Statistical Inference and Multivariate Analysis (MA 324)

Class Notes
January – May, 2021

Instructor
Ayon Ganguly
Department of Mathematics
IIT Guwahati

Contents

Chapter 5

Regression Analysis

Most of the contains of the chapter is taken from Montgomery, Peck, and Vining, Introduction to Linear Regression Analysis, *Wiley*, 2003

5.1 Regression and Model Building

Regression analysis is a statistical tool for investigating and modeling the relationship between variables. In fact, regression analysis may be the most widely use statistical technique. This technique is used in all most all the areas of science and technology, social sciences, economics, management.

In a typical problem of regression, we are interested in one particular variable. This variable is called target, response, or dependent variable. We have a set of k other variables which might be useful to model or explain the response. These variables are called predictors, regressors or independent variables. We will use y to denote the response and x_1, x_2, ..., x_k to denote the predictors. Note that the sense in which independent/dependent used in regression is different then that us used for independent/dependent of random variables.

For example, we may be interested in sales of a particular product, sale price of a home, voting preference of a particular voter, or delivery time of bottles of a particular soft drinks. These variables are response. In different problems (*i.e.*, analyzing of different response variables), we may have different predictors. For example, when sales of a particular product is response, predictors may include the price of the product, the prices of the competitor products, etc. Similarly, to model the sale price of a home useful predictors might include lot size, number of bedrooms, number of bathrooms, etc. For voter preference, age, sex, income, party membership, etc. could be considered as predictor. Typically, a regression analysis is used for one (or more) of three purposes:

- Modeling the relationship between predictors and response;
- Prediction of the target variable (forecasting);
- Testing of hypotheses.

We will discuss these purposes in the next sections.

5.2 Simple Linear Regression

Let us start with an example.

Example 5.1 (The Rocket Propellant Data)**.** A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a metal housing. The shear strength of the bond between the two types of propellant is an important quality characteristic. It has been suspected that the shear strength depends on the age of the batch of the sustainer propellant. Therefore, 20 observations on shear strength and age in weeks of the corresponding batch of sustainer propellant are made and given in the following table.

Table 5.1: Propellant Data

Sl. no.	Shear Strength (psi)	Age (in weeks)
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	1799.80	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	21.50

Let us denote the shear strengths by y_i's and ages by x_i's. In any regression analysis plotting of data is very important. We will come back to this with an example. Note that in this case sample correlation coefficient is -0.948. The scatter plot of shear strength versus propellant age is provided in the Figure 5.1. This figure suggests that there is a relationship among shear strength and propellant age. The impression is that the data points generally, but not exactly, fall along a straight line with negative slope.

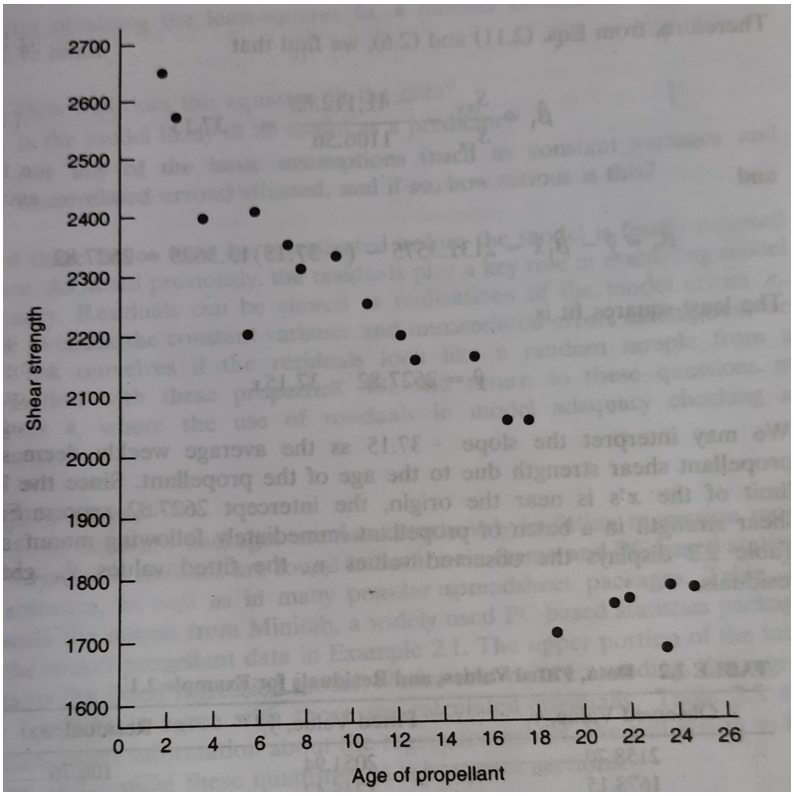

Figure 5.1: Scatter diagram of shear strength versus propellant age

Denoting shear strength by y and propellant age by x, the equation of a straight line relating these two variables may be presented by

$$y = \beta_0 + \beta_1 x,$$

where β_0 is intercept and β_1 is the slope. Note that data points do not fall exactly on the straight line. Therefore, we should modify the equation so that it can take this into account. Thus a more plausible model for shear straight is

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{5.1}$$

where $\varepsilon = y - (\beta_0 + \beta_1 x)$ is the difference between the observed value y and the straight line $(\beta_0 + \beta_1 x)$. Thus, ε is called an error. ||

Definition 5.1 (Simple Linear Regression)**.** *Equation* (5.1) *is called a linear regression model. Also, as the* (5.1) *includes only one predictor, it is called simple linear regression.*

It is convenient to assume that ε is a statistical error, *i.e.*, it is a random variable that accounts for the failure of the model to fit th data exactly. The error may be made up for the effects of other variables on the response like measurement errors. We will also assume that we can fix the value of the predictor x and observe the corresponding value of the response y. As x is fixed, the probabilistic properties of y will be determined by the random error ε. Thus, we make following assumptions:

1. The regressor x is controlled (thus is not a RV) by the analyst and measured with negligible error.

2. The random errors are assumed to have mean zero and variance σ^2. Note that on an average we do not want to commit any error, and hence, the mean zero is a meaningful assumption.

3. We assume that the errors are uncorrelated.

As we assume that the error is a RV, y is also a RV. Thus, for each x, we have a distribution of y. The mean and variance of this distribution are

$$E(y) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x,$$

and

$$Var(y) = Var(\varepsilon) = \sigma^2,$$

respectively. Thus, mean of y is a linear function of x and variance of y does not depend on x. Moreover, as the errors are assumed to be uncorrelated, responses are uncorrelated.

The parameters β_0 and β_1 are called regression coefficients and they have useful practical interpretation in many cases. For example, the slope β_1 is the amount by which the mean of the response variable changes with a unit change in regressor variable. If the range on x includes zero, then the intercept β_0 is the mean of y when $x = 0$. Of course, β_0 does not have any practical interpretation when the range of x does not include zero.

5.2.1 Least Squares Estimation of the Parameters

The method of least squared can be used to estimate regression coefficients β_0 and β_1. This method is described below. Assume that we have n pairs of data point

$$(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$$

on response and predictor, respectively. We estimate the regressions coefficients β_0 and β_1 such that the sum of the squares of the differences between the responses y_i and the straight line $\beta_0 + \beta_1 x_i$ is a minimum. Thus, the least square criterion is

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

Then, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are least squares estimators of β_0 and β_1, respectively, if

$$S\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \min_{\beta_0, \beta_1} S(\beta_0, \beta_1) .$$

Thus, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ must satisfy

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\widehat{\beta}_0, \widehat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = 0 \implies n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{5.2}$$

and

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\widehat{\beta}_0, \widehat{\beta}_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = 0 \implies \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{5.3}$$

Equations (5.2) and (5.3) are called the least squares normal equations (or simply normal equations). The solutions to the normal equations are

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \quad \text{and} \quad \widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{5.4}$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i\right)^2$ and $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} (x_i - \overline{x}) y_i = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$. The difference between the observed value y_i and its' fitted value $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ is called residual. Thus, the ith residual is

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i, \quad \text{for } i = 1, 2, \ldots, n.$$

Example 5.2 (The Rocket Propellant Data)**.** It seems reasonable to fit a linear regression form the Figure 5.1. Therefore, we want to fit the model

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

It can be easily seen that $S_{xx} = 1106.56$ and $S_{xy} = -41112.65$. Thus, using (5.1), $\widehat{\beta}_1 = -37.15$ and $\widehat{\beta}_0 = 2627.82$. The Table 5.2 provides the fitted values $\widehat{y}_i$ and residuals e_i. ||

Table 5.2: Fitted Values and Residuals

Sl. No.	Fitted Value ($\widehat{y}_i$)	Residual (e_i)
1	2051.94	106.76
2	1745.42	-67.27
3	2330.59	-13.59
4	1996.21	65.09
⋮		
19	2553.52	100.68
20	1829.02	-75.32

5.2.2 Properties of Least Squares Estimators

Theorem 5.1. *$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are linear combinations of the observations y_i.*

Proof: Easy to see. For example $\widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$, where $c_i = \frac{x_i - \overline{x}}{S_{xx}}$. □

Theorem 5.2. *$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are UE of the parameters β_0 and β_1, respectively.*

Proof:

$$E\left(\widehat{\beta}_1\right) = E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{\infty} c_i \left(\beta_0 + \beta_1 x_i\right) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1,$$

as $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$. Also,

$$E\left(\widehat{\beta}_0\right) = E\left(\overline{y} - \widehat{\beta}_1 \overline{x}\right) = E\left(\overline{y}\right) - \overline{x} E\left(\widehat{\beta}_1\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i\right) - \beta_1 \overline{x} = \beta_0.$$ □

Theorem 5.3. *The variance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are $\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)$ and $\frac{\sigma^2}{S_{xx}}$, respectively.*

Proof:

$$Var\left(\widehat{\beta}_1\right) = Var\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 Var(y_i) = \sigma^2 \sum_{i=1}^{n} \frac{(x_i-\overline{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

The second equality above is true as y_i are uncorrelated. The third equality holds true as $Var(y_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$.

$$Var\left(\widehat{\beta}_0\right) = Var\left(\overline{y} - \widehat{\beta}_1\overline{x}\right) = Var\left(\overline{y}\right) + \overline{x}^2 Var\left(\widehat{\beta}_1\right) - 2\overline{x}Cov\left(\overline{y},\, \widehat{\beta}_1\right).$$

Now,

$$Var\left(\overline{y}\right) = \frac{\sigma^2}{n}$$

and

$$Cov\left(\overline{y},\, \widehat{\beta}_1\right) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} y_i,\, \sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} \frac{c_i}{n} Var(y_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0.$$

Therefore,

$$Var\left(\widehat{\beta}_0\right) = \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right). \qquad \square$$

Definition 5.2 (Linear Estimator)**.** *An estimator $\widehat{\theta}$ is called a linear estimator of θ if $\widehat{\theta}$ is a linear combination of random observations.*

Definition 5.3 (BLUE)**.** *An estimator $\widehat{\theta}$ is called best linear unbiased estimator (BLUE) of a parameter θ if $\widehat{\theta}$ is a linear estimator and UE of θ and $\widehat{\theta}$ has minimum variance among all linear unbiased estimator of θ.*

Theorem 5.4 (Gauss-Markov Theorem)**.** *The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are best linear unbiased estimator of β_0 and β_1, respectively.*

Proof: Proof is skipped. $\square$

Theorem 5.5. $\sum_{i=1}^{n}(y_i - \widehat{y}_i) = \sum_{i=1}^{n} e_i = 0.$

Proof: It follows directly form the first normal equation. $\square$

Corollary 5.1. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i.$

Theorem 5.6. *The least squared regression line always passes through the centroid (the point $(\overline{x}, \overline{y})$) of the data.*

Proof: The proof is trivial. $\square$

Theorem 5.7. $\sum_{i=1}^{n} x_i e_i = 0.$

Proof: It follows directly form the second normal equation. $\square$

Theorem 5.8. $\sum_{i=1}^{n} \widehat{y}_i e_i = 0.$

Proof: $\sum_{i=1}^{n} \widehat{y}_i e_i = \sum_{i=1}^{n}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) e_i = \widehat{\beta}_0 \sum_{i=1}^{n} e_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i e_i = 0.$ $\square$

5.2.3 Estimation of Error Variance

In the previous couple of subsections, we have discussed estimation of two regression parameters. For many purpose, it is important to estimate the error variance σ^2, which can be estimated unbiasedly as follows. The estimator of σ^2 can be obtained from the residual or error sum of square

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 .$$

It can be shown that

$$E\left(SS_{Res}\right) = \left(n-2\right)\sigma^2.$$

Therefore, an unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}.$$

The quantity MS_{Res} is called the residual mean square. Note that $\widehat{\sigma}^2$ depends on the residual sum of squares, which in turn depends on model assumption. Therefore, any violation of assumptions may have serious damage on the usefulness or $\widehat{\sigma}^2$ as an estimator of σ^2.

A convenient computing formula for SS_{Res} may be found as follows.

$$\begin{aligned}
SS_{Res} &= \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \overline{y} - \widehat{\beta}_1 \overline{x} - \widehat{\beta}_1 x_i\right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 + \widehat{\beta}_1^2 S_{xx} - 2\widehat{\beta}_1 S_{xy} \\
&= SS_T - \widehat{\beta}_1 S_{xy},
\end{aligned}$$

where $SS_T = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$ is called total sum of square.

Example 5.3 (The Rocket Propellant Data)**.** It can be seen that $SS_T = 1693737.60$. Hence $SS_{Res} = 166402.65$. Therefore, the estimate of σ^2 is $\widehat{\sigma}^2 = \frac{166402.65}{20-2} = 9244.59$. ||

5.2.4 Hypothesis Testing on the Slope and Intercept

In this section we will discuss hypothesis testing related to simple linear regression. Note that if $\beta_1 = 0$, then $y = \beta_0 + \varepsilon$. That means that the regression is not meaningful if $\beta_1 = 0$. Therefore, it is one of the fundamental thing to test in case of simple linear regression.

To perform the hypothesis testing and to construct interval estimator, we need an additional assumption, *viz.*, the errors are normally distributed. Thus, the complete set of assumptions are as follows: the errors ε_i are *i.i.d.* RVs having a normal distribution with mean zero and variance σ^2.

Suppose that we want to test if the slope equals to a constant, say β_{10}. Therefore, appropriate hypotheses are $H_0 : \beta_1 = \beta_{10}$ against $H_1 : \beta_1 \neq \beta_{10}$. Based on the assumption on errors, we can see that $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and y_i's are independent. Therefore, $\widehat{\beta}_1$, being a linear combination of y_i's, follows a normal distribution with mean β_1 and variance $\frac{\sigma^2}{S_{xx}}$. Thus, the statistic

$$Z = \frac{\widehat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$$

follows a $N(0, 1)$ distribution if $\beta_1 = \beta_{10}$. If σ is known, we can use Z to to test the hypothesis. However, generally σ is unknown in practice. Hence, Z cannot be use for testing purpose. We can replace σ^2 with its estimator $\widehat{\sigma}^2$. It can be shown that $\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$. Also, MS_{Res} and $\widehat{\beta}_1$ are independent RVs. Therefore, the statistic

$$t = \frac{\widehat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$$

under the null hypothesis $H_0 : \beta_1 = \beta_{10}$. The null hypothesis is rejected at the level α if $|t| > t_{n-2,\,\frac{\alpha}{2}}$.

Similarly, we may want to test $H_0 : \beta_0 = \beta_{00}$ against $H_1 : \beta_0 \neq \beta_{00}$. We can use the test statistic

$$t = \frac{\widehat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}.$$

Under the null hypothesis $H_0 : \beta_0 = \beta_{00}$, t follows a t-distribution with degrees of freedom $n - 2$. Therefore, the null hypothesis may be rejected at level α if $|t| > t_{n-2,\,\frac{\alpha}{2}}$.

Example 5.4 (The Rocket Propellant Data)**.** We will test for the significance of the regression in the rocket propellant data. That means we want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The observed value of the test statistic is

$$t = \frac{-37.15}{\sqrt{9244.59/1106.56}} = -12.85.$$

If we choose $\alpha = 0.05$, $t_{18,\,0.025} = 2.101$. Thus, the null hypothesis $H_0 : \beta_1 = 0$ is rejected and we conclude that there is a linear relationship between shear strength and the age of the propellant. ||

5.2.5 Interval Estimation

In this subsection we will discuss about interval estimation of β_0, β_1, σ^2 and mean response $E(y)$. To construct CI for β_0 and β_1, we can use pivots

$$\frac{\widehat{\beta}_0 - \beta_0}{se(\widehat{\beta}_0)} \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{se(\widehat{\beta}_1)},$$

where $se(\widehat{\beta}_0) = \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$ and $se(\widehat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}}$. Note that both the pivots follow t_{n-2} distribution. Therefore, a $100(1-\alpha)\%$ CI for β_0 is

$$\left[\widehat{\beta}_0 - t_{n-2,\,\frac{\alpha}{2}}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)},\ \widehat{\beta}_0 + t_{n-2,\,\frac{\alpha}{2}}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}\right],$$

and a $100(1-\alpha)\%$ CI for β_1 is

$$\left[\widehat{\beta}_1 - t_{n-2,\,\frac{\alpha}{2}}\sqrt{\frac{MS_{Res}}{S_{xx}}},\ \widehat{\beta}_1 + t_{n-2,\,\frac{\alpha}{2}}\sqrt{\frac{MS_{Res}}{S_{xx}}}\right].$$

Under the assumption of normal errors, it can be shown that

$$\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}.$$

Therefore, we can use it as a pivot to construct CI for σ^2. A $100(1-\alpha)\%$ CI for σ^2 is

$$\left[\frac{(n-2)MS_{Res}}{\chi^2_{n-2,\,\frac{\alpha}{2}}},\ \frac{(n-2)MS_{Res}}{\chi^2_{n-2,\,1-\frac{\alpha}{2}}}\right].$$

Now, we will discuss CI for mean response for a particular value of regressor. In case of forecasting it is meaningful. For example, we may want to know the estimate of mean shear strength of a propellant that is 10 weeks old. In general, let x_0 be the level of the regressor variable for which we wish to estimate the mean response $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 = \mu_{y|x_0}$, say. Note that $\mu_{y|x_0}$ follows a normal distribution as it is a linear combination of y_i's. The mean of $\mu_{y|x_0}$ is $E(y|x_0) = \beta_0 + \beta_1 x_0$ and variance of $\mu_{y|x_0}$ is

$$Var(\mu_{y|x_0}) = Var\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_0\right) = Var\left(\overline{y} + \widehat{\beta}_1(x_0 - \overline{x})\right) = \sigma^2\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right).$$

Here, the second can be found by replacing $\widehat{\beta}_0$ by $\overline{y} - \widehat{\beta}_1\overline{x}$. The third equality holds true as $Cov(\overline{y}, \widehat{\beta}_1) = 0$. Thus, the distribution of

$$\frac{\mu_{y|x_0} - E(y|x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)}}$$

is t with $n-2$ degrees of freedom. Therefore, it can be used as a pivot. A $100(1-\alpha)\%$ CI for mean response at $x = x_0$ is given by

$$\left[\mu_{y|x_0} - t_{n-2,\,\frac{\alpha}{2}}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)},\ \mu_{y|x_0} + t_{n-2,\,\frac{\alpha}{2}}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right)}\right].$$

Example 5.5 (The Rocket Propellant Data)**.** We will construct 95% CI on β_1. The standard error $\widehat{\beta}_1$ is $se(\widehat{\beta}_1) = 2.89$ and $t_{18,0.025} = 2.101$. Therefore, a 95% CI for β_1 is $[-43.22, -31.08]$. Similarly, a 95% CI for mean response at $x = 13.3625$ becomes $[2086.230, 2176.571]$. ||

5.2.6 Prediction of New Observation

An important application of the regression model is prediction of new observations y corresponding to a specified level of the regressor variable x. If x_0 is the value of the regressor variable of interest, then

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

is the point estimate of the new value of the response y_0.

Now, consider interval estimation of this future response y_0. The CI on the mean response at $x = x_0$ is inappropriate as it is interval estimate for mean response, not a probability statement on future observation. Note that the random variable $\psi = y_0 - \widehat{y}_0$ follows a normal distribution with mean zero and variance

$$Var(\psi) = Var(y_0 - \widehat{y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right],$$

as y_0 and $\widehat{y}_0$ are independent. Thus, a $100(1-\alpha)\%$ predictive interval on a future observation at x_0 is

$$\left[\widehat{y}_0 - t_{n-2,\frac{\alpha}{2}}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)},\ \widehat{y}_0 + t_{n-2,\frac{\alpha}{2}}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}\right].$$

Example 5.6 (The Rocket Propellant Data)**.** We will find a 95% prediction interval on a future value of propellant shear strength in a motor made from a batch of sustainer propellant that is 10 weeks old. Using the previous formula, the prediction interval becomes [2048.32, 2464.32]. ||

5.2.7 Coefficient of Determination

The quantity

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

is called the coefficient of determination. Note that

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

where $SS_R = \sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2$. Also, note that the cross-product term

$$\sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)\left(\widehat{y}_i - \overline{y}\right) = 0$$

using normal equation. Since, SS_T is a measure of the variability in y with out considering the effect of the regressor variable x and SS_{Res} is a measure of the variability in y remaining after x has been considered, R^2 is often called the proportion of variation explained by the regressor x. It is clear 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.

Example 5.7 (The Rocket Propellant Data)**.** For the regression model for the rocket propellant data, we have $R^2 = 0.9018$. That means that 90.18% of the variability in strength is accounted for by the regression model. ||

However, the statistic R^2 should be used with caution, since it is always possible to make R^2 large by adding new regressor in the model. But, it may happen that adding new regressor may not improve the quality of the regression significantly. In a matter of fact, adding new regressor may damage the quality of regression.