Recall:

The simple linear regression model and be formed in the matrix form as

$$Y = X\beta + \varepsilon,$$
$$E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2 I,$$
$$E(Y) = X\beta, \quad Var(Y) = \sigma^2 I,$$

where

$$\underset{n\times1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \underset{n\times2}{\mathbf{X}} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad \underset{2\times1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \underset{n\times1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \widehat{\beta} = (X'X)^{-1}X'Y$$

The Fitted Values and Residuals

Let the vector of the fitted values $\hat{Y}_i$ be denoted by $\hat{\mathbf{Y}}$:

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

In matrix notation, we then have:

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \underset{n\times 2}{\mathbf{X}} \; \underset{2\times 1}{\mathbf{b}}$$

because:

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix}$$

Example

For the Toluca Company example, we obtain the vector of fitted values using the matrices

$$\hat{\mathbf{Y}} = \mathbf{Xb} = \begin{bmatrix} 1 & 80 \\ 1 & 30 \\ \vdots & \vdots \\ 1 & 70 \end{bmatrix} \begin{bmatrix} 62.37 \\ 3.5702 \end{bmatrix} = \begin{bmatrix} 347.98 \\ 169.47 \\ \vdots \\ 312.28 \end{bmatrix}$$

It can be easily calculated using R as: fits=x%*%b

Hat Matrix. We can express the matrix result for $\hat{\mathbf{Y}}$

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

or, equivalently:

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \underset{n\times n}{\mathbf{H}}\ \underset{n\times 1}{\mathbf{Y}}$$

where:

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

The matrix H is symmetric and has the special property (called idempotency):

$$HH=H$$

In general, a matrix M is said to be idempotent if MM = M.

Residuals

Let the vector of the residuals $e_i = Y_i - \hat{Y}_i$ be denoted by **e**:

$$\underset{n\times 1}{\mathbf{e}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

In matrix notation, we then have:

$$\underset{n\times 1}{\mathbf{e}} = \underset{n\times 1}{\mathbf{Y}} - \underset{n\times 1}{\hat{\mathbf{Y}}} = \underset{n\times 1}{\mathbf{Y}} - \underset{n\times 1}{\mathbf{Xb}}$$

For the Toluca Company example, we obtain the vector of the residuals by using the results

$$\mathbf{e} = \begin{bmatrix} 399 \\ 121 \\ \vdots \\ 323 \end{bmatrix} - \begin{bmatrix} 347.98 \\ 169.47 \\ \vdots \\ 312.28 \end{bmatrix} = \begin{bmatrix} 51.02 \\ -48.47 \\ \vdots \\ 10.72 \end{bmatrix}$$

It can be easily calculated using R as: Res= y-fits

Variance-Covariance Matrix of Residuals. The residuals e_i, like the fitted values $\hat{Y}_i$, can be expressed as linear combinations of the response variable observations Y_i, using the result for $\hat{\mathbf{Y}}$:

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

We thus have the important result:

$$\underset{n\times 1}{\mathbf{e}} = \underset{n\times n}{(\mathbf{I} - \mathbf{H})} \underset{n\times 1}{\mathbf{Y}}$$

Now

$$Var(e) = Var((I-H)Y) = (I-H)\,\text{var}(Y)(I-H)$$
$$= (I-H)\sigma^2 I (I-H)'$$
$$= \sigma^2 (I-H) I (I-H)'$$
$$= \sigma^2 (I-H) = MSE(I-H),$$

where

$$MSE = SSE/(n-2) = e'e/(n-2).$$

$$SSE = e'e.$$

Analysis of Variance Results

Sums of Squares

To see how the sums of squares are expressed in matrix notation, we begin with the total sum of squares $SSTO$, It will be convenient to use an algebraically equivalent expression:

$$SSTO = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

We know from (5.13) that:

$$\mathbf{Y}'\mathbf{Y} = \sum Y_i^2$$

The subtraction term $(\sum Y_i)^2/n$ in matrix form uses $\mathbf{J}$, the matrix of 1s

$$\frac{(\sum Y_i)^2}{n} = \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

For instance, if $n = 2$, we have:

$$\left(\frac{1}{2}\right)[Y_1 \quad Y_2]\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \frac{(Y_1 + Y_2)(Y_1 + Y_2)}{2}$$

Hence, it follows that:

$$SSTO = \mathbf{Y'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$$

Just as $\sum Y_i^2$ is represented by $\mathbf{Y'Y}$ in matrix terms, so $SSE = \sum e_i^2 = \sum(Y_i - \hat{Y}_i)^2$ can be represented as follows:

$$SSE = \mathbf{e'e} = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y'Y} - 2\mathbf{b'X'Y} + \mathbf{b'X'Xb}$$

hence

$$SSE = Y'Y - b'X'Y .$$

Finally, it can be shown that:

$$SSR = \mathbf{b'X'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$$

Remark: SSTO=SSE+SSR

Example

For Toluca Company example, find the SSE, SSTO and SSR

$$\mathbf{Y'Y} = [399 \quad 121 \quad \cdots \quad 323]\begin{bmatrix} 399 \\ 121 \\ \vdots \\ 323 \end{bmatrix} = 2{,}745{,}173$$

$$\mathbf{b}'\mathbf{X}'\mathbf{Y} = [62.37 \quad 3.5702] \begin{bmatrix} 7,807 \\ 617,180 \end{bmatrix} = 2,690,348$$

Hence:

$$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} = 2,745,173 - 2,690,348 = 54,825$$

$$SSTO = Y'Y - \frac{1}{n}(Y'JY) = 2745173 - 2437970$$
$$= 307203$$

$$\text{SSR}=\text{b'X'Y}-\frac{1}{n}(Y'JY) = 2690348\text{-}243790\text{=}252378$$

Remark:

SSTO=t(y)%*%y-t(y)%*%J%*%y/25

SSR=t(b)%*%t(x)%*%y-t(y)%*%J%*%y/25

Sums of Squares as Quadratic Forms

The ANOVA sums of squares can be shown to be *quadratic forms*. An example of a quadratic form of the observations Y_i when $n = 2$ is:

$$5Y_1^2 + 6Y_1Y_2 + 4Y_2^2$$

Note that this expression is a second-degree polynomial containing terms involving the squares of the observations and the cross product. We can express in matrix terms as follows:

$$[Y_1 \quad Y_2]' \begin{bmatrix} 5 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \mathbf{Y}'\mathbf{A}\mathbf{Y}$$

where $\mathbf{A}$ is a symmetric matrix.

In general, a quadratic form is defined as:

$$\underset{1\times 1}{\mathbf{Y}'\mathbf{A}\mathbf{Y}} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}Y_iY_j \qquad \text{where } a_{ij} = a_{ji}$$

$\mathbf{A}$ is a symmetric $n \times n$ matrix and is called the *matrix of the quadratic form.*

The ANOVA sums of squares $SSTO$, SSE, and SSR are all quadratic forms, as can be seen by reexpressing $\mathbf{b}'\mathbf{X}'$. From (5.71), we know,

$$\mathbf{b}'\mathbf{X}' = (\mathbf{X}\mathbf{b})' = \hat{\mathbf{Y}}'$$

We now use the result in (5.73) to obtain:

$$\mathbf{b}'\mathbf{X}' = (\mathbf{H}\mathbf{Y})'$$

Since $\mathbf{H}$ is a symmetric matrix so that $\mathbf{H}' = \mathbf{H}$, we finally obtain

$$\mathbf{b}'\mathbf{X}' = \mathbf{Y}'\mathbf{H}$$

This result enables us to express the ANOVA sums of squares as follows:

$$SSTO = \mathbf{Y}'\left[\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y}$$

$$SSE = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$SSR = \mathbf{Y}'\left[\mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y}$$

Each of these sums of squares can now be seen to be of the form $\mathbf{Y}'\mathbf{A}\mathbf{Y}$, where the three $\mathbf{A}$ matrices are:

$$\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}$$

$$\mathbf{I} - \mathbf{H}$$

$$\mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J}$$