fit the data set. The reason is that the predictor variables X_1 and X_2 are perfectly related, say, $X_2 = 5 + 0.5X_1$. There are two key implications:

1. the perfect relation between X_1 and X_2 did not inhibit our ability to obtain a good fit to the data;

2. Since many different response functions are provided the same good fit, we cannot interpret any one set of regression coefficients as reflecting the effects of the different predictor variables. Thus, in response function, $b_1 << b_2$ do not imply that X_2 is the key predictor variable and X_1 plays a little role, because response function provides an equally good fit and its regression coefficients have opposite comparative magnitudes.

3.5.1 Effects of Multi-collinearity in General

1. The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean response or prediction of new observations, provided these inferences are made within the region of observations;

2. The counterpart in real life to the many different regression functions providing equally good fits to the data in our idealized example is that the estimated regression coefficients tend to have large sampling variability when the predictor variables are highly correlated. Thus, the estimated regression coefficients tend to vary widely from one sample to the next when the predictor variables are correlated. As a result, only imprecise information may be available about the individual true regression coefficients. Indeed, many of the estimated regression coefficients individually may be statistically not significant even though a definite statistical relation exists between the response variable and the set of predictor variables;

3. The common interpretation of a regression coefficient as measuring the change in thee expected value of the response variable when the given predictor variable is increased by one unit while all other predictor variables are held constant is not fully applicable

when multi-collinearity exists. It may be conceptually feasible to think of varying one predictor variable and holding the other constant, but it may not be possible in practice to do so for predictor variables that are highly correlated.

3.5.2 Effects of Extra Sums of Squares

When predictor variables are correlated, the marginal contribution of any one predictor variable in reducing the error sum of squares varies, depending on which other variables are already in the regression model, just as for regression coefficients. For example, if

$$\begin{aligned}\text{SSR}(X_1) &= 352.27,\\ \text{SSR}(X_1|X_2) &= 3.47,\end{aligned}$$

The reason why $\text{SSR}(X_1|X_2)$ is so small compared with $\text{SSR}(X_1)$ is that X_1 and X_2 are highly correlated with each other and with response variable. Thus, when X_2 is already in the regression model, the marginal contribution of X_1 in reducing the error sum of squares is comparatively small because X_2 contains much of the same information as X_1.

Multi-collinearity also affects the coefficients of partial determination through its effects on the extra sums of squares. For example,

$$\begin{aligned}R^2_{Y1} &= \frac{\text{SSR}(X_1)}{\text{SST}} = 0.71,\\ R^2_{Y1|2} &= \frac{\text{SSR}(X_1|X_2)}{\text{SSE}(X_2)} = 0.03\end{aligned}$$

The reason for the small coefficient of partial determination here is, as we have seen, that X_1 and X_2 are highly correlated with each other and with the response variable. Hence, X_1 provides only relatively limited additional information beyond furnished by X_2.

Remark 3.5.1 The extra sum of squares for a predictor variable after other correlated predictor variables are in the model need not necessarily be smaller than before these other variables are in the model.

3.5.3 Effects on $s(b_k)$, Fitted values and Predictions

The high degree of multi-collinearity among the predictor variables is responsible for the inflated variability of the estimated regression coefficients. The precision of fitted values and prediction, i.e., variance and $\hat{Y}_h$ are not so influenced.

3.5.4 Effects on Simultaneous Test of β_k

A not infrequent abuse in the analysis of multiple regression model is to examine the t^* statistic for each regression coefficient in rutn to decide whether $\beta_k = 0$, $k = 1, ..., p-1$.

Even if a simultaneous inference procedure is used, and often it is not, problems still exist when the predictor variables are highly correlated.

Suppose we wish to test whether $\beta_1 = 0$ and $\beta_2 = 0$. For level of significance 0.05, by *Bonferroni method*, each of the two t tests be conducted with level of significance 0.025. Hence, we need, for example, $t(0.9875; 17) = 2.46$. If both t^* statistics have absolute value that do not exceed 2.46, we would conclude form two separate tests that $\beta_1 = 0$ and $\beta_2 = 0$. Yet the proper F test for $H_0 : \beta_0 = \beta_1 = 0$ would lead to conclusion H_α. The reason for this apparently paradoxical result is that each t^* test is a marginal test, from perspective of the general linear test approach. Thus, a small $\mathrm{SSR}(X_1|X_2)$ here indicates that X_1 does not provide much additional information beyond X_2, which already is in the model. Hence, we are led to the conclusion that $\beta_1 = 0$. Similarly, we are led to conclude $\beta_2 = 0$ because $\mathrm{SSR}(X_2|X_1)$ is small, indicating that X_2 does not provide much more additional information when X_1 is already in the model. But the two tests of the marginal effects of X_1 and X_2 together are not equivalent to testing whether there is a regression relation between Y and two predictor variables. The reason is that the reduced model for each of the separate tests contains the other predictor variable, whereas the reduced model for testing whether $\beta_1 = 0$ and $\beta_2 = 0$ would contain neither predictor variable. The proper F test shows that there is a definite regression relation between Y and X_1 and X_2.

Chapter 4

Building The Regression Model – Model Selection and Validation

4.1 Criterion for Model Selection

Model selection procedures, also known as subset selection or variables selection procedures, have been developed to identify a small group of regression models that are "good" according to a specified criterion. A detailed examination can be then made of limited number of the more promising or "candidate" models, leading to the selection of the final regression model to be employed. This limited number might consist of three to six "good" subsets according to the criteria specified, so the investigator can then carefully study these regression models for choosing the final model.

While many criteria for comparing the regression models have been developed, we will focus on six: R_p^2, $R_{a,p}^2, C_p, \text{AIC}_p, \text{SBC}_p, \text{PRESS}_p$. Before doing so, we will need to develop some notation. We shall denote the number of potential X variables in the pool by $P-1$. We assume throughout this chapter that all regression models contain an intercept term β_0. Hence, the regression function containing all potential X variables contains P parameters, and the function with no X variables contains one parameter β_0.

The number of X variables in a subset will be denoted by $p-1$, as always, so that there are p parameters in the regression function for this subset of X variables. Thus, we have

$$1 \leq p \leq P$$

We will assume that the number of observations exceeds the maximum number of potential parameters:

$$n > p$$

and, indeed, it is highly desirable that n be substantially larger than P, as we noted earlier, so that sound results can be obtained.

4.1.1 R_p^2 or SSE$_p$ Criterion

The R_p^2 criterion calls for the use of the coefficient of multiple determination $\mathbb{R}^2$ in order to identify several "good" subsets of X variables – in other words, subsets for which R^2 is high. We show the number of parameters in the regression model as a subscript of R^2. Thus R_p^2 indicates that there are p parameters, or $p-1$ X variables, in the regression function on which R_p^2 is based.

The R_p^2 criterion is equivalent to using the error sum of squares SSE$_p$ as the criterion. With the SSE$_p$ criterion, subsets for which SSE$_p$ is small are considered "good". The equivalence of the R_p^2 and SSE$_p$ criteria follows from:

$$R_p^2 = 1 - \frac{\text{SSE}_p}{\text{SST}}$$

Since the denominator SST is constant for all possible regression models, R_p^2 varies inversely with SSE$_p$.

The R_p^2 criterion is not intended to identify the subsets that maximize this criterion. We know that R_p^2 can never decrease as additional X variables are included in the model. Hence, R_p^2 will be a maximum when all $P-1$ potential X variables are included in the regression model. The intent in using the R_p^2 criterion is to find the point where adding more X variables is not worthwhile because it leads to a very small increase in R_p^2. Often, this point is reached when only a limited number of X variables is included in the regression model. Clearly, the determination of where diminishing returns set in is a judgmental one.

4.1.2 $R_{a,p}^2$ or MSE$_p$ Criterion

Since R_p^2 does not take account of the number of parameters in the regression model and since $\max(R_p^2)$ can never decrease as p increases, the adjusted coefficient of multiple determination $R_{a,p}^2$ has been suggested as an alternative criterion:

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}_p}{\text{SST}}$$

This coefficient takes the number of parameters in the regression model into account through the degrees of freedom. It can be seen that $R_{a,p}^2$ increases if and only if MSE$_p$ decreases since SST$/(n-1)$ is fixed for the given Y observations. Hence, $R_{a,p}^2$ and MSE$_p$ provide equivalent information. We shall consider here the criterion $R_{a,p}^2$, again showing the number of parameters in the regression model as a subscript of the criterion. The largest $R^2_{a,p}$ for a given number of parameters in the model, $\max(R^2_{a,p})$, can, indeed, decrease as p increases. This occurs when the increase $\max)(R^2_{a,p})$ becomes so small that it is not sufficient to offset the loss of an additional degree of freedom. Users of the $R^2_{a,p}$ criterion seek to find a few subsets for which $R^2_{a,p}$ is at the maximum or so close to the maximum that adding more variables is not worthwhile.

4.1.3 AIC$_p$ and SBC$_p$ Criteria

$R^2_{a,p}$ is model selection criteria that penalize models having large numbers of predictors. Two popular alternatives that also provide penalties for adding predictors are *Akaike's information criterion* and *Schwarz' Bayesian Criterion.* We search for models that have small values of AIC$_p$ or SBC$_p$, where these criteria are given by:

$$\mathrm{AIC}_p = n \ln \mathrm{SSE}_p - n \ln n + 2p$$
$$\mathrm{SBC}_p = n \ln \mathrm{SSE}_p - n \ln n + [\ln n]p$$

Notice that for both of theses measures, the first term is $n\,\mathrm{lnSSE}_p$, which decreases as p increases. The second term is fixed, and the third term increases with th number of parameters, p. Models with small SSE$_p$ will do well by these criteria, as long as the penalties – $2p$ for AIC$_p$ and $[\ln n]p$ for SCB$_p$ – are not too large. If $n \geq 8$ the penalty for SBC$_p$ is larger than that for AIC$_p$; hence the SCB$_p$ criterion tends to favor more parsimonious models.

4.1.4 PRESS$_p$ Criterion

The PRESS$_p$ criterion is a measure of how well the use of the fitted values for a subset model can predict the observed response Y_i. The error sum of squares, $\mathrm{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)$, is also such a measure. The PRESS measure differs from SSE in that each fitted value $\hat{Y}_i$ for the PRESS criterion is obtained by deleting the i-th case from the data set, estimating the regression function for the subset model from the remaining $n-1$ cases, and then using the fitted regression function to obtain the predicted value $\hat{Y}_{i(i)}$ for the i-th case. We use the notation $\hat{Y}_{i(i)}$ now for the fitted value to indicate, by the first subscript i, that is a predicted value for the i-th case and, by the second subscript (i), that the i-th case was omitted when the regression function was fitted.

The PRESS prediction error for the i-th error for the i-th case then is:

$$Y_i - \hat{Y}_{i(i)}$$

and the PRESS$_p$ criterion is the sum of squared prediction errors over all n cases:

$$\mathrm{PRESS}_p = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2$$

Models with small PRESS_p values are considered good candidate models. The reason is that when the prediction errors $Y_i - \hat{Y}_{i(i)}$ are small, so are the squared prediction errors and the sum of the squared prediction errors. Thus, models with small PRESS_p values fit well in the sense of having small prediction errors.

PRESS values can be calculated without requiring n separate regression runs, each time deleting one of the n cases. later on we will explain, we can do everything just in one run.

4.2 Automatic Search Procedures for Model Selection

1. "Best" Subsets Algorithm: Time-saving algorithms have been developed in which the best subsets according to a specified criterion are identified without requiring the fitting of all of the possible subset regression models. In fact, these algorithms require the calculation of only a small fraction of all possible regression models. For instance, if the C_p criterion is to be employed and the five best subsets according to this criterion are to be identified, these algorithms search for the five subsets of X variables with the smallest C_p values using much less computational effort than when all possible subsets are evaluated. These algorithms are called "best" subsets algorithms. Not only do these algorithms provide the best subsets according to the specified criterion, but they often also identify several "good" subsets for each possible number of X variables in the model to given the investigator additional helpful information in making the final selection of the subset of X variables to be employed in the regression.

2. Step-wise Regression Method: In those occasional cases when the pool of potential X variables contains 30 or 40 or even more variables, use of a "best" subsets algorithm may not be feasible. An automatic search procedure that develops the "best" subset of X variables sequentially may then be helpful. The forward stepwise regression procedure is probably the most widely used of the automatic search methods. It was developed to economize on computational efforts, as compared with the various all-possible-regressions procedures. Essentially, this search method develops a sequence of regression models, at each step adding or deleting an X variable. The criterion for adding or deleting X variable can be stated equivalently in terms of error sum of squares reduction, coefficient of partial correlation, t^* statistic, or F^* statistic.

An essential difference between step-wise procedures and the "best" subsets algorithm is that stepwise search procedures end with the identification of a single regression models can be identified as "good" for final consideration. The identification of a single regression model as "best" by the stepwise procedures is a major weakness of these procedures. Experience has shown that each of the stepwise search procedures can sometimes err by identifying a suboptimal regression model as "best". In addition, the identification of a single regression model may hide the fact that several other regression models may also be "good". Finally, the "goodness" of a regression model can only be established by a through examination using a variety of diagnostics.

What then can we do on those occasions when the pool of potential X variables is very large and an automatic search procedure must be utilized? Basically, we should use the subset identified by the automatic search procedure as a starting point for searching for other "good" subsets. One possibility is to treat the number of X variables in the regression model identified by the automatic search procedure as being about the right subset size and then use the "best" subsets procedure for subsets of this and nearby sizes.

3. Forward Stepwise Regression:

1. The stepwise regression routine first fits a simple linear regression model for each of the $P-1$ potential X variables. For each simple linear regression model, the t^* statistic for testing whether or not the slope is zero is obtained by: $t^*_k = \frac{b_k}{s(b_k)}$. The X variable with the largest t^* value is the candidate for first addition. If this t^* value exceeds a predetermined level, or if the corresponding P-value is less than a predetermined α, the X variables is added. Otherwise, the program terminates with no X variable considered sufficiently helpful to enter the regression model;

2. Assume X_7 is the variable entered at step 1. The stepwise regression routine now fits all regression models with two X variables, where X_7 is one of the pair. For each such regression model, the t^* test statistic corresponding to the newly added predictor X_k is obtained. This is the statistic for testing whether or not $\beta_k = 0$ when X_7 and X_k are the variables in the model. The X variable with the largest t^* value is the candidate for addition at the second state;

3. Suppose X_3 is added at the second stage. Now the stepwise regression routine examines whether any of the other X variables already in the model should be dropped. For our illustration, there is at this stage only one other X variable in the model, X_7, so that only one t^* test statistic is obtained. At later stages, there would be a number of these t^* statistics, one for each of the variables in the model besides the one last added. The variable for which this t^* value is smallest is the candidate for deletion;

4. Suppose X_7 is retained so that both X_3 and X_7 are now in the model. The stepwise regression routine now examines which X variable is the next candidate for addition, then examines whether any of the variables already in the model should now be dropped, and so on until no further X variables can either be added or deleted, at which point the search terminates.

4.3 Model Adequacy

4.3.1 Added-Variable Plots

There are several ways to look at an added variable plot. We outline below what an added variable plot is and some of the different ways to approach it.

The added variable plot basically plots the residuals from the regression of the response on a subset of the regressors versus the residuals from the regression of the new regressors on the same subset of regressors. For example our response is Y and we have 4 possible predictors X_1, X_2, X_3 and X_4. We want to know whether X_4 adds anything over and above X_1, X_2 and X_3, which are already in our model. For this one can use an added variable plot. One proceeds as follows for the plot.

1. Fit a regression of Y on X_1, X_2 and X_3 and save the residuals, say e_1;
2. Fit a regression of X_4 on X_1, X_2 and X_3 and save the residuals, say e_2;
3. Plot e_1 against e_2.

An added variable plot is used when you want to know what a particular regressor's contribution is over and above those already present in the model. It also frequently serves as a guide to the right functional form that the said regressor should enter into the regression equation.

The way to look at this is the following. When you regress Y on X_1, X_2 and X_3, what you have left as the residual is the part of Y unexplained by X_1, X_2 and X_3 for which you seek an explanation from X_4. But in a regression setup, X_4 as a whole will not contribute instead what will contribute is the part of X_4 unexplained by X_1, X_2 and X_3. That we can get as the residual of X_4 on X_1, X_2 and X_3. Hence the added variable plot is the true picture on the contribution of X_4 if it is entered.

There is also a *Layman*'s explanation. Suppose we have the regression we talked about previously, namely Y on X_1, X_2, X_3 and X_4. We want to fit a multiple regression to this data, but let us suppose that we unfortunately have forgotten how to do this. Luckily, however we do remember our simple regression. Is there a way to proceed? Yes! Proceed as follows : Fit Y on X_1, and save the residuals, say e_1. Regress these residuals on X_2, and save them say e_2. Regress these on X_3, and save the residuals as e_3. Finally do the same for X_4 as well. Whatever coefficients you got for X_1, X_2, X_3 and X_4 constitute your multiple regression!

If the plot shows a linear band with a non-zero slope. This plot indicates that a linear term in new added variable, here X_4, may ba a helpful addition to the regression model already containing X_1, X_2, X_3. If it shows a curvilinear band, it indicates that the addition of new variable to regression model maybe helpful and suggesting the possible nature of the curvature effect by the patter shown.

4.3.2 Identifying Outlying Y observations – Studentized Deleted Residuals

A basic step in any regression analysis is to determine if the regression model under consideration is heavily influenced by one or a few cases in the data set. Visualization in lower dimension by box plot, stem-and-leaf plots, scatter plots are possible, but in higher dimension we need some refined measures for identifying cases with outlying Y observations.

The detection of outlying or extreme Y observations based on an examination of the residuals has been considered in earlier chapters. We utilized there either the residual e_i:

$$e_i = Y_i - \hat{Y}_i \tag{4.1}$$

or the semi-studentized residuals e_i^*:

$$e_i^* = \frac{e_i}{\sqrt{MSE}} \tag{4.2}$$

We introduce now two refinements to make the analysis of residuals more effective for identifying outlying Y observations. Recall the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

And the fitted value $\hat{\mathbf{Y}}$ can be expressed as:

$$\hat{\mathbf{Y}} = \mathbf{HY}$$

and also

$$\mathbf{e} = \textbf{(I-H)}\mathbf{Y}$$

Then the variance-covariance matrix of the residuals is

$$\sigma^2(\mathbf{e}) = \sigma^2\textbf{(I-H)}$$

Therefore, the variance of residual e_i, denoted by $\sigma^2(e_i)$, is:

$$\sigma^2(e_i) = \sigma^2(1 - h_{ii})$$

where h_{ii} is the i-th element on the main diagonal of the hat matrix, and the covariance between residuals e_i and e_j is:

$$\sigma(e_i, e_j) = \sigma^2(0 - h_{ij}) = -h_{ij}\sigma^2,\ i \neq j$$

where h_{ij} is the element in the i-th element in the i-th row and j-th column of the hat matrix. These variances and covariances are estimated by using MSE as the estimator of the error variance σ^2:

$$s^2(e_i) = \text{MSE}(1 - h_{ii}); \tag{4.3}$$
$$s(e_i, e_j) = -h_{ij}(\text{MSE}),\ i \neq j \tag{4.4}$$

Let's define the ration of e_i to $s(e_i)$, which is called the *studentized residual*

$$r_i = \frac{e_i}{s(e_i)}$$

while the residuals e_i will have substantially different sampling variations if their standard deviations differ markedly, the studentzied-residuals r_i have constant variance(when the model is appropriate). Studentized residuals often are called *internally studentized residuals.*

The second refinement to make residuals more effective for detecting outlying Y observations is to measure the i-th residual $e_i = Y_i - \hat{Y}_i$ when the fitted regression is based on all of the cases except i-th one. The reason for this refinement is that if Y_i is far outlying, the fitted least square regression function based on all cases including the i-th one may be influenced to come close to Y_i, yielding a fitted value $\hat{Y}_i$ near Y_i. In that event, the residual e_i will be small and will not disclose that Y_i is outlying. On the other hand, if the i-th case is excluded before the regression function is fitted, the least squares fitted value $\hat{Y}_i$ is not influenced by the outlying Y_i observation, and the residual for the i-th case will then tend to be larger and therefore more likely to disclose the outlying Y observation.

The procedure then is to delete the i-th case, fit the regression function to the remaining $n - 1$ cases, and obtain the point estimate of the expected value when X levels are those on the i-th case, to be denoted by $\hat{Y}_{i(i)}$. The difference between the actual observed value Y_i and the estimated expected value $\hat{Y}_{i(i)}$ will be denoted by d_i:

$$d_i = Y_i - \hat{Y}_{i(i)}$$

The difference d_i is called the *deleted residual* for the i-th case. An algebraically equivalent expression for d_i that does not require a re-computation of the fitted regression function omitting the i-th case is:

$$d_i = \frac{e_i}{h_{ii}}$$

where e_i is the ordinary residual for the i-th case and h_{ii} the i-th diagonal element in the hat matrix. Thus, deleted residuals will at time identify outlying Y observations when ordinary residuals would not identify these; at other times deleted residuals lead to the same identification as ordinary residuals.

Note that a deleted residual also corresponds to the prediction error for a new observation. There, we are predicting a new $n + 1$ observation from the fitted regression function based on the earlier n case. Modifying the earlier notation for the context of deleted residuals, where $n-1$ cases are used for predicting the new n-th case, we can restate the estimated variance of d_i,

$$s^2(d_i) = \text{MSE}_{(i)} \left(1 + \mathbf{X}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_i\right)$$

where $\mathbf{X}_i$ is the X observations vector for the i-th case, $\text{MSE}_{(i)}$ is the mean square error when the i-th case is omitted in fitting the regression function, and $\mathbf{X}_i$ is the $\mathbf{X}$ matrix with the i-th case deleted. An algebraically equivalent expression for $s^2(d_i)$ is:

$$s^2(d_i) = \frac{\text{MSE}_{(i)}}{1-h_{ii}}$$

It follows that

$$\frac{d_i}{s(d_i)} \sim t(n-p-1)$$

Combining the above two refinements, we utilize for diagnosis of outlying or extreme Y observations the deleted residual d_i and studentize it by dividing it by its estimated standard deviation given. The *studentized deleted residual*, denoted by t_i, therefore is:

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{\text{MSE}_{(i)}(1-h_{ii})}}$$

The studentized deleted residual t_i is also called an *externally studentized residual*, in constrast to the internally studentized residual. We know that each studentized deleted residual follows the t distribution with $n-p-1$ degrees of freedom. The t_i, however, are not independent. Fortunately, t_i can be calculated without having to fit new regression functions each time a different case is omitted. A simple relationship exists between MSE and $\text{MSE}_{(i)}$ and use this relation, we can re-express t_i

$$t_i = e_i\left[\frac{n-p-1}{\text{SSE}(1-h_{ii})-e_i^2}\right]$$

We identify as outlying Y observations those case whose studentized deleted residuals are large in absolute value. In addition, we can conduct a formal test by means of the Bonferroni test procedure of whether the case with the largest absolute studentized deleted residual is an outlier.

4.3.3 Identifying Oultying X Observations and Hidden Extrapolation – Hat Matrix

The diagonal element h_{ii} of the hat matrix have some useful properties. In particular, their values are always between 0 and 1 and their sum is p:

$$0 \le h_{ii} \le 1, \;\; \sum_{i=1}^{n} h_{ii} = p$$

where p is number of regression parameters in the regression function including the intercept term. In addition, it can be shown that h_{ii} is a measure of the distance between the X values for the i-th case and the means of the X values for all n cases. Thus, a large value h_{ii} indicates that the i-th case is distant from the center of all X observations. The diagonal element h_{ii} in this context is called the *leverage* of the i-th case.

A leverage value h_{ii} is usually considered to be large if it is more than twice as large as the mean leverage value, denoted by $\bar{h}$, which is

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n} \tag{4.5}$$

Hence, leverage value greater than $\frac{2p}{n}$ are considered by this rule to indicate outlying cases with regard to their X values.

The hat matrix is also useful after the model has been selected and fitted for determining whether an inference for a mean response or a new observation involves a substantial extrapolation beyond the range of the data. To spot hidden extrapolations, we can utilize the direct leverage calculation for the new set of X values for which inferences are to be made:

$$h_{\text{new,new}} = \mathbf{X}'_(\text{new})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{\text{new}}$$

where $\mathbf{X}_{\text{new}}$ is the vector containing the X values for which an inference about a mean response or a new observation is to be made, and the $\mathbf{X}$ matrix is the one based on the data set used for fitting the regression model. If $h_{\text{new,new}}$ is well within the range of leverage values h_{ii} for the cases in the data set, no extrapolation is involved. On the other hand, if $h_{\text{new,new}}$ is much larger than the leverage values for the cases in the data set, an extrapolation is indicated.

4.3.4 Identifying Influential Cases – DFFITS, Cook's Distance and DFBETAS Measures

After identifying cases that are outlying with respect to their Y values and/or their X values, the next step is to ascertain whether or not these outlying cases are influential. We shall consider a case to be *influential* if its exclusion cause major changes in the fitted regression function. We take up three measures influence that are widely used in practice.

1. Influence on Single Fitted Value – DFFITS A useful measure of the influence that case i has on the fitted value $\hat{Y}_i$ is given by:

$$(\text{DFFITS})_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}}$$

The letter DF stand for the difference between the fitted value $\hat{Y}_i$ for the i-th case when all n cases are used in fitting the regression function and the predicted value $Y_{i(i)}$ for the

i-th case obtained when the i-th case is ommited in fitting the regression function. The denominator is the estimated standard deviation of $\hat{Y}_i$, but it uses the error mean square when the i-th case is omitted in fitting the regression function for estimating the error variance σ^2. The denominator provides a standardization so that the value DFFITS$_i$ for the i-th case represents the number of estimtaed standard deviations of $\hat{Y}_i$ that the fitted value $\hat{Y}_i$ increases or decreases with the inclusion of the i-th case in fitting the regression model. As a guideline for identifying influential cases, we suggest considering a case influential if the absolute value of DFFITS exceeds 1 for small to medium data set and $2\sqrt{p/n}$ for large data sets.

2. Influence on All Fitted Values – Cook's Distance: In contrast to DFFITS measure, which considers the influence of th i-th case on the fitted value $\hat{Y}_i$ for this case, Cook's distance measure considers the influence of the i-th case on all n fitted values. Cook's distance measure, denoted by D_i, is an aggregate influence measure, showing the effect of the i-th case on all n fitted values:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})}{p\text{MSE}}$$

Note that the numerator involves similar differences as in the DFFITS measure, but here each of the n fitted values $\hat{Y}_j$ is compared with the corresponding fitted value $\hat{Y}_{j(i)}$ when the i-th case is deleted in fitting the regression model. These differences are then square and summed, so that the aggregate influence of the i-th case is measured without regard to the signs of the effects. Finally, the denominator serves as a standardizing measure. In matrix terms, Cook's distance measure can be expressed as follows:

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{p\text{MSE}}$$

For interpreting the Cook's distance measure, it has been found useful to related D_i to the $F(p, n-p)$ distribution and ascertain the corresponding percentile value. If the percentile value is less than about 10 or 20 percent, the i-th case has little apparent influence on the fitted values. If, on the other hand, the percentile value is near 50 percent or more, the fitted values obtained with and without the i-th case should be considered to differ substantially, implying that the i-th case has a major influence on the fit of the regression function.

3. Influence on the Regression Coefficients – DFBETAS A measure of the influence of the i-th case on each regression coefficient b_k, $k = 0, 1, ..., p-1$, is the difference between the estimated regression coefficient b_k based on all n cases and the regression coefficient obtained when the i-th case is omitted, to be denoted by $b_{k(i)}$. When this difference is divided by an estimate of the standard deviation of b_k, we obtain the measure DFBETAS:

$$\text{DFBETAS}_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{\text{MSE}_{(i)} c_{kk}}}, \ k = 0, 1, ..., p-1$$

where c_{kk} is the k-th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. DFBETAS value by its sign indicates whether inclusion of a case leads to an increase or a decrease in the estimated regression coefficient, and its absolute magnitude shows the size of the difference relative to the estimated standard deviation of the regression coefficient. A large absolute value is indicative of a large impact of the i-th case on the k-th regression coefficient. As a guideline for identifying influential cases, we recommend considering a case influential if the absolute value of DFBETAS exceeds 1 for small to medium data set and $2/\sqrt{n}$ for large data sets.

4.4 Multi-collinearity Diagnostics – Variance Inflation Factor

Indications of the presence of serious multi-collinearity are given by the following informal diagnostics:

1. Large changes in the estimated regression coefficients when a predictor variable is added or deleted, or when an observation is altered or deleted;
2. Non-significant results in individual tests on the regression coefficients for important predictor variables;
3. Estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical considerations or prior experience;
4. large coefficients of simple correlation between pairs of predictor variables in the correlation matrix r_{XX};
5. Wide confidence intervals for the regression coefficients representing important predictor variables.

A formal method of detecting the presence of multi-collinearity that is widely accepted is the use of VIFs. These factors measure how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables are not linear related.

To understand the significance of VIFs, we begin with the precision of least squares estimated regression coefficients, which is measured by their variance. We know that the variance-covariance matrix of the estimated regression coefficients is:

$$\sigma^2(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

For purposes of measuring the impact of multi-collinearity, it is useful to work with the standardized regression model, which is obtained by transforming the variables by means of the correlation transformation. When the standardized regression model is fitted, the estimated regression coefficients b_k are standardized coefficients that are related to the estimated regression coefficients for the untransformed variables. The variance-covariance matrix of the estimated standardized regression coefficients can be obtainned, which states that the $\mathbf{X}'\mathbf{X}$ matrix for the transformed variables is the correlation matrix of the X variables $\mathbf{r}_{XX}$. Hence, we obtain:

$$\sigma^2(\mathbf{b}_k^*) = (\sigma^*)^2 \mathbf{r}_{XX}$$

where $\mathbf{r}_{XX}$ is the matrix of the pairwise simple correlation coefficients among the X variables and $(\sigma^*)^2$ is the error term variance for the transformed model. Let VIF_k denote the k-th diagonal element of the matrix $\mathbf{r}_{XX}^{-1}$:

$$\sigma^2(b_k^*) = (\sigma^*)^2 (\mathrm{VIF})_k$$

VIF_k is called the variance inflation factor for b_k^*. It can be shown that this variance inflation factor is equal to:

$$\mathrm{VIF}_k = (1 - R_k^2)^{-1}$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the $p-2$ other X variables in the model. Hence, we have

$$\sigma^2(b_k^*) = \frac{(\sigma^*)^2}{1 - R_k^2}$$

The VIF_k is equal to 1 when $R_k^2 = 0$, i.e., when X_k is not linearly related to the other X variables. When $R_k^2 \neq 0$, then VIF_k is greater than 1, indicating an inflated variance for b_k^* as a result of the inter-correlations among the X variables. When X_k has a perfect linear association with the other X variables in the model so that $R_k^2 = 1$, then VIF_k and $\sigma^2(b_k^*)$ are unbounded.

The largest VIF value among all X variables is often used as an indicator of the severity of multi-collinearity. A maximum VIF value in excess of 10 is frequently taken as an indication that multi-collinearity may be unduly influencing the least squares estimates. The mean of the VIF values also provides information about the severity of the multi-collinearity in terms of how far the estimated standardized regression coefficients b_k^* are from the true values β_k^*. Thus, large VIF values result, on the average, in larger differences between the estimated and true standardized regression coefficients.

$$\bar{\mathrm{VIF}} = \frac{\sum_{k=1}^{p-1} (\mathrm{VIF})_k}{p-1}$$

If the value is large than 1, it is indicative of serious multi-collinearity.