REGRESSION ANALYSIS AND INDICATOR VARIABLES

Thesis Submitted in partial fulfillment of the requirements for the award of degree of

Masters of Science
in
Mathematics and Computing

Submitted by
Sweety Arora
Reg. No. – 301403021

Under the guidance of
Dr. Anil Gaur

School of Mathematics
Thapar University
Patiala-147004(PUNJAB)
INDIA
July, 2016

CERTIFICATE

I hereby certify that the dissertation entitled, **"Regression Analysis and Indicator Variables"**, which is being submitted by **Sweety Arora** (Roll No. 301403021), in the partial fulfillment of the requirement for the award of the degree of Master of Science in the School of Mathematics, Thapar University, Patiala, comprises of candidate's own research work carried out under the supervision and guidance of **Dr. Anil Gaur** during the period from January 2016 to June 2016.

The part of the work presented in this dissertation has not been submitted either in part or in full to this or any other University/ Institute for the award of any degree.

Sweety Arora

Sweety Arora
(301403021)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

02/07/2016

Dr. Anil Gaur
Assistant Professor
School of Mathematics
Thapar University, Patiala

Countersigned by:

Dr. A.K. Lal
Associate Professor and Head
School of Mathematics
Thapar University, Patiala

Dr. S.S. Bhatia
Dean of Academic Affairs
Thapar University, Patiala

Acknowledgement

First of all, I would like to thank the almighty for granting perseverance. I would like to express my gratitude to my honorable supervisor **Dr. Anil Gaur, Assistant Professor, SOM, Thapar University, Patiala,** for their patient guidance and support throughout this work. I was truly very fortunate to have the opportunity to work under him as a student. It was both an honor and privilege to work with him. He also provides help in technical writing and presentation style and I found this guidance to be extremely valuable.

I take this opportunity to express my sincere thanks to **Dr. A.K. Lal, Head, SOM, Thapar University, Patiala,** for their valuable support and help without which it would not have been possible for me to complete this work.

I would like to thank my beloved parents for their years of unyielding love and encouragement. They have always wanted the best for me and I admire my parent's determination and sacrifice to put me through college.

Finally, I am also thankful to all my friends who devoted their valuable time and helped me in all possible ways towards successful completion of this work.

Sweety Arora.

Sweety Arora

Patiala

July, 2016

Abstract

Regression analysis is statistical model that is concerned with describing and evaluating the relationship between a given variable known as dependent variable and one more other variable known as independent variable.

The present thesis entitled ***"Regression Analysis and Indicator Variables"***. This exposition comprises four chapters and each chapter is divided into various subsections.

Chapter 1 includes introduction about Bivariate distribution. The main focus is on knowing the nature and relationship between two these variables. In this chapter there are two techniques used for this, one is correlation analysis and other is regression analysis.

In **Chapter 2**, simple linear regression is discussed. The main focus in this chapter is on least squares-fit to estimate the model parameters. This chapters includes definition of simple linear regression, examples, properties of least squares estimators and the fitted regression model, estimation of σ^2, hypothesis testing on model parameters, use of t-tests, testing significance of regression, analysis of variance and coefficient of determination.

In **Chapter 3**, we have discussed multiple linear regression. In this chapter, we have done least squares-estimation of model parameters as described in Chapter 2. This chapters includes basic definition of multiple linear regression, examples, properties of least-square estimators, estimation of σ^2, testing significance of regression, test on individual regression coefficients and coefficient of determination- R^2 and adjusted R^2.

Chapter 4 contains introduction about indicator variables. This chapter includes example of indicator variable, an indicator variable with more than two levels, more than one indicator variable.

Contents

List of Tables

Chapter 1

Introduction

Consider (X, Y) where X exhibits one variable and Y exhibits second variable. If we take sample of size n, then it may look like $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. X and Y are two members of the population. A population or distribution exhibiting two variables jointly is called Bivariate distribution. For example, X represents the age of bride and Y represents the age of bridegroom. Then the distribution of age of bride and bridegroom jointly will be bivariate distribution.

We are interested to know the nature and strength of relationship between the two variables X and Y. It is done with the help of two statistical techniques:

1. Correlation Analysis
2. Regression Analysis

1.1 Correlation Analysis

Correlation analysis is concerned with measuring the strength of linear relationship between two variables and we measure it through the correlation coefficient. One is interested in the degree of correlation between two variables. Correlation tells us about the linear relationship between two variables.

When an increase (or decrease) in one variable results in increase (or decrease) in other variable then the correlation is said to be positive correlation. If an increase

in one variable leads to the decrease in other variable or vice-versa, the correlation is said to be negative. If an increase or decrease in one variable results in no change in other variable, then there is zero correlation.

The correlation between two variables X and Y is measured by Karl Pearson Correlation coefficient or simply correlation coefficient, denoted by r, is defined as

$$\begin{aligned} r &= \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \\ \text{where } Cov(X,Y) &= \sigma_{XY} = \frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y}) \\ Var(X) &= \sigma_X^2 = \frac{1}{n}\sum(x_i-\bar{x})^2 \\ Var(Y) &= \sigma_Y^2 = \frac{1}{n}\sum(y_i-\bar{y})^2 \end{aligned}$$

So correlation coefficient r becomes

$$\begin{aligned} r &= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \\ r &= \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \end{aligned}$$

1.1.1 Properties of correlation coefficient

1. Correlation Coefficient always between -1 and 1 i.e., $-1 \leq r \leq 1$.
2. correlation coefficient between X and Y is independent of change of origin and scale.
3. If two variables X and Y are independent, coefficient of correlation between them will be zero.
4. Coefficient of correlation measures only linear correlation between X and Y.

1.2 Regression Analysis

Regression analysis is used to know the nature of the relationship between two variables, that is, probable form of mathematical relation between X and Y. Regression is also used to predict or estimate the nature of one variable (dependent variable) corresponding to a given value of another variable(independent variable) i.e., to estimate X when Y is given as y_j Or to estimate Y when X is given as x_j.

In scatter diagrams, quite often, it is seen that there is a tendency for the point (x,y) to cluster around some curve, called the curve of regression. If the curve is straight line, it is called line of regression and it tells there is a linear regression among variables.

Chapter 2

Simple linear regression

A model that involves a single regressor variable x which has a relationship with a response y that is a straight line is simple linear regression model. The model is

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.0.1}$$

where the intercept β_0 and slope β_1 are unknown constants and ε is random error disturbance or error. It is assumed that errors have mean zero and unknown variance σ^2. Additionally we also assume that the errors are uncorrelated. That is the value of one error is independent of the value of another error.

It is important to view the regressor x as measured with negligible error and controlled by the data analyst, while the response y is a random variable. This means that there will be a probability distribution for y at each value for x. The mean of this distribution is

$$E(y|x) = \beta_0 + \beta_1 x \tag{2.0.2}$$

and the variance is

$$Var(y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \tag{2.0.3}$$

Hence, the mean of y is approximately a linear function of x although the variance of y is independent of the value of x. Furthermore the responses are uncorrelated because the errors are also uncorrelated.

The parameters β_0 and β_1 are usually known as model regression coefficients. These coefficients have useful and simple interpretation. The slope β_1 may be interpreted as the change in the mean of the distribution of y for unit change in x. The intercept β_0 is the predicted value of mean of the distribution of response y when $x = 0$. β_0 has no practical interpretation if the range of x does not include zero.

2.1 Least Squares Estimation of Parameters

The parameters β_0 and β_1 which are unknown coefficients must be estimated with the help of sample data. Let us assume that we have n pairs of data say $(y_i, x_i),\ i = 1, 2..., n$. These data may result either from an observational study, or from a controlled experiment designed specially to collect the data, or from existing historical records.

We estimate β_0 and β_1 using the popular least squares method, that gives the line which minimizes the sum of the squares of the differences between the straight line and the observations y_i. From equation (2.0.1) we have

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ ; \quad i = 1, 2, 3...., n \tag{2.1.1}$$

Equation (2.0.1) can be viewed as a population regression model while equation (2.1.1) is a sample regression model, which is written in terms of the n pairs of data $(y_1, x_1), (y_2, x_2), ...(y_n, x_n)$. Hence, least squares criterion is

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.1.2}$$

The estimators β_0 and β_1, say $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

and

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = 2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0$$

Simplifying these equations:-

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned}$$

Above equations are called the least-squares normal equations.

The solution to the normal equations is

$$\begin{aligned} \hat{\beta}_0 &= \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \\ \hat{\beta}_0 &= \sum_{i=1}^{n} \frac{y_i}{n} - \hat{\beta}_1 \sum_{i=1}^{n} \frac{x_i}{n} \end{aligned}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.1.3}$$

and

$$\begin{aligned} \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i \\ \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i\right)}{n} + \hat{\beta}_1 \sum_{i=1}^{n} \frac{x_i}{n} \sum_{i=1}^{n} x_i \\ \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2 \right) &= \sum_{i=1}^{n} y_i x_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i \end{aligned}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{2.1.4}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the averages of x_i and y_i respectively. Hence the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in equation (2.1.3) and (2.1.4) are known as the least-squares estimators of the intercept β_0 and slope β_1, respectively. The fitted simple linear regression model is then

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.1.5}$$

Above equation gives a point estimate of the mean of y for a particular x.

Since the the numerator of equation(2.1.4) is the corrected sum of cross products of the x_i and y_i and the denominator is the corrected sum of squares of the x_i, we can write these quantities in a more compact notation as

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} \\ S_{xy} &= \sum_{i=1}^{n} y_i \left[x_i - \frac{\sum_{i=1}^{n} x_i}{n}\right] \\ S_{xy} &= \sum_{i=1}^{n} y_i (x_i - \bar{x}) \end{aligned}$$

Also

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

Hence a convenient way to write equation (2.1.4) is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

The difference between the observed value and the corresponding fitted value, that is y_i and $\hat{y}_i$ respectively is known as a residual. Mathematically the i_{th} residual is given by

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, 2, .., n$$

Residuals plays a vital role in detecting departures from the underlying assumptions and also in investigating the fitted regression model's adequacy.

2.2 Alternate form of model parameters

We know that

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
S_{xy} &= \sum_{i=1}^{n} y_i \left[x_i - \frac{\sum_{i=1}^{n} x_i}{n} \right] = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{1}{n} \sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i}{n} \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i x_i - \bar{x}\bar{y} \\
&= Cov(x, y) \\
S_{xx} &= \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = \sigma_x^2 \\
\therefore \quad \hat{\beta}_1 &= \frac{Cov(x, y)}{\sigma_x^2} \\
&= \frac{Cov(x, y)}{\sigma_x \sigma_y} \frac{\sigma_y}{\sigma_x}
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_1 &= r\frac{\sigma_y}{\sigma_x} \\
\because \quad r &= \frac{Cov(x,y)}{\sigma_x \sigma_y} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_0 &= \bar{y} - r\frac{\sigma_y}{\sigma_x}\bar{x}
\end{aligned}$$

So equation of the regression equation of y on x becomes

$$\begin{aligned}
y &= \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x \\
&= \bar{y} + \hat{\beta}_1 (x - \bar{x}) = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \\
y - \bar{y} &= r\frac{\sigma_y}{\sigma_x}(x - \bar{x})
\end{aligned}$$

This equation used to estimate the value of y when x is given.
Similarly

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

used to predict x when y is given.
Hence there are two regression lines y on x and x on y.

2.3 Example

A rocket motor is manufactured by an igniter propellant and a sustainer propellent together inside a metal housing. The shear strength is related to the age in weeks of the batch of sustainer propellent. Twenty observations on shear strength and the age of corresponding batch of propellent have been collected and are shown in table.

Table 2.1: Data for Example 2.3

Observation i	Shear Strength(psi) y_i	Age of Propellant(weeks) x_i
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.75

To estimate the model parameters, first calculate

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 4677.69 - \frac{71,422.56}{20} = 1106.56$$

and

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 528,492.64 - \frac{(267.25)(42,627.15)}{20} = -41,112.65$$

Therefore we find out that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-41,112.65}{1106.56} = -37.15$$

$$\text{and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2131.3575 - (-37.15)13.3625 = 2627.82$$

Given table shows the Data Fitted values, and residuals for above example:

Table 2.2: Data, Fitted value, and Residuals For example(2.3)

Observed Value, y_i	Fitted Value, $\hat{y}_i$	Residual, e_i
2158.70	2051.94	106.76
1678.15	1745.42	-67.27
2316.00	2330.59	-14.59
2061.30	1996.21	65.09
2207.50	2423.48	-215.98
1708.30	1921.90	-213.60
1784.70	1736.14	48.56
2575.00	2534.94	40.06
2357.90	2349.17	8.73
2256.70	2219.13	37.57
2165.20	2144.83	20.37
2399.55	2488.50	-88.95
1779.80	1698.98	80.82
2336.75	2265.58	71.17
1765.30	1810.44	-45.14
2053.50	1959.06	94.44
2414.40	2404.90	9.50
2200.50	2163.40	37.10
2654.20	2553.52	100.68
1753.70	1829.02	-75.32
$\sum y_i = 42627.15$	$\sum \hat{y}_i = 42627.15$	$\sum e_i = 0.00$

The least squares fit is

$$\hat{y} = 2627.82 - 37.15x$$

we may intercept the slope -37.15 as the average weekly decrease in propellent shear strength due to the age of the propellent. Since the lower limit of the x's is near the origin, the intercept 2627.82 represents the shear strength in batch of propellent immediately following manufacture. Table displays the observed values y_i, and the residuals.

2.4 Properties of least squares estimators and the Fitted Regression Model

The least-squares estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ have many important properties.
$\hat{\beta}_1$ and $\hat{\beta}_0$ are linear combinations of the observations y_i.
For example,

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i \\ \text{where } c_i &= \frac{(x_i - \bar{x})}{S_{xx}} \quad for\ i = 1, 2, ..., n \\ S_{xy} &= \sum_{i=1}^{n} y_i (x_i - \bar{x}) \ \text{ and } \ S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

1. The least-squares estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are the unbiased estimators of model parameters β_0 and β_1 respectively.

$$i.e\ E(\hat{\beta}_1) = \beta_1 \ \text{ and } \ E(\hat{\beta}_0) = \beta_0$$

To Prove: $E(\hat{\beta}_1) = \beta_1$

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\sum_{i=1}^{n} c_i y_i \right) = \sum_{i=1}^{n} c_i E(y_i) \\ &= \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \\ \text{Now} \quad \sum_{i=1}^{n} c_i &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0 \\ \therefore \sum_{i=1}^{n} c_i (x_i - \bar{x}) &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 1 \quad and \quad \sum_{i=1}^{n} c_i x_i = 1 \end{aligned}$$

$$\begin{aligned} \text{So,} \quad E(\hat{\beta}_1) &= \beta_0(0) + \beta_1(1) \\ \Rightarrow E(\hat{\beta}_1) &= \beta_1 \end{aligned} \tag{2.4.1}$$

To Prove $E(\hat{\beta}_0) = \beta_0$

we know that

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \varepsilon \\ \therefore E(y) &= \beta_0 + \beta_1 x \\ y &= \hat{\beta}_0 + \hat{\beta}_1 x + e \\ \beta_0 + \beta_1 x &= E(\hat{\beta}_0) + xE(\hat{\beta}_1) \\ \Rightarrow E(\hat{\beta}_0) &= \beta_0 \end{aligned}$$

2. The variance of $\hat{\beta}_1$ is found as :-

$$Var(\hat{\beta}_1) = Var\big(\sum_{i=1}^{n} c_i y_i\big) = \sum_{i=1}^{n} c_i^2 Var(y_i)$$

We know that the observations y_i are uncorrelated hence the variance of sum is just the sum of variances. The variance of each term in the sum is $c_i^2 Var(y_i)$ and we have already assumed that $Var(y_i) = \sigma^2$.

$$\begin{aligned} \therefore \quad Var(\hat{\beta}_1) &= \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{S_{xx}^2} \\ Var(\hat{\beta}_1) &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

The variance of $\hat{\beta}_0$ is :-

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} Cov(\bar{y}, \hat{\beta}_1) \end{aligned}$$

Now the variance of $\bar{y}$ is just $Var(\bar{y}) = \frac{\sigma^2}{n}$ and the covariance between $\hat{\beta}_1$ and $\bar{y}$ is zero.

Therefore

$$Var(\hat{\beta}_0) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1)$$

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \tag{2.4.2}$$

The least-squares fit have many useful properties:

1. In any regression model, the sum of residuals which contains an intercept β_0 is always zero, that is,

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$$

   This property is followed directly from the first normal equation.

2. The sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i$ or

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

3. The line of least-squares always passes through $(\bar{y}, \bar{x})$ which is the point of intersection.

4. The sum of residuals weighted by the corresponding value of the regressor variables always equals zero, that is

$$\sum_{i=1}^{n} x_i e_i = 0$$

5. The sum of residuals weighted by the corresponding fitted value always equals

zero, that is,

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

6. In case there is a perfect correlation, both the regression lines coincide. That is there is only one regression line.

$$\begin{aligned} y - \bar{y} &= r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \\ \frac{y - \bar{y}}{\sigma_y} &= \pm\frac{x - \bar{x}}{\sigma_x} \\ \text{and} \quad \frac{x - \bar{x}}{\sigma_x} &= \pm\frac{y - \bar{y}}{\sigma_y} \end{aligned}$$

That is both the regression lines reduces to

$$\begin{aligned} \frac{y - \bar{y}}{\sigma_y} &= \pm\frac{x - \bar{x}}{\sigma_x} \\ \Rightarrow \; y - \bar{y} &= \pm\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \\ \text{and} \;\; x - \bar{x} &= \pm\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \end{aligned}$$

7. r is geometric mean of both the regression coefficients.

$$r = \sqrt{\beta_1 xy \beta_1 yx}$$

where $\beta_1 yx$ is slope of regression line of y on x and $\beta_1 xy$ is the slope of regression line of x on y.

8. One of the regression coefficient is greater than unity and then other must be less than unity.

$$\begin{aligned} \text{Let} \;\; \beta_1 yx &> 1 \;\Rightarrow\; \frac{1}{\beta_1 yx} < 1 \\ \text{Also} \;\; r^2 &\leq 1 \;\Rightarrow\; \beta_1 yx \beta_1 xy \leq 1 \\ \Rightarrow \quad \beta_1 xy &\leq \frac{1}{\beta_1 yx} < 1 \Rightarrow \beta_1 xy < 1 \end{aligned}$$

9. Angle between two regression lines :-

For a bivariate data (x, y) the equations of two regression lines are

$$y - \bar{y} = \beta_1 yx(x - \bar{x}) \tag{2.4.3}$$

$$\text{and} \quad x - \bar{x} = \beta_1 xy(y - \bar{y}) \tag{2.4.4}$$

Now slope of equation (2.4.3) and equation (2.4.4) is given by

$$m_1 = \beta_1 yx \quad \text{and} \quad m_2 = \beta_1 xy$$

If θ is the angle between two regression lines, then

$$\begin{aligned} tan\theta &= \frac{m_1 - m_2}{1 + m_1 m_2} = \frac{\beta_1 yx - \beta_1 xy}{1 + \beta_1 yx \beta_1 xy} \\ &= \frac{r\frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + r\frac{\sigma_y}{\sigma_x}\frac{\sigma_y}{\sigma_x}} \end{aligned}$$

$$\begin{aligned} &= \frac{\sigma_x \sigma_y (r^2 - 1)}{r(\sigma_x^2 + \sigma_y^2)} \\ \text{If } \ r &= \pm 1 \\ \text{then } \ tan\theta &= 0 \\ \Rightarrow \theta &= 0 \text{ or } \Pi \end{aligned}$$

In this case, two regression lines either coincides or are parallel to each other but they are not parallel to each other since they passed through the point $(\bar{x}, \bar{y})$.

$$\text{If } \ r = 0, \quad tan\theta = \infty \quad \Rightarrow \quad \theta = \frac{\Pi}{2}$$

If two variables are uncorrelated, then the two regression lines are perpendicular to each other.

2.5 Estimation of σ^2

In addition to estimate β_1 and β_0, an estimate of σ^2 is also required to construct interval estimates pertinent to regression model and test hypothesis. Ideally we would like this estimate not to depend on the adequacy of the fitted model. This is only possible when there are several observations on y for at least one value of x or when prior information concerning σ^2 is available. When this approach cannot be used, the estimate of σ^2 is obtained from the residuals or error sum of squares,

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.5.1}$$

A convenient computing formula for SS_{Res} may be found by substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into above equation and simplifying, yielding

$$\begin{aligned}
SS_{Res} &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2 \\
SS_{Res} &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} \\
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = SS_T
\end{aligned}$$

is just the corrected sum of squares of the response observations, so

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy} \tag{2.5.2}$$

The residuals sum of squares has $n - 2$ degrees of freedom, because two degrees of freedom are associated with the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ involved in obtaining y_i. Unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res} \tag{2.5.3}$$

The quantity MS_{Res} is known as the residual mean square. The square root of $\hat{\sigma}^2$ is sometimes called the standard error of regression, and its units are same as the response y.

2.6 Example: The Rocket Propellent Data

To estimate σ^2 for rocket propellant data in example (2.3), first find

$$\begin{aligned} SS_T &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\ &= 92,547,433.45 - \frac{(42,627.15)^2}{20} = 1,693,737.60 \end{aligned}$$

From equation (2.5.3), the residual sum of squares is

$$\begin{aligned} SS_{Res} &= SS_T - \hat{\beta}_1 S_{xy} \\ &= 166,402.65 \end{aligned}$$

Hence, the estimate of σ^2 is computed from equation (2.5.3) as

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = 9244.59$$

This estimate of σ^2 is model dependent.

2.7 Hypothesis Testing On β_0 and β_1

We are mostly interested in testing the hypothesis and constructing the confidence intervals about the model parameters. Hypothesis testing deals with confidence intervals. These procedures require that we make an additional assumption that the model error ε_i are normally distributed. Thus the complete assumptions are that the errors are normally and independently distributed with mean zero and variance σ^2, abbreviated $NID(0, \sigma^2)$.

2.7.1 Use of t-Tests

Suppose that we want to test the hypothesis that the slope is equal to a constant, say β_{10}. The appropriate hypothesis are

$$\begin{aligned} H_0 &: \beta_1 = \beta_{10} \\ H_1 &: \beta_1 \neq \beta_{10} \end{aligned} \tag{2.7.1}$$

where we have specified a two-sided alternative. Since the errors ε_i are $NID(0,\ \sigma^2)$, the observations y_i are $NID(\beta_0 + \beta_1 x_i,\ \sigma^2)$. since $\hat{\beta}_1$ is linear combination of the observations, so $\hat{\beta}_1$ is distributed normally with mean β_1 and variance $\frac{\sigma^2}{S_{xx}}$. Hence, the statistic

$$Z_0 = \frac{\beta_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$$

is distributed normally with mean zero and variance 1 if the null hypothesis is true. If σ^2 was known, we could use Z_0 to test the hypothesis (2.7.1). Typically σ^2 is not known. We know that MS_{Res} is an unbiased estimator of σ^2. By definition of t-statistic,

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \tag{2.7.2}$$

follows a t_{n-2} distribution if the null hypothesis $H_0 : \beta_1 = \beta_{10}$ is true. The degrees of freedom associated with t_0 are the number of degrees of freedom associated with MS_{Res}. Hence, the ratio t_0 is the test statistic used to test $H_0 : \beta_1 = \beta_{10}$. The test procedure computes t_0 and compares the observed value of t_0 from equation(2.7.2) with the upper $\frac{\alpha}{2}$ percentage point of the t_{n-2} distribution ($t_{\frac{\alpha}{2},\ n-2}$). This procedure rejects the null hypothesis if

$$|t_0| \geq t_{\frac{\alpha}{2}, n-2} \tag{2.7.3}$$

The denominator of the test statistic, t_0, in Eq(2.7.2) is often called the estimated standard error, or more simply, the standard error of the slope. That is,

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{MS_{Res}}{S_{xx}}} \tag{2.7.4}$$

Hence, we mostly see that t_0 is written as

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta_1})} \tag{2.7.5}$$

A similar procedure can be used to test hypothesis about the intercept. To test

$$\begin{aligned} H_0 &: \beta_0 = \beta_{00} \\ H_1 &: \beta_0 \neq \beta_{00} \end{aligned} \tag{2.7.6}$$

We will use the test statistic

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_{00}}{se(\hat{\beta}_0)} \tag{2.7.7}$$

where $se(\hat{\beta}_0) = \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$ is the standard error of the intercept. We reject the null hypothesis $H_0 = \beta_0 = \beta_{00}$ if $|t_0| > t_{\frac{\alpha}{2},\, n-2}$.

2.8 Testing Significance of Regression

A very important special case of the hypothesis in equation(2.7.2) is

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

These hypothesis relate to the significance of regression. Failing to reject $H_0 : \beta_1 = 0$ implies that there is no linear relationship between x and y. Note that this may

imply either that x is of little value in explaining the variation in y and the best estimator of y for any x is $\hat{y} = \bar{y}$ or the true relationship between x and y is not linear. Therefore, failing to reject $H_0 : \beta_1 = 0$ is equivalent to saying that there is no linear relationship between y and x.

Alternatively if $H_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in y. However rejecting $H_0 : \beta_1 = 0$ could mean either that the straight line model is adequate or that even though there is a linear effect of x, better results could be obtained with the addition of higher order polynomial terms in x.

The test procedure for $H_0 : \beta_1 = 0$ may be developed from two approaches. The first approach simply makes use of t statistic in eqn (2.7.4) with $\beta_{10} = 0$; or

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

The hypothesis of significance of regression would be neglected if $|t_0| \geq t_{\frac{\alpha}{2},n-2}$

2.9 Example: The Rocket Propellent Data

We will test for significance of regression in the rocket propellent regression model of example (2.3). The estimate of the slope is $\hat{\beta}_1 = -37.15$ and in example (2.6), we computed the estimate of σ^2 to be $MS_{Res} = \hat{\sigma}^2 = 9244.59$. The standard error of slope is

$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}} = 2.89$$

Hence, the statistic is

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = -12.85$$

If we choose $\alpha = 0.05$, the critical value of t is $t_{0.025,18} = 2.101$. Hence, we would

reject $H_0 : \beta_1 = 0$ and conclude that there is a linear relationship between shear strength and the age of propellent.

2.10 The Analysis of Variance

We may also use an analysis of variance approach to test significance of regression. The analysis of variance is based on a partitioning of total variability in the response variable y. To obtain this partitioning, begin with the identity.

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \tag{2.10.1}$$

Squaring both sides of equation(2.10.1) and summing over all n observations produces

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

Note that third term on the right- hand side of this expression can be written as

$$\begin{aligned} 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\sum_{i=1}^{n}\bar{y}(y_i - \hat{y}_i) \\ &= 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\hat{y}\sum_{i=1}^{n}e_i \end{aligned}$$

Because we know that sum of residuals is always zero and the sum of residuals weighted by the corresponding fitted value $\hat{y}_i$ is also zero. Therefore,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.10.2}$$

The left- hand side of above equation is the corrected sum of squares of the observations, SS_T, which measures the total variability in the observations. The two components of SS_T measure respectively, the amount of variability in the observations y_i accounted for by the regression line and the residual variation left unexplained by

the regression line. We recognize $SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ as the residual or error sum of squares. It is customary to call $SS_{Res} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ the regression or model sum of squares.

Equation (2.10.2) is the fundamental analysis of variance identity for a regression model. Symbolically, we usually write

$$SS_T = SS_R + SS_{Res} \tag{2.10.3}$$

We see that the regression sum of squares may be computed as

$$SS_R = \hat{\beta}_1 S_{xy} \tag{2.10.4}$$

The degree-of-freedom breakdown is determined as follows. The total sum of squares, SS_T has $df_T = n - 1$ degrees of freedom because one degree of freedom is lost as a result of the constraint $\sum_{i=1}^{n}(y_i - \bar{y})$ on the deviations $(y_i - \bar{y})$. The model or regression sum of squares, SS_R has $df_T = 1$ degree of freedom because SS_R is completely determined by one parameter, namely, degrees of freedom because two constraints are imposed on the deviations $(y_i - \bar{y})$ as a result of estimating $\hat{\beta}_0$ and $\hat{\beta}_1$. Note that the degrees of freedom have additive property:

$$\begin{aligned} df_T &= df_R + df_{Res} \\ n - 1 &= 1 + (n - 2) \end{aligned} \tag{2.10.5}$$

We can use the usual analysis of variance F test to test the hypothesis $H_0 : \beta_1 = 0$. If the null hypothesis is true, then SS_R follows chi- square distribution SS_{Res} and SS_R are independent. By the definition of an F statistic:

$$F_0 = \frac{\frac{SS_R}{df_R}}{\frac{SS_{Res}}{df_{Res}}} = \frac{\frac{SS_R}{1}}{\frac{SS_{Res}}{n-1}} = \frac{MS_R}{MS_{Res}} \tag{2.10.6}$$

follows the $F_{1,n-2}$ distribution. Expected values of these mean squares are

$$E(MS_{Res}) = \sigma^2$$
$$E(MS_{Res}) = \sigma^2 + \beta_1^2 S_{xx}$$

2.11 Example: The Rocket Propellent Data

We will test for significance of regression in the model developed in example(2.3) for the rocket propellent data. The fitted model is $\hat{y} = 2627.82 - 37.15x, SS_T = 1,693,737.60$ *and* $S_{xy} = -41,112.65$. The regression sum of squares is computed from equation(2.10.4) as

$$SS_R = \hat{\beta}_1 S_{xy} = (-37.15)(-41,112.65) = 1,527,334.95$$

The analysis of variance is summarized in following table. The computed value of $F_0 = 165.21$,

Table 2.3: Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

Table 2.4: Analysis of Variance Table for Rocket Propellent Regression Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P value
Regression	1,527,334.95	1	1,527,334.95	165.21	1.66×10^{-10}
Residual	166,402.65	18	9,244.59		
Total	1,693,737.60	19			

2.12 Coefficient of Determination

The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \tag{2.12.1}$$

is called the coefficient of determination. Since SS_T is a measure of the variability in y without considering the effect of regressor variable x and SS_{Res} is a measure of variability in y remaining after x has been considered, R^2 is often called the proportion of variation explained by the regressor x.

Because $0 \leq SS_{Res} \leq SS_T$, it follows that $0 \leq R^2 \leq 1$. Values of R^2 that are close to 1 imply that most of the variability in y is explained by the regression model. For the regression model for the rocket propellant data in example 2.3, we have

$$R^2 = \frac{SS_R}{SS_T} = \frac{1,527,334.95}{1,693,737.60} = 0.9018$$

that is, 90.18 percentage of the variability in strength is accounted for by the regression model.

Chapter 3

Multiple linear regression

A regression model with more than one regressor variable is known as multiple regression model. Let us suppose that in a chemical process, the yield in pounds of conversion depends on the catalyst concentration and temperature. This relationship can be described by following multiple regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where x_1 denotes the temperature, x_2 denotes the catalyst concentration and y denotes the yield. The above model is a multiple linear regression model with two regressor variables. Here the term linear is used as the above equation is a linear function of the unknown coefficients β_0, β_1 and β_2.

In general, the relationship between response y and k predictor or regressor variables is formulated as a linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k + \varepsilon \tag{3.0.1}$$

The parameters $\beta_j,\ j = 0, 1, .., k$ are constants referred to as the regression coefficients. These coefficients have useful and simple interpretations. The coefficients β_j may be interpreted as the expected change in the response y produced by a unit change in x_j when all of the other regressor variables $x_i (i \neq j)$ are held constant.

This is the reason that the parameters $\beta_j,\ j = 1, 2, .., k$, are mostly known as partial regression coefficients.

3.1 Least-squares Estimation of the regression coefficients

The least squares method is used to estimate the model regression coefficients β_j in equation (3.0.1), that is we minimize the sum of sqaures of errors. Let us suppose that there are $n > k$ observations and let x_{ij} be the i^{th} observation or level of regressor x_j and y_i be the i^{th} observed response. The error term ε in the model assumed to have mean zero and variance σ^2. Additionally we assume that the errors are uncorrelated.

We assume that the regressors $x_1, x_2, ...x_k$ are fixed variables, measured without error. We may write the regression model corresponding to equation (3.0.1) as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_k x_{ik} + \varepsilon_i \\ y_i &= \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, ..., n \end{aligned} \tag{3.1.1}$$

The least-squares function is

$$\begin{aligned} S(\beta_0, \beta_1, ...\beta_K) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ S(\beta_0, \beta_1, ...\beta_K) &= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{i=1}^{k} \beta_j x_{ij}\right)^2 \end{aligned} \tag{3.1.2}$$

The function S must be minimized with respect to $\beta_0\ , \beta_1\ , ..., \beta_k$. The least-squares

estimators of $\beta_0, \beta_1, ..., \beta_k$ must satisfy

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{i=1}^{k}\hat{\beta}_j x_{ij}\right) = 0 \tag{3.1.3}$$

$$\left.\frac{\partial S}{\partial \beta_j}\right|_{\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{i=1}^{k}\hat{\beta}_j x_{ij}\right)x_{ij} = 0, \quad j = 1, 2, ...k \tag{3.1.4}$$

simplifying these equations we obtain least square normal equations

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{i2} + ...\hat{\beta}_k\sum_{i=1}^{n}x_{ik} &= \sum_{i=1}^{n}y_i \\
\hat{\beta}_0\sum_{i=1}^{n}x_{i1} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^{n}x_{i1}x_{i2} + ...\hat{\beta}_k\sum_{i=1}^{n}x_{i1}x_{ik} &= \sum_{i=1}^{n}x_{i1}y_i \\
&\vdots \\
\hat{\beta}_0\sum_{i=1}^{n}x_{ik} + \hat{\beta}_1\sum_{i=1}^{n}x_{ik}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{ik}x_{i2} + ... + \hat{\beta}_k\sum_{i=1}^{n}x_{ik}^2 &= \sum_{i=1}^{n}x_{ik}y_i
\end{aligned} \tag{3.1.5}$$

There are $p = k + 1$ normal equations, one for each of the unknown regression parameters. The estimators $\beta_j,\ j = 1, 2, .., n$ are least squares estimators because they are the solution to the least- squares method. If multiple regression models are expressed in matrix notation, then it will be more convenient to deal with these models. This allows a very compact display of the data, model and results. The model in matrix notation is given by equation (3.1.1) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.1.6}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}.$$

In general, $\mathbf{X}$ is an $n \times p$ matrix of the levels of the regressor variables, $\mathbf{y}$ is an $n \times 1$ vector of the observations, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors and $\boldsymbol{\beta}$ is a $p \times 1$ vector of the regression coefficients.
we want to find the vector of least-squares estimators, $\hat{\boldsymbol{\beta}}$, which minimizes

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\varepsilon_i^2} = \boldsymbol{\varepsilon^T \varepsilon} = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Note that $S(\boldsymbol{\beta})$ may be expressed as

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta^T}\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta^T}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta^T}\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta^T}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

We know that $\boldsymbol{\beta^T}\mathbf{X}^T\mathbf{y}$ is a 1×1 matrix, or a scaler, therefore its transpose $(\boldsymbol{\beta^T}\mathbf{X}^T\mathbf{y})^T = \mathbf{y}^T\mathbf{X}\boldsymbol{\beta}$ will be the same scaler. The least-squares estimators must satisfy

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$

which simplifies to

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \tag{3.1.7}$$

Above equations are the Least-squares normal equations. These are the matrix analogue of the scaler presentation .

To solve these normal equations, multiply both sides of (3.1.7) by the inverse of

$\mathbf{X}^T\mathbf{X}$. Hence, the least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.1.8}$$

provided that inverse matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ exists. If the regressors are linearly independent, this means that if no column of the $\mathbf{X}$ matrix is a linear combination of the other columns then the $(\mathbf{X}^T\mathbf{X})^{-1}$ matrix will always exist.

It will be more easy to see that the matrix form of the normal equations (3.1.7) is identical to the scaler form (3.1.5). Writing (3.1.7) in detail, we obtain

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \dots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \dots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \dots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

3.2 Example: The delivery Time Data

A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. This service activity includes stocking the machine with products and minor maintenance or housekeeping. The industrial engineer responsible for the study suggested that the two most important variables affecting the delivery time (y) are the number of cases of product stocked (x_1) and the distance walked by the route driver (x_2). The engineer has collected 25 observations on delivery time. We will fit the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

to the delivery time data in following Table.

Table 3.1: Delivery Time Data for Example (3.2)

Observation number	Delivery Time (y)	Number of cases (x_1)	Distance(x_2)
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.37	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

To fit the multiple regression model we first form the $\mathbf{X}$ matrix and $\mathbf{y}$ vector

$$X = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ 1 & 4 & 80 \\ 1 & 6 & 150 \\ 1 & 7 & 330 \\ 1 & 2 & 110 \\ 1 & 7 & 210 \\ 1 & 30 & 1460 \\ 1 & 5 & 605 \\ 1 & 16 & 688 \\ 1 & 10 & 215 \\ 1 & 4 & 255 \\ 1 & 6 & 462 \\ 1 & 9 & 448 \\ 1 & 10 & 776 \\ 1 & 6 & 200 \\ 1 & 7 & 132 \\ 1 & 3 & 36 \\ 1 & 17 & 770 \\ 1 & 10 & 140 \\ 1 & 26 & 810 \\ 1 & 9 & 440 \\ 1 & 8 & 635 \\ 1 & 4 & 150 \end{bmatrix} \qquad y = \begin{bmatrix} 16.68 \\ 11.50 \\ 12.03 \\ 14.88 \\ 13.75 \\ 18.11 \\ 8.00 \\ 17.83 \\ 79.24 \\ 21.50 \\ 40.33 \\ 21.00 \\ 13.50 \\ 19.75 \\ 24.00 \\ 29.00 \\ 15.35 \\ 19.00 \\ 9.50 \\ 35.10 \\ 17.90 \\ 52.32 \\ 18.75 \\ 19.83 \\ 10.75 \end{bmatrix}$$

The $\mathbf{X}'\mathbf{X}$ matrix is

$$X'X = \begin{bmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{bmatrix}$$

and the $\mathbf{X}'\mathbf{y}$ vector is

$$X'Y = \begin{bmatrix} 559.60 \\ 7375.44 \\ 337072 \end{bmatrix}$$

The least-square estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

or

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} &= \begin{bmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{bmatrix}^{-1} \begin{bmatrix} 559.60 \\ 7375.44 \\ 337072 \end{bmatrix} \\ &= \begin{bmatrix} 2.34123115 \\ 1.61590712 \\ 0.01438483 \end{bmatrix} \end{aligned}$$

The least-squares fit(with regression coefficients reported to five decimals) is

$$\hat{y} = 2.34123 + 1.61591 + 0.01438x_2$$

3.3 Properties of least-square Estimators

1. The statistical properties of the least-square estimator $\hat{\boldsymbol{\beta}}$ may be easily demonstrated. Consider first bias

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] \\ \Rightarrow E(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} \end{aligned}$$

   Since $E(\boldsymbol{\varepsilon}) = 0$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$. Thus $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

2. The variance property of $\hat{\boldsymbol{\beta}}$ is expressed by the covariance matrix

$$Cov(\hat{\boldsymbol{\beta}}) = E\{[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})]\}$$

Which is a $p \times p$ symmetric matrix whose j^{th} diagonal element is the variance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ and whose ij^{th} off-diagonal element is the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$. The covariance matrix of $\hat{\beta}$ is

$$Cov(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Therefore, if we let $C = (\mathbf{X}^T\mathbf{X})^{-1}$, the variance of $\hat{\beta}_j$ is $\sigma^2 C_{ij}$ and the covariance between $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ is $\sigma^2 C_{ij}$.

3.4 Estimation of σ^2

We can develop an estimator of σ^2 from the residual sum of squares

$$SS_{Res} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$$

substituting $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, we have

$$\begin{aligned} SS_{Res} &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

Since $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ then this last equation becomes

$$SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \tag{3.4.1}$$

The residual sum of squares has $n - p$ degrees of freedom associated with it since p

parameters are estimated in the regression model. The residual mean square is

$$MS_{Res} = \frac{SS_{Res}}{n-p} \tag{3.4.2}$$

The expected value of MS_{Res} is σ^2, so an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = MS_{Res} \tag{3.4.3}$$

This σ^2 is model independent as noted in simple linear regression.

3.5 Example: The Delivery Time Data

We will estimate the error variance σ^2 for the multiple regression model fit to the soft drink delivery time data in example (3.2). Since

$$\begin{aligned} \mathbf{y}'\mathbf{y} &= \sum_{i=1}^{25} y_i^2 = 18,310.6290 \\ \text{and } \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} &= 18,076.90304 \end{aligned}$$

The residual sum of squares is

$$\begin{aligned} SS_{Res} &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= 18,310.6290 - 18,076.90304 = 233.7260 \end{aligned}$$

Hence, the estimate of σ^2 is the residual mean square

$$\hat{\sigma^2} = \frac{SS_{Res}}{n-p} = \frac{233.7260}{25-3} = 10.6239$$

3.6 Hypothesis Testing in Multiple Linear Regression

Once we have estimated the parameters in the model, we face two immediate questions:

1. What is overall adequacy of the model ?
2. Which specific regressors seem important ?

Several hypothesis testing procedures prove useful for addressing these questions. The formal tests require that our random errors be independent and follow a normal distribution with mean zero and variance σ^2.

3.6.1 Test for Significance of Regression

The test for significance of regression is a test to determine if there is a linear relationship between the response y and any of regressor variables $x_1, x_2, ..., x_k$. This procedure is often thought of as an overall or global test of model adequacy. The appropriate hypothesis are

$$H_o : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \text{ for at least one j}$$

Rejection of this null hypothesis implies that at least one of regressors $x_1, x_2, ..., x_k$ contributes significantly to the model.

The test procedure is a generalization of the analysis of variance used in simple linear regression. The total sum of squares SS_T is partitioned into a sum of squares due to regression, SS_R, and a residual sum of squares, SS_{Res}.
Hence,

$$SS_T = SS_R + SS_{Res}$$

If null hypothesis is true, then $\frac{SS_R}{\sigma^2}$ follows a chi- square distribution, which has the same number of degrees of freedom as number of regressor variables in the model. Also SS_{Res} and SS_R are independent. By the definition of F statistic

$$F_0 = \frac{\frac{SS_R}{k}}{\frac{SS_{Res}}{n-k-1}} = \frac{MS_R}{MS_{Res}}$$

follows the $F_{k,\ n-k-1}$ distribution. We have

$$E(MS_{Res}) = \sigma^2$$
$$E(MS_R) = \sigma^2 + \frac{\boldsymbol{\beta}^{*'}\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}^*}{k\sigma^2}$$

where $\boldsymbol{\beta}^* = (\beta_1, \beta_2, ..., \beta_k)'$ and $\mathbf{X}_c$ is the centered model matrix given by

$$\begin{bmatrix} x_{11} - \bar{x_1} & x_{12} - \bar{x_2} & \dots & x_{1k} - \bar{x_k} \\ x_{21} - \bar{x_1} & x_{22} - \bar{x_2} & \dots & x_{2k} - \bar{x_k} \\ \vdots & \vdots & & \vdots \\ x_{i1} - \bar{x_1} & x_{i2} - \bar{x_2} & \dots & x_{ik} - \bar{x_k} \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x_1} & x_{n2} - \bar{x_2} & \dots & x_{nk} - \bar{x_k} \end{bmatrix}$$

These expected mean squares indicates that if the observed value of F_0 is large, then it is likely that at least one $\beta_j \neq 0$, then F_0 follows a noncentral F distribution with k and $n-k-1$ degrees of freedom and a non centrality parameter of

$$\lambda = \frac{\boldsymbol{\beta}^{*'}\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}^*}{\sigma^2}$$

This non centrality parameter also indicates that if the observed value of F_0 is large, then it is likely that at least one $\beta_j \neq 0$. Therefore, to test the hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$, compute the test statistic F_0 and reject H_0 if

$$F_0 > F_{\alpha,\ k,\ n-k-1}$$

The test procedure is usually summarized in an analysis of variance table such as following table.

A computational formula for SS_R is found by starting with

$$SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \tag{3.6.1}$$

and since

$$SS_T = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Table 3.2: Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	n-1		

We may write the above equation as

$$SS_{Res} = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - \left[\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] \tag{3.6.2}$$

$$\text{or } SS_{Res} = SS_T - SS_R \tag{3.6.3}$$

Therefore, the regression sum of squares is

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{3.6.4}$$

the residual sum of squares is

$$SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \tag{3.6.5}$$

and the total sum of squares is

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{3.6.6}$$

3.7 Example The Delivery Time Data

We will test for significance of regression using the delivery using the delivery time data from Example (3.2). Some of the numerical quantities required are calculated in Example (3.5). Note that

$$\begin{aligned} SS_T &= \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\ &= 18,310.6290 - \frac{(559.60)^2}{25} = 5784.5426 \end{aligned}$$

$$\begin{aligned} SS_R &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \\ &= 18,076.9030 - \frac{(559.60)^2}{25} = 5550.8166 \end{aligned}$$

and

$$\begin{aligned} SS_{Res} &= SS_T - SS_R \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = 233.7260 \end{aligned}$$

The analysis of variance is shown in Table 3.5. To test $H_0 : \beta_1 = \beta_2 = 0$, we calculate the statistic

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{2775.4083}{10.6239} = 261.24$$

Since the P value is very small, we conclude that delivery time is related to delivery volume and/or distance. However, this does not necessarily imply that the relationship found is an appropriate one for predicting delivery time as a function of volume and distance. Further tests of model adequacy are required.

3.8 R^2 and Adjusted R^2

Two other ways to assess the overall adequacy of the model are R^2 and the adjusted R^2, denoted R^2_{Adj}. The R^2 for the multiple regression model for the delivery time data as $R^2 = 0.9596$. R^2 always increases when a regressor is added to the model, regardless of the value of the contribution of that variable. Therefore, it is difficult to judge whether an increase in R^2 is really telling us anything important.

Some regression model builders prefer to use an adjusted R^2 statistic, defined as

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)} \tag{3.8.1}$$

Table 3.3: Test for significance of Regression for Example(3.7)

Source Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P Value
Regression	5550.8166	2	2775.4083	261.24	4.7×10^{-16}
Residual	233.7260	22	10.6239		
Total	5784.5426	24			

Since $SS_{Res}/(n-p)$ is the residual mean square and $SS_T/(n-1)$ is constant regardless of how many variables are in the model, R^2_{Adj} will only increase on adding a variable to the model if the addition of the variable reduces the residual mean square.

3.9 Tests on Individual Regression Coefficients

Once we have determined that at least one of the regressors is important, a logical question becomes which one(s). Adding a variable to a regression model always causes the sum of squares for regression to increase and the residual sum of squares to decrease. We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional regressor in the model. The addition of a regressor also increases the variance of the fitted value $\hat{y}$, so we must be careful to include only regressors that are of real value in explaining the response. Furthermore,

adding an unimportant regressor may increase the residual mean square, which may decrease the usefulness of the model.

The hypothesis for testing the significance of any individual regression coefficient, such as β_j are

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{3.9.1}$$

If $H_0 : \beta_j = 0$ is not rejected, then this indicates that the regressor x_j can be deleted from the model. The test statistic for this hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{3.9.2}$$

where C_{jj} is the diagonal element of $(X'X)^{-1}$ corresponding to $\hat{\beta}_j$. The null hypothesis $H_0 : \beta_j = 0$ is rejected if $|t_0| > t_{\alpha/2,n-k-1}$. This is really a partial or marginal test because the regression coefficient $\hat{\beta}_j$ depends on all of the other regressor variable $x_j (i \neq j)$ that are in the model. Thus, this is a test of the contribution of x_j given the other regressors in the model.

3.10 Example The Delivery Time Data

To illustrate the procedure, consider the delivery time data in Example (3.2). Suppose we wish to assess the value of the regressor variable DISTANCE (x_{12}) given that the regressor CASES(x_1) is in the model. The hypothesis are

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_1 &: \beta_2 \neq 0 \end{aligned}$$

The main diagonal element of $(X'X)^{-1}$ corresponding to β_2 is $C_{22} = 0.00000123$, so

that t statistic (3.9.2) becomes

$$t_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \frac{0.01438}{\sqrt{(10.6239)(0.00000123)}} = 3.98$$

Since $t_{0.025,\ 22} = 2.074$, we reject $H_0 : \beta_2 = 0$ and conclude that the regressor DISTANCE, or x_2, contributes significantly to the model given that CASES, or x_1, is also in the model.

Chapter 4

Indicator Variables

Those variables which are usually employed in regression analysis are quantitative variables, it means, the variables that have a well defined scale of measurement. Variables such as distance, income, pressure and temperature are the quantitative variables. Sometimes it becomes important to use categorical or qualitative variables as predictor variables in regression analysis. Examples of categorical or qualitative variables are employment status (employed or unemployed), operators, sex (male or female) and shifts (day, evening, or night). Generally these qualitative variable have no natural scale of measurement. So, we need to assign a set of levels to the qualitative variables to account for the effect that the variable may have on the response. This can be done through the use of indicator variables. Sometimes indicator variables are also known as dummy variables.

Let us suppose that a mechanical engineer wants to relate the effective life of a cutting tool (y) used on a lathe to the lathe speed in revolutions per minute (x_1) and the type of cutting tool used. The second regressor variable, tool type, is qualitative, and has two levels (e.g., tool types A and B). We are using an indicator variable that takes on the values 0 and 1 to identify the classes of the regressor variable "tool

type". Let

$$x_2 = \begin{cases} 0 & \text{if the observation is from tool type A} \\ 1 & \text{if the observation is from tool type B} \end{cases}$$

The choice of 0 and 1 to identify the levels of a qualitative variable is arbitrary. Any two distinct values for x_2 would be satisfactory, although 0 and 1 are usually best.

Assuming that a first-order model is appropriate, we have

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{4.0.1}$$

To interpret the parameters in this model, let us suppose first tool type A, for which $x_2 = 0$ then the regression model becomes

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \varepsilon \end{aligned} \tag{4.0.2}$$

Hence, the relationship between the tool life and lathe speed for tool type A is a straight line with intercept β_0 and slope β_1 . For tool type B, we have $x_2 = 1$, and

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(1) + \varepsilon \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon \end{aligned} \tag{4.0.3}$$

That means, for tool type B the relationship between tool life and lathe speed will also be a straight line with intercept $\beta_0 + \beta_2$ and slope β_1.

The models (4.0.2) and (4.0.3) describe two parallel regression lines, that is, two lines with different intercepts but a common slope β_1. Also the variance of errors ε is assumed to be the same for both tool types A and B. The parameter β_2 expresses the difference in heights between the two regression lines, that is β_2 is a measure of the difference in mean tool life resulting from changing from tool type A to tool

type B.

We can generalize this approach to qualitative factors with any number of levels. For example, suppose that there are three tool types A, B and C of interest. Then two indicator variables, such as x_2 and x_3, are required to incorporate the three levels of tool type into the model. The levels of the indicator variables are

x_2	x_3	
0	0	if the observation is from tool type A
1	0	if the observation is from tool type B
0	1	if the observation is from tool type C

and the regression model will be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Generally, a qualitative variable with n levels is represented by $n - 1$ indicator variables, each taking on the values 0 and 1.

4.1 Example: Tool Life Data

Twenty observations on tool life and lathe speed are presented in table (4.1). Hence, we will fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

If the observation is from tool type A then the indicator variable $x_2 = 0$ and if the observation is from tool type B then $x_2 = 1$.

Table 4.1: Data, Fitted Values, and Residuals for Example (4.1)

i	y_i(Hours)	x_{i1}(rpm)	Tool Type	$\hat{y}_i$	e_i
1	18.73	610	A	20.7552	-2.0252
2	14.52	950	A	11.7087	2.8113
3	17.43	720	A	17.8284	-0.3984
4	14.54	840	A	14.6355	- 0.0955
5	13.44	980	A	10.9105	2.5295
6	24.39	530	A	22.8838	1.5062
7	13.34	680	A	18.8927	-5.5527
8	22.71	540	A	22.6177	0.0923
9	12.68	890	A	13.3052	-0.6252
10	19.32	730	A	17.5623	1.7577
11	30.16	670	B	34.1630	-4.0030
12	27.09	770	B	31.5023	-4.4123
13	25.40	880	B	28.5775	-3.1755
14	26.05	1000	B	25.3826	0.6674
15	33.49	760	B	31.7684	1.7216
16	35.62	590	B	36.2916	-0.6716
17	26.07	910	B	27.7773	-1.7073
18	36.78	650	B	34.6952	2.0848
19	34.95	810	B	30.4380	4.5120
20	43.67	500	B	38.6862	4.9838

The $\mathbf{X}$ matrix and $\mathbf{y}$ vector for fitting the model are

$$\mathbf{X} = \begin{bmatrix} 1 & 610 & 0 \\ 1 & 950 & 0 \\ 1 & 720 & 0 \\ 1 & 840 & 0 \\ 1 & 980 & 0 \\ 1 & 530 & 0 \\ 1 & 680 & 0 \\ 1 & 540 & 0 \\ 1 & 890 & 0 \\ 1 & 730 & 0 \\ 1 & 670 & 1 \\ 1 & 770 & 1 \\ 1 & 880 & 1 \\ 1 & 1000 & 1 \\ 1 & 760 & 1 \\ 1 & 590 & 1 \\ 1 & 910 & 1 \\ 1 & 650 & 1 \\ 1 & 810 & 1 \\ 1 & 500 & 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 18.73 \\ 14.52 \\ 17.43 \\ 14.54 \\ 13.44 \\ 24.39 \\ 13.34 \\ 22.71 \\ 12.68 \\ 19.32 \\ 30.16 \\ 27.09 \\ 25.40 \\ 26.05 \\ 33.49 \\ 35.62 \\ 26.07 \\ 36.78 \\ 34.95 \\ 43.67 \end{bmatrix}$$

The least squares fit is

$$\hat{y} = 36.986 - 0.027x_1 + 15.004x_2$$

Because two different regression lines are employed to model the relationship between tool life and lathe speed, so we could have fit two separate straight line models initially instead of a single model with an indicator variable. However the single model approach is preferred because the analyst has only one final equation to work with instead of two, a much simpler practical result. Furthermore, because both the straight lines are assumed to have the same slope, it makes sense to combine

the data from both tool types to produce a single estimate of this common parameter. This approach also gives one estimate of the common error variance σ^2 and more residual degrees of freedom than would result from fitting two separate regression lines.

Now consider that we expect the regression lines relating tool life to lathe speed to differ in both slope and intercept. It is possible to model this situation with a single regression equation by using indicator variables. The model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \tag{4.1.1}$$

Comparing Eq.(4.1.1) with Eq.(4.0.1), we note that a cross product between lathe speed x_1 and the indicator variable denoting tool type x_2 has been added to the model. To interpret the parameters in this models, first suppose tool type A, for which $x_2 = 0$. Model (4.1.1) becomes

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \varepsilon \end{aligned} \tag{4.1.2}$$

This is a straight line having intercept β_0 and slope β_1. For tool type B, we have $x_2 = 1$, and

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) + \varepsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \varepsilon \end{aligned} \tag{4.1.3}$$

which is a straight line model with intercept $\beta_0 + \beta_2$ and slope $\beta_1 + \beta_3$. Observe that eqn (4.1.1) defines the two regression lines having different intercepts and slopes. Therefore, the parameter β_2 indicates the change in the intercept associated with changing from tool type A to tool type B (the classes 0 and 1 for the indicator variable x_2), and β_3 reflects the change in the slope associated with changing from tool type A to tool type B.

Fitting model (4.1.1) is equivalent to fitting two different regression equations. With the use of indicator variables, there is an advantage that tests of hypothesis can be performed directly using the extra-sum-of-squares method. For example, to test whether or not the two regression models are identical, we would test

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

If $H_0 : \beta_2 = \beta_3 = 0$ is not rejected, then it means that a single regression model can explain the relationship between tool life and lathe speed. To test that the two regression lines have possibly different intercepts but a common slope, the hypothesis are

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

With the use of model (4.1.1), both regression lines can be fitted and these tests performed with one computer run, provided the program producers the sums of squares $SS_R(\beta_1|\beta_0)$, $SS_R(\beta_2|\beta_0,\ \beta_1)$, and $SS_R(\beta_3|\beta_0,\ \beta_1,\ \beta_2)$.

4.2 Example : The Tool Life Data

we will fit the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

to the tool life in table (4.1). The $\mathbf{X}$ matrix and $\mathbf{y}$ vector for this model are

$$\mathbf{X} = \begin{bmatrix} 1 & 610 & 0 & 0 \\ 1 & 950 & 0 & 0 \\ 1 & 720 & 0 & 0 \\ 1 & 840 & 0 & 0 \\ 1 & 980 & 0 & 0 \\ 1 & 530 & 0 & 0 \\ 1 & 680 & 0 & 0 \\ 1 & 540 & 0 & 0 \\ 1 & 890 & 0 & 0 \\ 1 & 730 & 0 & 0 \\ 1 & 670 & 1 & 670 \\ 1 & 770 & 1 & 770 \\ 1 & 880 & 1 & 880 \\ 1 & 1000 & 1 & 1000 \\ 1 & 760 & 1 & 760 \\ 1 & 590 & 1 & 590 \\ 1 & 910 & 1 & 910 \\ 1 & 650 & 1 & 650 \\ 1 & 810 & 1 & 810 \\ 1 & 500 & 1 & 500 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 18.73 \\ 14.52 \\ 17.43 \\ 14.54 \\ 13.44 \\ 24.39 \\ 13.34 \\ 22.71 \\ 12.68 \\ 19.32 \\ 30.16 \\ 27.09 \\ 25.40 \\ 26.05 \\ 33.49 \\ 35.62 \\ 26.07 \\ 36.78 \\ 34.95 \\ 43.67 \end{bmatrix}$$

The fitted regression model is

$$\hat{y} = 32.775 - 0.021x_1 + 23.971x_2 - 0.012x_1x_2$$

To test the hypothesis that the two regression lines are identical ($H_0 : \beta_2 = \beta_3 = 0$), use the statistic

$$F_0 = \frac{SS_R(\beta_2,\ \beta_3|\beta_1,\ \beta_0)/2}{MS_{Res}}$$

Since

$$\begin{aligned} SS_R(\beta_2,\ \beta_3|\beta_1,\ \beta_0) &= SS_R(\beta_1,\ \beta_2,\ \beta_3\ |\beta_0) - SS_R(\beta_1|\beta_0) \\ &= 1434.112 - 293.005 \\ &= 1141.107 \end{aligned}$$

the test statistic is

$$F_0 = \frac{SS_R(\beta_2,\ \beta_3|\beta_1,\ \beta_0)/2}{MS_{Res}} = \frac{1141.107/2}{8.811} = 64.75$$

4.3 Example : An Indicator Variable with More Than Two Levels

An electric utility is investigating the effect of the size of a single-family house and the type of air conditioning used in the house on the total electricity consumption during warm weather months. Let y denote the total electricity consumption (in kilowatt-hours) during the period June through September and x_1 be the size of the house (square feet of floor space). There are four types of air conditioning systems: (1) no air conditioning, (2)window units, (3) heat pump, and (4) central air conditioning. The four levels of this factor can be modeled by three indicator variables, x_2, x_3, and x_4, defined as follow

Type of Air Conditioning	x_2	x_3	x_4
No air conditioning	0	0	0
Window units	1	0	0
Heat pump	0	1	0
Central air conditioning	0	0	1

The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \tag{4.3.1}$$

If the house has no air conditioning, Eq. (4.3.1) becomes

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

If the house has window units, then

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

If the house has a heat pump, the regression model is

$$y = (\beta_0 + \beta_3) + \beta_1 x_1 + \varepsilon$$

while if the house has central air conditioning, then

$$y = (\beta_0 + \beta_4) + \beta_1 x_1 + \varepsilon$$

Hence model (4.3.1) assumes that the relationship between warm weather electricity consumption and the size of house is linear and the slope is independent of the type of air conditioning system employed. The parameters β_2, β_3, and β_4 modify the height(or intercept) of the regression model for the different types of air conditioning systems. That is β_2, β_3, and β_4 measure the effects of window units, a heat pump, and a central air conditioning system, respectively, compared to no air conditioning. Furthermore other effects can be determined by directly comparing the appropriate regression coefficients. For example, $\beta_3 - \beta_4$ reflects the relative efficiency of a heat pump compared to central air conditioning. Note also the assumption that the variance of energy consumption is independent of the type of air conditioning system used. This assumption may be inappropriate.

In this problem it would seem unrealistic to assume that the slope of the regression function relating mean electricity consumption to the size of the house is independent of the type of air conditioning system. For example, we would expect the mean electricity consumption to increase with the size of the house, but the rate of increase should be different for a central air conditioning system than for window

units for larger houses. That is, there should be interaction between the size of the house and the type of air conditioning system. This can be incorporated into the model by expanding model (4.3.1) to include interaction terms. The resulting model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \varepsilon \qquad (4.3.2)$$

The four regression model corresponding to the four types of air conditioning systems are a follows:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \quad \text{(no air conditioning)} \\ y &= (\beta_0 + \beta_2) + (\beta_1 + \beta_5) x_1 + \varepsilon \quad \text{(window units)} \\ y &= (\beta_0 + \beta_3) + (\beta_1 + \beta_6) x_1 + \varepsilon \quad \text{(heat pump)} \\ y &= (\beta_0 + \beta_4) + (\beta_1 + \beta_7) x_1 + \varepsilon \quad \text{(central air conditioning)} \end{aligned}$$

we must note that the model (4.3.2) implies that each type of air conditioning system can have a separate regression line with a unique slope and intercept.

4.4 Example : More Than One Indicator Variable

Frequently there are many different qualitative variables which must be incorporated into the model. To understand, suppose that in Example (4.1) a second qualitative factor, the type of cutting oil used, must be considered. Suppose that this factor has two levels, we can define the second indicator variable, x_3, as follows:

$$x_3 = \begin{cases} 0 & \text{if low- viscosity oil used} \\ 1 & \text{if medium viscosity oil used} \end{cases}$$

A regression model relating tool life (y) to cutting speed (x_1), tool life (x_2), and

type of cutting oil (x_3) is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{4.4.1}$$

Clearly the slope β_1 of regression model relating tool life to cutting speed is independent of either the type of tool or the type of cutting oil. The intercept of the regression line depends on these factors in an additive fashion.

Several types of interaction effects may be added to the model. For example, assume that we consider interactions between cutting speed and the two qualitative factors, so that model (4.4.1) becomes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon \tag{4.4.2}$$

This implies the following situation

Tool Type	Cutting Oil	Regression Model
A	Low viscosity	$y = \beta_0 + \beta_1 x_1 + \varepsilon$
B	Low viscosity	$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \varepsilon$
A	Medium Viscosity	$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \varepsilon$
B	Medium viscosity	$y = (\beta_0 + \beta_2 + \beta_3 +) + (\beta_1 + \beta_4 + \beta_5) x_1 + \varepsilon$

Note that each combination of tool type and cutting oil results in a different regression line, having different intercepts and slopes. However, the model is still additive with respect to the levels of the indicator variables. This means that, changing from low viscosity to medium viscosity cutting oil changes the the slope by β_5 and intercept by β_3 regardless of the type of tool used.

Consider that we add a cross-product term involving the two indicator variables x_2 and x_3 to the model, resulting in

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_6 x_2 x_3 + \varepsilon \tag{4.4.3}$$

We then have the following

Tool Type	Cutting Oil	Regression Model
A	Low viscosity	$y = \beta_0 + \beta_1 x_1 + \varepsilon$
B	Low viscosity	$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1 + \varepsilon$
A	Medium Viscosity	$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1 + \varepsilon$
B	Medium viscosity	$y = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5)x_1 + \varepsilon$

The addition of cross-product term $\beta_6 x_2 x_3$ in (4.4.3) results in the effect of one indicator variable on the intercept depending on the level of the other indicator variable. Meaning that changing from low viscosity cutting oil to medium viscosity cutting oil changes the intercept by β_3 if the tool type A is used, but if the tool type B is used and same changes in cutting oil changes the intercept by $\beta_3 + \beta_6$. If an interaction term $\beta_7 x_1 x_2 x_3$ were added to the model (4.4.3), then changing from low to medium viscosity cutting oil would have an effect on both the intercept and the slope, which depends on the type of tool used.

Unless prior information is available concerning the anticipated effect of tool type and cutting oil viscosity on tool life, we will have to let the data guide us in selecting the correct form of the model. This may generally be done by testing hypothesis about individual regression coefficients using the partial F-test. For example, testing $H_0 : \beta_6 = 0$ for model (4.4.3) would allow us to discriminate between the two candidate models (4.4.3) and (4.4.2).

Bibliography

[1] DC. Montgomery, E A. Peck, G.G Vining (2001) *"Intoduction To Linear Regression Analysis"*, Wiley-India.

[2] John T. McArthur, Leo H. T. West, *"Multiple regression as a general data analysis technique"* Research in Science education., 4(1), 185-192, 1974.

[3] N R. Draper, Harry Smith (1998) *Applied Regression Analysis*, 3rd ed., Wiley Series in Probability and Statistics.

[4] A.Agresti 2nd Ed. (2007), *"An Introduction to Categorial Data Analysis"*.

[5] B.G. Tabachnick, L.S Fidell (1996), 3rd Ed., *" Using Multivariate Statistics"* Harper Collins College Publishers, New York.

[6] C.R.Rao, H.Toutenburg, Shalabh, C.Heumann,Springer(2008) *"Linear Model and Generalizations"*

[7] F. Anscombe and J. W.Turky (1963), *The Examination and Analysis of Residuals*, Technometrics, pp. 141-160.

[8] William Feller, *An Introduction to Probability Theory and Its Applications, Volume 2*, 2rd ed., John Wiley and Sons, Inc., New York, 1971.

[9] Chakravarti, Laha, and Roy, (1967). *"Handbook of Methods of Applied Statistics, Volume I*, John Wiley and Sons, pp. 392-394.

[10] Evans, Hastings, and Peacock (2000), *"Statistical Distributions"*, 3rd. Ed., John Wiley and Sons.