Multiple Regression - Extra Sums of Squares

- Big Pic: Model selection, First for inclusion, "opinion" enters

Basic Idea: 2 views

1) An extra sum of squares measures the marginal reduction in the error sum of squares when var's are added to the model
2) Equivalently, an extra sum of squares measures the marginal increase in the regression sum of squares

Example: Female body fat amount vs. several predictor var's
X_1 thigh circumference, X_3 midarm circ., X_2 thigh circ.

Consider ~~3~~ 4 different regression models ← model selection

1) Y regressed on X_1 alone ← note intercept, does this make sense?

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

		df	MS
SSR	352.27	1	352.27
SSE	143.12	18	7.95
SSTO	495.39	19	$\boxed{\Rightarrow n=20}$

Var	Est reg. coeff	Est. Std.	t^*
X_1	$b_1 = .8572$	$s\{b_1\} = .1288$	6.66

2) Y regressed on X_2 $\qquad \hat{Y} = -23.634 + .8565 X_2$

SSR	381.97	1	381.97
SSE	113.42	18	6.30
SSTO	495.39	19	

Var			
X_2	$b_2 = .8565$	$s\{b_2\} = .1100$	7.79

3) Regress. of Y on X_1 & X_2

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

		df	MS
SSR	385.44	2	192.72
SSE	109.95	17	6.47
SSTO	495.39	19	

Var		Est std. dev	t^*
X_1	$b_1 = .2224$	$s\{b_1\} = .3034$	.73
X_2	$b_2 = .6594$	$s\{b_2\} = .2912$	2.26

4) Regress of Y on X_1, X_2 & X_3

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

SSR	396.98	3
SSE	98.41	16
SSTO	495.39	19

Var			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

Remember $t^* = b_i / s\{b_i\}$

Notice: when X_1 & X_2 are in model, $SSE(X_1, X_2) = 109.95$ which is less than $SSE(X_1) = 143.12$ and $SSE(X_2) = 113.42$. This difference is an

extra sum of squares

$$SSR(X_2 \mid X_1) \equiv SSE(X_1) - SSE(X_1, X_2)$$

decrease in error

$$= 143.12 - 109.95 = 33.17$$

Equivalently

$$SSR(X_2 \mid X_1) \equiv SSR(X_1, X_2) - SSR(X_1)$$

$$= 385.44 - 352.27 = 33.17$$

increase in regression sum of squares

$$X'Xb = X'Y \qquad H = X(X'X)^{-1}X'$$
$$b = (X'X)^{-1}X'$$

Remember

$$\begin{aligned} SSTO &= Y'\left(I - \tfrac{1}{n}J\right)Y \\ SSE &= Y'(I - H)Y \\ SSR &= Y'\left(H - \tfrac{1}{n}J\right)Y \end{aligned}$$

~~$SSTO = SSR(X_1, X_2, X_3) +$~~

~~$SSE(X_1)$~~

$$SSTO = \cancel{SSR(X_1, X_2, X_3) + SSE(X_1, X_2, X_3)}$$
$$= \cancel{SSR(}$$
$$= SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2)$$

$$SSTO = Y'\left(H_1 - \tfrac{1}{n}J\right)Y + Y'(H_{2|1})Y + Y'(H_{3|1,2})Y + Y'(H_d \mid_{1:d-1})Y + Y'\left(H - \tfrac{1}{n}J\right)Y$$

slight problem were leaving out the intercept

If $\left(H_1 + H_{2|1} + H_{3|1,2} + H_{d:1:d-1}\right) = H$ and ranks sum then we're golden

$$H_1 = X_1(X_1'X_1)^{-1}X_1' + X_2(X_2'X_2)^{-1}X_2'$$

if $X_1 \perp X_2$ then $H = H_1 + H_{2|1}$

sketch

$$\begin{array}{c} \\ n \end{array}\left[\begin{array}{c|c} X_1 & X_2 \end{array}\right] \; p\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \left[\begin{array}{c} X_1 \\ \hline X_2 \end{array}\right]$$

if diag's are zero then

$$H = \frac{1}{(X_1'X_1)}X_1X_1' + \frac{1}{(X_2'X_2)}X_2X_2' = H_1 + H_{2|1}$$

Re decomposition, obviously

$$SSTO = \uparrow SSR + SSE \downarrow$$

any reduction in ~~SSR~~ SSE must be accompanied by an increase in SSR, (SSTO is fixed)

It can also be seen that

$$SSR(X_3 | X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$$
$$= 109.95 - 98.41 = 11.54$$

and equiv

$$SSR(X_3 | X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$$
$$= 396.98 - 385.44 = 11.54$$

We can also consider adding multiple variables at once

ie.

$$SSR(X_2, X_3 | X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$$
$$= 143.12 - 98.41 = 44.71$$

and equiv

$$SSR(X_2, X_3 | X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$$
$$= 396.98 - 352.27 = 44.71$$

Def's

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

equiv.

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

note, opposite order of args

Conversely

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

and

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

3-var & more extensions straightforward

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

and

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

Decomposition of SSR into Extra Sums of Squares

Think - want error with all vars included, but would like to know impact of adding vars to regression model.

To start, consider

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

remember

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

substituting

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

and so on and so forth, ie.

$$SSR(X_1, X_2 \ldots X_d) = SSR(X_1) + SSR(X_2|X_1) + \ldots + SSR(X_d | X_1, X_2 \ldots, X_{d-1})$$

note - the order of this decomposition is arbitrary (how many? d!)

ANOVA

Anova tables can be constructed for these decompositions. ie.

Source of Var.	SS	df	MS
Reg.	$SSR(X_1 X_2 X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \| X_1$	$SSR(X_2 \| X_1)$	1	$MSR(X_2 \| X_1)$
$X_3 \| X_1, X_2$	$SSR(X_3 \| X_1, X_2)$	1	$MSR(X_3 \| X_1, X_2)$
Error	$SSE(X_1, X_2, X_3$	$n-4$	$MSE(X_1 X_2, X_3)$
Total	$SSTO$	$n-1$	

The order of these vars is arbitrary.

What use? Tests.

Test whether a Single $\beta_k = 0$

To test whether $\beta_k X_k$ can be dropped from a multiple regression model we are interested in

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

The test stat for this is

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

this is one way.

- F-test way

Consider full model

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

and testing the alternatives

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

Recipe: fit full model and compute SSE, etc

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

df in full model SSE is $n-4$

Reduced model when H_0 holds is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad \text{(reduced model)}$$

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

df in reduced model is $n-3$

The general linear test stat is

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

here is

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

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

i.e.

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

$$= \frac{MSR(X_3 | X_1, X_2)}{MSE(X_1, X_2, X_3)}$$

So ANOVA table with extra-sums of squares can be used to do model selection efficiently.

Similar technique(s) can be used to test whether several $\beta_k = 0$

Review tests in 7.3

Multi-collinearity

Comments

1) Correlated predictor variables do not inhibit getting a good model fit nor prediction.

2) Correlated predictor var's lead to large sampling intervals for the estimated regress. coeff's. Individual predictors might be deemed statistically insignificant even though there $\exists$ a relationship

3) Interpretation gets difficult: if predictors are multicollinear then the interpretation of linear rate of change of output given fixed other covariates no longer valid.