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:

Summary of Tests Concerning Regression Coefficients

Test whether All $\beta_k = 0$

This is the *overall F test* of whether or not there is a regression relation between the response variable Y and the set of X variables. The alternatives are:

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$
$$H_a\colon \text{not all } \beta_k \ (k = 1, \ldots, p-1) \text{ equal zero}$$

and the test statistic is:

$$\begin{aligned} F^* &= \frac{SSR(X_1, \ldots, X_{p-1})}{p-1} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p} \\ &= \frac{MSR}{MSE} \end{aligned}$$

If H_0 holds, $F^* \sim F(p-1, n-p)$. Large values of F^* lead to conclusion H_a.

Test whether a Single B$_k$ = 0

This is a *partial F test* of whether a particular regression coefficient β_k equals zero. The alternatives are:

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

and the test statistic is:

$$F^* = \frac{SSR(X_k|X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})}{1} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p}$$
$$= \frac{MSR(X_k|X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})}{MSE}$$

If H_0 holds, $F^* \sim F(1, n-p)$. Large values of F^* lead to conclusion H_a. Statistics packages that provide extra sums of squares permit use of this test without having to fit the reduced model.

An equivalent test statistic is

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

If H_0 holds, $t^* \sim t(n-p)$. Large values of $|t^*|$ lead to conclusion H_a.

Test whether Some B$_k$ = 0

This is another *partial F test*. Here, the alternatives are:

$$H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$
$$H_a: \text{not all of the } \beta_k \text{ in } H_0 \text{ equal zero}$$

where for convenience, we arrange the model so that the last $p-q$ coefficients are the ones to be tested. The test statistic is:

$$F^* = \frac{SSR(X_q, \ldots, X_{p-1}|X_1, \ldots, X_{q-1})}{p-q} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p}$$
$$= \frac{MSR(X_q, \ldots, X_{p-1}|X_1, \ldots, X_{q-1})}{MSE}$$

If H_0 holds, $F^* \sim F(p-q, n-p)$. Large values of F^* lead to conclusion H_a.

Remark:

The partial F* for several B$_k$=0 can be formed in terms of R^2 as

$$F^* = \frac{MSR(X_q,\ldots,X_{p-1}\,|\,X_1,\ldots,X_{q-1})}{MSE(X_1,\ldots,X_{p-1})} = \frac{\dfrac{SSR(X_q,\ldots,X_{p-1}\,|\,X_1,\ldots,X_{q-1})}{p-q}}{\dfrac{SSE(X_1,\ldots,X_{p-1})}{n-p}}$$

$$= \frac{\dfrac{SSR(X_1,\ldots,X_{p-1}) - SSR(X_1,\ldots,X_{q-1})}{p-q}}{\dfrac{SSE(X_1,\ldots,X_{p-1})}{n-p}}$$

$$= \frac{\dfrac{SSR(X_1,\ldots,X_{p-1}) - SSR(X_1,\ldots,X_{q-1})}{(p-q)SST}}{\dfrac{SSE(X_1,\ldots,X_{p-1})}{(n-p)SST}}$$

$$= \frac{\dfrac{R^2(X_1,\ldots,X_{p-1}) - R^2(X_1,\ldots,X_{q-1})}{(p-q)}}{\dfrac{1-R^2(X_1,\ldots,X_{p-1})}{(n-p)}} = \frac{\dfrac{R_F^2 - R_R^2}{df_R - df_F}}{\dfrac{1-R_F^2}{df_F}}$$

Coefficients of Partial determination

Two Predictor Variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

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

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

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

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

Thus, $R^2_{Y1|2}$ measures the proportionate reduction in the variation in Y remaining after X_2 is included in the model that is gained by also including X_1 in the model.

The coefficient of partial determination between Y and X_2, given that X_1 is in the model, is defined correspondingly:

Thus, $R^2_{Y1|2}$ measures the proportionate reduction in the variation in Y remaining after X_2 is included in the model that is gained by also including X_1 in the model.

The coefficient of partial determination between Y and X_2, given that X_1 is in the model, is defined correspondingly:

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

General Case

The generalization of coefficients of partial determination to three or more X variables in the model is immediate. For instance:

$$R^2_{Y1|23} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)}$$

$$R^2_{Y2|13} = \frac{SSR(X_2|X_1, X_3)}{SSE(X_1, X_3)}$$

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)}$$

$$R^2_{Y4|123} = \frac{SSR(X_4|X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)}$$

Note that in the subscripts to R^2, the entries to the left of the vertical bar show in turn the variable taken as the response and the X variable being added. The entries to the right of the vertical bar show the X variables already in the model.

Example

For the body fat example, we can obtain a variety of coefficients of partial determination. Here are three (Tables 7.2 and 7.4):

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{33.17}{143.12} = .232$$

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{11.54}{109.95} = .105$$

$$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3.47}{113.42} = .031$$

We see that when X_2 is added to the regression model containing X_1 here, the error sum of squares $SSE(X_1)$ is reduced by 23.2 percent. The error sum of squares for the model containing both X_1 and X_2 is only reduced by another 10.5 percent when X_3 is added to the model. Finally, if the regression model already contains X_2, adding X_1 reduces $SSE(X_2)$ by only 3.1 percent.

Coefficients of Partial Correlation

Example

For the body fat example, we have:

$$r_{Y2|1} = \sqrt{.232} = .482$$

$$r_{Y3|12} = -\sqrt{.105} = -.324$$

$$r_{Y1|2} = \sqrt{.031} = .176$$

Note that the coefficients $r_{Y2|1}$ and $r_{Y1|2}$ are positive because we see from $b_2 = .6594$ and $b_1 = .2224$ are positive. Similarly, $r_{Y3|12}$ is negative because we see from Table 7.2d that $b_3 = -2.186$ is negative.

Comment

Coefficients of partial determination can be expressed in terms of simple or other partial correlation coefficients. For example:

$$R^2_{Y2|1} = [r_{Y2|1}]^2 = \frac{(r_{Y2} - r_{12}r_{Y1})^2}{\left(1 - r^2_{12}\right)\left(1 - r^2_{Y1}\right)}$$

$$R^2_{Y2|13} = [r_{Y2|13}]^2 = \frac{(r_{Y2|3} - r_{12|3}r_{Y1|3})^2}{\left(1 - r^2_{12|3}\right)\left(1 - r^2_{Y1|3}\right)}$$

where r_{Y1} denotes the coefficient of simple correlation between Y and X_1, r_{12} denotes the coefficient of simple correlation between X_1 and X_2, and so on. Extensions are straightforward.

```
install.packages("asbio")

library(asbio)

lm1=lm(y~x1)

lm12=lm(y~x1+x2)

partial.R2(lm1, lm12)
```

Standardized Multiple Regression Model

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 $N(0, \sigma^2)$
$i = 1, \ldots, n$

The standardized regression model is as follows:

$$Y_i^* = \beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*$$

where the response variable Y^* and the independent variables X_i^* are given by

$$Y^{*} = \frac{Y - \bar{Y}}{S_Y} \quad , \quad X_i^{*} = \frac{X_i - \bar{X}_i}{S_{X_i}}, \quad i = 1, 2 ...., p - 1,$$

where $\bar{Y}$ and $\bar{X}_k$ are the respective means of the Y and the X_k observations, and s_Y and s_k are the respective standard deviations defined as follows:

$$s_Y = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}}$$

$$s_k = \sqrt{\frac{\sum_i (X_{ik} - \bar{X}_k)^2}{n-1}} \qquad (k = 1, \ldots, p-1)$$

The correlation transformation is a simple function of the standardized variables

$$Y_i^{*} = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

$$X_{ik}^{*} = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{s_k} \right) \qquad (k = 1, \ldots, p-1)$$

The relation between the coefficients of the original model and standardized model are

$$\beta_k = \left(\frac{s_Y}{s_k} \right) \beta_k^{*} \qquad (k = 1, \ldots, p - 1)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \cdots - \beta_{p-1} \bar{X}_{p-1}$$

We see that the standardized regression coefficients β_k^{*} and the original regression coefficients β_k $(k = 1, \ldots, p-1)$ are related by simple scaling factors involving ratios of standard deviations.

Estimated Standardized Regression Coefficients

Let

$$\underset{n\times(p-1)}{\mathbf{X}} = \begin{bmatrix} X_{11}^{*} & \cdots & X_{1,p-1}^{*} \\ X_{21}^{*} & \cdots & X_{2,p-1}^{*} \\ \vdots & & \vdots \\ X_{n1}^{*} & \cdots & X_{n,p-1}^{*} \end{bmatrix}$$

and

$$Y = \begin{bmatrix} y_{1}^{*} \\ y_{2}^{*} \\ \vdots \\ y_{n}^{*} \end{bmatrix}$$

Then

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

It can be shown that for the transformed variables, X'Y and X'X become

$$X\ 'X\ = r_{XX} \quad \mathit{and} \quad X\ 'Y\ = r_{yx}$$

$$r_{XX} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{12} & 1 & r_{23} & r_{2,p-1} \\ & & \ddots & \\ r_{1,p-1} & r_{2,p-1} & & r_{(1-p)(1-p)} \end{bmatrix}, \quad r_{YX} = \begin{bmatrix} r_{y1} \\ r_{y2} \\ \vdots \\ r_{y,p-1} \end{bmatrix},$$

$$r_{ij} = corr(X_i, X_j),\ r_{yi} = corr(Y, X_i)$$

and hence

$$\mathbf{b} = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$$

$$\underset{(p-1)\times 1}{\mathbf{b}} = \begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{p-1}^* \end{bmatrix},$$

The regression coefficients $b_1^*, \ldots, b_{p-1}^*$ are often called *standardized regression coefficients.*

The return to the estimated regression coefficients for regression model original variables is accomplished by employing the relations:

$$b_k = \left(\frac{s_Y}{s_k}\right) b_k^* \qquad (k = 1, \ldots, p-1)$$

$$b_0 = \bar{Y} - b_1\bar{X}_1 - \cdots - b_{p-1}\bar{X}_{p-1}$$

Example: In Dwaine Studios example data

$$Y_1^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_1 - \bar{Y}}{s_Y}\right) = \frac{1}{\sqrt{21-1}}\left(\frac{174.4 - 181.90}{36.191}\right) = -.04634$$

$$X_{11}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{11} - \bar{X}_1}{s_1}\right) = \frac{1}{\sqrt{21-1}}\left(\frac{68.5 - 62.019}{18.620}\right) = .07783$$

$$X_{12}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{12} - \bar{X}_2}{s_2}\right) = \frac{1}{\sqrt{21-1}}\left(\frac{16.7 - 17.143}{.97035}\right) = -.10208$$

$$\hat{Y}^* = .7484X_1^* + .2511X_2^*$$

and

$$b_1 = \left(\frac{s_Y}{s_1}\right) b_1^* = \frac{36.191}{18.620}(.7484) = 1.4546$$

$$b_2 = \left(\frac{s_Y}{s_2}\right) b_2^* = \frac{36.191}{.97035}(.2511) = 9.3652$$

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 181.90 - 1.4546(62.019) - 9.3652(17.143) = -68.860$$

R:Code

```
mydata=read.table("Dwaine Studios.txt",header=TRUE)
Y=mydata$Y
X1=mydata$X1
X2=mydata$X2
n=length(X1)
one=as.vector(rep(1, n))
X=cbind(one,X1,X2)
Model=lm(Y~X1+X2)
library(QuantPsyc)
lm.beta(Model)
```