Applied Regression

Frank Wood

March 4, 2010

Extra Sums of Squares

- A topic unique to multiple regression
- An extra sum of squares measures the marginal decrease in the error sum of squares when on or several predictor variables are added to the regression model, given that other 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

Example

- Multiple regression
  – Output: Body fat percentage – Input:
  1. triceps skin fold thickness(X_1)
  2. thigh circumference (X_2)
  3. midarm circumference (X_3)
- Aim
  –Replace cumbersome immersion procedure with model.
- Goal
  – Determine which predictor vafriables provide a good model.

The Data

Figure:

Subject i	Triceps Skinfold Thickness X_{i1}	Thigh Circumference X_{i2}	Midarm Circumference X_{i3}	Body Fat Y_i
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
...	...	...	...	...
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

Regression of Y on X_1

Figure:

(a) Regression of Y on X_1

$\hat{Y} = -1.496 + .8572X_1$

Source of Variation	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .8572$	$s\{b_1\} = .1288$	6.66

Regression of Y on X_2

Figure:

(b) Regression of Y on X_2

$$\hat{Y} = -23.634 + .8565X_2$$

Source of Variation	SS	df	MS
Regression	381.97	1	381.97
Error	113.42	18	6.30
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_2	$b_2 = .8565$	$s\{b_2\} = .1100$	7.79

Regression of Y on X_1 and X_2

Figure:

(c) Regression of Y on X_1 and X_2

$$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$$

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .2224$	$s\{b_1\} = .3034$	.73
X_2	$b_2 = .6594$	$s\{b_2\} = .2912$	2.26

(d) Regression of Y on X_1, X_2, and X_3

$$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$$

Regression of Y on X_1 and X_2 cont.

Figure:

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = 4.334$	$s\{b_1\} = 3.016$	1.44
X_2	$b_2 = -2.857$	$s\{b_2\} = 2.582$	−1.11
X_3	$b_3 = -2.186$	$s\{b_3\} = 1.596$	−1.37

Notation

- SSR X_1 only denoted by -SSR(X_1)=352.27
- SSE X_1 only denoted by -SSE(X_1)=143.12
- Accordingly,
  –SSR(X_1, X_2)=385.44
  –SSE(X_1, X_2)=109.95

More Powerful Model, Smaller SSE

- When X_1 and X_2 are in the model, SSE(X_1, X_2)=109.95 is smaller than when the model contains only X_1
- The difference is called an extra sum of squares and will be denoted by
  –$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2) = 33.17$
- The extra sum of squares $SSR(X_2|X_1)$ measure the marginal effect of adding X_2 to the regression model when X_1 is already in the model

SSR increase ¡-¿ SSE decrease

The extra sum of squares $SSR(X_1|X_2)$ can equivalently be viewed as the marginal increase in the regression sum of squares.
–$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$
–$= 385.44 - 352.27 = 33.17$

Why does this relationship exist?

- Remember $SSTO = SSR + SSE$
- SSTO measures only the variability of the Y's and does not depend on the regression model fitted.
- Any increase in SSR must be accompanied by a corresponding decrease in the SSE.

Example relations

$SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = 11.54$
or $SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2) = 11.54$
or with multiple variables included at time
–$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3) = 44.71$
–or $SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1) = 44.71$

Extra sums of squares

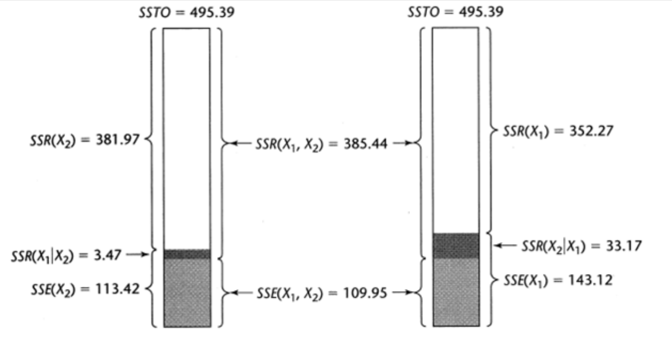
An extra sum of squares always involves the difference between the error sum of squares for the regression model containing the X variables in the model already the error sum of squares for the regression model containing both the original X variables and the new X variables.

Definitions

- Definition
  $-SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2)$
- Equivalently
  $-SSR(X_1|X_2) = SSR(X_1, X_2) - SSR(X_2)$
- We can switch the order of X_1 and X_2 in these expressions
- We can easily generalize these definitions for more than two variables
  $-SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$
  $-SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$

N! different partitions

Figure:



ANOVA Table

Various software packages can provide extra sums of squares for regression analysis. These are usually provided in the order in which the input variables are provided to the system, for instance

Figure:

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	$SSTO$	$n-1$	

Why? Who cares?

Extra sums of squares are of interest because they occur in a variety of tests about regression coefficients where the question of concern is whether certain X variables can be dropped from the regression model.

Test whether a single $\beta_k = 0$

- Does X_k provide statistically significant improvement to the regression model fit?
- We can use the general linear test approach
- Example
  –First order model with three predictor variables
  $Y_i = \beta_) + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$
  –We want to answer the following hypotheses test

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

Test whether a single $\beta_k = 0$

- For the full model we have $SSE(F) = SSE(X_1, X_2, X_3)$
- The reduced model is $Y_i = \beta_) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$
- And for this model we have $SSE(R) = SSE(X_1, X_2)$
- Where there are $df_r = n - 3$ degrees of freedom associated with the reduced model

Test whether a single $\beta_k = 0$

The general linear test statistics is

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

which becomes

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

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

Test whether a single $\beta_k = 0$

The general linear test statistics is

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

Extra sum of squares has one associated degree of freedom.

Example

Body fat: Can X_3 (midarm circumference) be dropped from the model?

Figure:

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
X_1	352.27	1	352.27
$X_2 \vert X_1$	33.17	1	33.17
$X_3 \vert X_1, X_2$	11.54	1	11.54
Error	98.41	16	6.15
Total	495.39	19	

$$F^* = \frac{SSR(X_3|X_1,X_2)}{1} \Big/ \frac{SSE(X_1,X_2,X_3)}{n-4} = 1.88$$

Example Cont.

- For $\alpha = .01$ we require $F(.99; 1, 16) = 8.53$
- We observe $F^* = 1.88$
- We conclude $H_0 : \beta_3 = 0$

Test whether $\beta_k = 0$

Another example
$H_0 : \beta_2 = \beta_3 = 0$
H_1 : not both β_2 and β_3 are zero
The general linear test can be used again

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

But $SSE(X_1) - SSE(X_1, X_2, X_3) = SSR(X_2, X_3 | X_1)$
so the expression can be simplified.

Tests concerning regression coefficients

Summary:

– General linear test can be used to determine whether or not a predictor variable(or sets of variables) should be included in the model

– The ANOVA SSE's can be used to compute F^* test statistics

– Some more general tests require fitting the model more than once unlike the examples given.

Standardized Multiple Regression

- Numerical precision errors can occur when
  - $(X'X)^{-1}$ is poorly conditioned near singular : colinearity
  - And when the predictor variables have substantially different magnitudes
- Solution
  – Regularization
  – Standardized multiple regression
- First, transformed variables

Correlation Transformation

Makes all entries in $X'X$ matrix for the transformed variables fall between -1 and 1 inclusive

Another motivation

– Lack of comparability of regression coefficients

$\hat{Y} = 200 + 20000X_1 + .2X_2$

Y in dollars, X_1 in thousand dollars, X_2 in cents

– Which is most important predictor?

Correlation Transformation

1. Centering

$$\frac{Y_i - \bar{Y}}{s_y}$$

$$\frac{X_{ik} - \bar{X}}{s_k}, k - 1, ..., p - 1$$

2. Scaling

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

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

Correlation Transformation

Transformed variables

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

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

Standardized Regression Model

Define the matrix consisting of the transformed X variables

$$X = \begin{pmatrix} X_{11} & \dots & X_{1,p-1} \\ X_{21} & \dots & X_{2,p-1} \\ \dots & & \\ X_{n1} & \dots & X_{n,p-1} \end{pmatrix}$$

And define $X'X = r_{xx}$

Correlation matrix of the X variables

Can show that

$$r_{xx} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1,p-1} \\ r_{21} & 1 & \dots & r_{2,p-1} \\ \dots & & & \\ r_{p-1,1} & r_{p-1,2} & \dots & 1 \end{pmatrix}$$

where each entry is just the coefficient of correlation between X_i and X_j

$$\begin{aligned} \sum x^*_{i1} x^*_{i2} &= \sum \left(\frac{X_{i1} - \bar{X}_1}{\sqrt{n-1} s_1}\right)\left(\frac{X_{i2} - \bar{X}_2}{\sqrt{n-1} s_2}\right) \\ &= \frac{1}{n-1} \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{s_1 s_2} \\ &= \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{[\sum (X_{i1} - \bar{X}_1)^2 \sum (X_{i2} - \bar{X}_2)^2]^{1/2}} \end{aligned}$$

Standardized Regression Model

- If we define in a similar way $X'Y = r_{yx}$, where r_{yx} is the coefficient of simple correlations between the response variable Y and X_j
- Then we can set up a standard linear regression problem

$$r_{xx}b = r_{yx}$$

Standardized Regression Model

The solution

$$\mathbf{b} = \begin{pmatrix} b_1^* \\ b_2^* \\ . \\ . \\ . \\ b_{p-1}^* \end{pmatrix}$$

can be related to the solution to the untransformed regression problem through the relationship

$$b_k = \left(\frac{s_y}{s_k}\right) b_k^*, k = 1, ..., p-1$$
$$b_0 = \bar{Y} - b_1 \bar{X}_1 - ... - b_{p-1} \bar{X}_{p-1}$$

Multi-colinearity

- Brief summary
- $(X'X)^{-1}$ must be full rank to compute the regression solution –$rank(AB) <= min(rank(A), rank(B))$
- Multi-colinearity means that rows of X are linearly dependent
- Regression solution is degenerate
- High degrees of colinearity produce numerical instability
- Very important to consider in real world applications