Multiple Regression II

regression model, with normal error terms, simply in terms of X variables:

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

where:

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
$X_{i1}, \ldots, X_{i,p-1}$ are known constants
ε_i are independent $N(0, \sigma^2)$
$i = 1, \ldots, n$

To express general linear regression model

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

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

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Note that the $\mathbf{Y}$ and $\boldsymbol{\varepsilon}$ vectors are the same as for simple linear regression. The $\boldsymbol{\beta}$ vector contains additional regression parameters, and the $\mathbf{X}$ matrix contains a column of 1s as well as a column of the n observations for each of the $p-1$ X variables in the regression model. The row subscript for each element X_{ik} in the $\mathbf{X}$ matrix identifies the trial or case, and the column subscript identifies the $\mathbf{X}$ variable.

In matrix terms, the general linear regression model

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{n\times p}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

where:

$\mathbf{Y}$ is a vector of responses

$\boldsymbol{\beta}$ is a vector of parameters

$\mathbf{X}$ is a matrix of constants

$\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with expectation

$\mathbf{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$ and variance-covariance matrix:

$$\underset{n\times n}{\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

Consequently, the random vector $\mathbf{Y}$ has expectation:

$$\underset{n\times 1}{\mathbf{E}\{\mathbf{Y}\}} = \mathbf{X}\boldsymbol{\beta}$$

and the variance-covariance matrix of $\mathbf{Y}$ is the same as that of $\boldsymbol{\varepsilon}$:

$$\underset{n\times n}{\boldsymbol{\sigma}^2\{\mathbf{Y}\}} = \sigma^2\mathbf{I}$$

Estimation of Regression Coefficients

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = \hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{Y} = Xb = X\hat{\beta} = HY, \quad H = X(X'X)^{-1}X'$$

$$E(\hat{\beta}) = \beta$$

$$Var(\hat{\beta}) = MSE(X'X)^{-1}$$

$$MSE = \frac{SSE}{n-p}$$

The Extra Sum of Squares

An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model. Equivalently, one can view an extra sum of squares as measuring the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model. We first utilize an example to illustrate these ideas, and then we present definitions of extra sums of squares and discuss a variety of uses of extra sums of squares in tests about regression coefficients.

Example (Book: page 256) Body fat example,

From the example, we define

$$SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2)$$

or, equivalently:

$$SSR(X_1|X_2) = SSR(X_1, X_2) - SSR(X_2)$$

If X_2 is the extra variable, we define:

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

or, equivalently:

$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$$

Extensions for three or more variables are straightforward. For example, we define:

$$SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$$

or:

$$SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$$

and:

$$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$$

or:

$$SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$$

and:

$$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$$

Decomposition of *SSR* into Extra Sums of Squares

In multiple regression, unlike simple linear regression, we can obtain a variety of decompositions of the regression sum of squares *SSR* into extra sums of squares.

$$SSTO = SSR(X_1) + SSE(X_1)$$

where the notation now shows explicitly that X_1 is the X variable in the model. Replacing $SSE(X_1)$ by its equivalent

$$SSTO = SSR(X_1) + SSR(X_2|X_1) + SSE(X_1, X_2)$$

We now make use of the same identity for multiple regression with two X variables as for a single X variable, namely:

$$SSTO = SSR(X_1, X_2) + SSE(X_1, X_2)$$

Solving (7.7) for $SSE(X_1, X_2)$

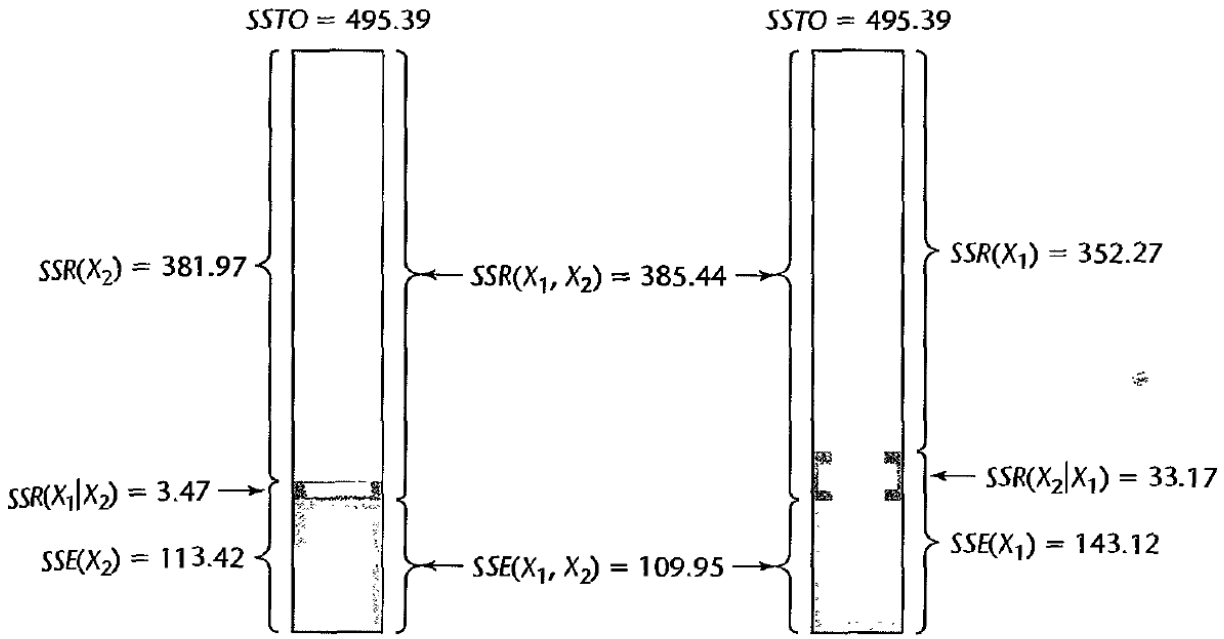
$$SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1)$$

Of course, the order of the X variables is arbitrary. Here, we can also obtain the decomposition:

$$SSR(X_1, X_2) = SSR(X_2) + SSR(X_1|X_2)$$

When the regression model contains three X variables, a variety of decompositions of $SSR(X_1, X_2, X_3)$ can be obtained. We illustrate three of these:

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2)$$
$$SSR(X_1, X_2, X_3) = SSR(X_2) + SSR(X_3|X_2) + SSR(X_1|X_2, X_3)$$
$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2, X_3|X_1)$$



Example of ANOVA Table with Decomposition of SSR for Three X Variables.

Source of Variation	SS	df	MS
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1	$MSR(X_2 \mid X_1)$
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1	$MSR(X_3 \mid X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	$n-4$	$MSE(X_1, X_2, X_3)$
Total	$SSTO$	$n-1$	

Uses of Extra Sums of Squares in Tests for Regression Coefficients

Test whether a Single $\beta_k = 0$

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

$$H_0: \beta_k = 0$$
$$H_a: \beta_k \neq 0$$

We already know that test statistic

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

is appropriate for this test.

We, now show that this can also be done using the extra sum of squares. Let us consider the first-order regression model with three predictor variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad \text{Full model}$$

To test the alternatives:

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

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

$$SSE(F) = SSE(X_1, X_2, X_3)$$

The degrees of freedom associated with $SSE(F)$ are $df_F = n - 4$ since there are four parameters in the regression function for the full model

The reduced model when H_0 holds is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad \text{Reduced model}$$

We next fit this reduced model and obtain:

$$SSE(R) = SSE(X_1, X_2)$$

There are $df_R = n - 3$ degrees of freedom associated with the reduced model.

The general linear test statistic

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

here becomes:

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

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

$$SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = SSR(X_3|X_1, X_2)$$

Hence the general linear test statistic here is:

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

And this can be compared with the critical region F(1, n-4) to have the decision.

Remark: F-statistic in this case also equal to (t-statistic)^2

Example:

In the body fat example, can we remove the X3 from the model?

Solution

1- H0: B3=0 vs H1: B3≠0

2-

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

$$= \frac{11.54}{1} \div \frac{98.41}{16} = 1.88$$

3-

For $\alpha = .01$, we require $F(.99; 1, 16) = 8.53$. Since $F^* = 1.88 \leq 8.53$, we conclude H_0, that X_3 can be dropped from the regression model that already contains X_1 and X_2.

Remark: if we use t-test we see that

Since $(t^*)^2 = (-1.37)^2 = 1.88 = F^*$, we see that the two test statistics are equivalent, just as for simple linear regression.

Test whether Several Coefficients

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

H_0: $\beta_2 = \beta_3 = 0$

H_a: not both β_2 and β_3 equal zero

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \qquad \text{Reduced model}$$

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

$$SSE(R) = SSE(X_1)$$

This error sum of squares has $df_R = n - 2$ degrees of freedom associated with it.

The general linear test statistic (2.70) thus becomes here:

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

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

$$SSE(X_1) - SSE(X_1, X_2, X_3) = SSR(X_2, X_3|X_1)$$

Hence, the test statistic becomes:

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

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

Example:

In the body fat example, can we remove the X2 and X3 from the model?