Multiple Regression - Extra Sums of Squares

Big Pic: Model selection. First for inclusion, "opinion" effect.

Basic Idea: 2 views

1) An extra sum of squares measures the marginal reduction in the error sum of squares when vars 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 content vs. several predictor vars $X_1$: thigh circumference, $X_2$: midarm circumference, $X_2$: thigh circumference.

Consider 8 different regression models $\to$ model selection.

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

\[
\hat{y} = -1.496 + 0.8572 x_1
\]

<table>
<thead>
<tr>
<th>SS</th>
<th>352.27</th>
<th>1</th>
<th>352.27</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE</td>
<td>143.12</td>
<td>18</td>
<td>7.95</td>
</tr>
<tr>
<td>SSTO</td>
<td>495.39</td>
<td>19</td>
<td>$\Rightarrow n = 20$</td>
</tr>
</tbody>
</table>

$\hat{\beta}_1 = .8572$

Vor $\hat{\beta}_1 = .54$, $\hat{\sigma} = .66$

2) $Y$ regressed on $X_2$

\[
\hat{y} = -23.634 + .8565 x_2
\]

<table>
<thead>
<tr>
<th>SS</th>
<th>381.97</th>
<th>1</th>
<th>381.97</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE</td>
<td>111.42</td>
<td>18</td>
<td>6.30</td>
</tr>
<tr>
<td>SSTO</td>
<td>495.39</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

Vor $\hat{\beta}_2 = .8565$

$\hat{\sigma} = .1100$, $\hat{\sigma}_y = 7.79$
3) Regress of $Y \sim X_1 \sim X_2$

\[ \hat{Y} = -19.174 + 0.2224X_1 + 0.6594X_2 \]

\[ \begin{array}{c|ccc}
SSR & 385.44 & 2 & 192.72 \\
SSE & 109.95 & 17 & 6.47 \\
SSTO & 495.39 & 19 & \\
\end{array} \]

- Var

\[ \bar{b}_1 = 0.2224 \hspace{2cm} \bar{b}_2 = 0.6594 \]

\[ \bar{b}_1 \text{ St. Dev} = 0.73 \]

\[ \bar{b}_2 \text{ St. Dev} = 0.26 \]

4) Regress of $Y \sim X_1 \sim X_2 \sim X_3$

\[ \hat{Y} = 117.09 + 4.334X_1 - 2.857X_2 - 2.186X_3 \]

\[ \begin{array}{c|ccc}
SSR & 396.98 & 3 & \\
SSE & 47.41 & 14 & \\
SSTO & 495.39 & 19 & \\
\end{array} \]

- Var

\[ \bar{b}_1 = 4.334 \hspace{2cm} \bar{b}_2 = -2.857 \hspace{2cm} \bar{b}_3 = -2.186 \]

\[ \bar{b}_1 \text{ St. Dev} = 1.44 \hspace{2cm} \bar{b}_2 \text{ St. Dev} = -1.11 \hspace{2cm} \bar{b}_3 \text{ St. Dev} = -1.34 \]

\[ R\text{-}\text{value} \quad t^* = \frac{b}{s_b} \]

Notice: when $X_1 \sim X_2$ are in model, $SSE(X_1, X_2) = 109.95$ which is less than $SSE(X_i) = 143.12$ and $SSE(X_2) = 113.42$. This difference is an extra sum of squares

\[ \text{decrease in error sum} \quad = 143.12 - 109.95 = 33.17 \]

\[ \text{Equivalently} \]

\[ \text{SSR}(X_2 | X_1) = \text{SSR}(X_1, X_2) - \text{SSR}(X_1) \]

\[ = 385.44 - 752.27 = 36.17 \]

\[ \text{Increase in regression sum of squares} \]
\[ x' x b = x' x, \quad b = (x' x)^{-1} x' \]

\[
SSTO = y' \left( I - \frac{1}{n} J \right) y
\]
\[
SSE = y' (I - H) y
\]
\[
SSR = y' \left( I - \frac{1}{n} J \right) y
\]

\[
SSR(x_1, x_2, x_3)
\]

\[
SSE(x_i)
\]

\[
SSTO = SSR(x_1, x_2, x_3) - SSE(x_1, x_1, x_3) - SSR(x_3 | x_1, x_2)
\]

\[
SSR = y' \left( I - \frac{1}{n} J \right) y + y' \left( H_{21} - \frac{1}{n} J \right) y + y' \left( H_{31} - \frac{1}{n} J \right) y
\]

If \( (H_1 + H_{21} + H_{31}) = H \)

Then we're golden

\[
H_i = x_1 (x' x)^{-1} x - x_2 \left( x_2' x_2 \right)^{-1} x_2
\]

If \( x_1 \parallel x_2 \) then \( H = H_1 + H_{21} \)

\[
H = \frac{1}{2} (x_1' x_1) x_1 + \frac{1}{2} (x_2' x_2) x_2 = H_1 + H_{21}
\]
Re: decomposition, obviously

\[ SSSTO \rightarrow SSR + SSE \downarrow \]

Any reduction in SSE must be accompanied by an increase in SSR (SSSTO is fixed).

It can also be seen that

\[ SSR(X_3 | X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \]

and

\[ SSR(X_3 | X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \]

= 109.95 - 99.41 = 11.54

We can also consider adding multiple variables at once

\[ SSR(X_2, X_3 | X_1) = SSE(X_1) - SSE(X_1, X_2, X_3) \]

= 143.12 - 99.41 = 44.71

and

\[ SSR(X_2, X_3 | X_1) = SSR(X_1, X_2, X_3) - SSR(X_1) \]

= 156.98 - 352.27 = 44.71

**Def's**

\[ SSR(X_1 | X_2) = SSE(X_2) - SSE(X_1, X_2) \]

\[ SSR(X_1 | X_2) = SSR(X_1, X_2) - SSR(X_2) \]

Conversely

\[ SSR(X_2 | X_1) = SSE(X_1) - SSE(X_1, X_2) \]

\[ SSR(X_2 | X_1) = SSR(X_1, X_2) - SSR(X_1) \]

3-way or more extensions straightforward

\[ SSR(X_3 | X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \]

\[ SSR(X_3 | X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \]
Decomposition of SSR in Extra Sum of Squares

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

To start, consider

$$SS_{TD} = SSR(x_1) + SSE(x_1)$$

Remember

$$SSE(x_1) = SSE(x_1, x_2) + SSR(x_2 | x_1)$$

Substituting

$$SS_{TD} = SSR(x_1) + SSR(x_2 | x_1) + SSE(x_1, x_2)$$

and so on and so forth, etc.

$$SSR(x_1, x_2, \ldots, x_d) = SSR(x_1) + SSR(x_2 | x_1) + \ldots + SSR(x_d | x_1, x_2, \ldots, x_{d-1})$$

Note: the order of this decomposition is arbitrary (why?)

ANOVA

Anova tables can be constructed from these decompositions, i.e.

<table>
<thead>
<tr>
<th>Source of Var.</th>
<th>$SS$</th>
<th>$df$</th>
<th>$MS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reg.</td>
<td>SSR(x_1, x_2, x_3)</td>
<td>3</td>
<td>MSR(x_1, x_2, x_3)</td>
</tr>
<tr>
<td>$x_1$</td>
<td>SSR(x_1)</td>
<td>1</td>
<td>MSR(x_1)</td>
</tr>
<tr>
<td>$x_2</td>
<td>x_1$</td>
<td>SSR(x_2</td>
<td>x_1)</td>
</tr>
<tr>
<td>$x_3</td>
<td>x_1, x_2$</td>
<td>SSR(x_3</td>
<td>x_1, x_2)</td>
</tr>
<tr>
<td>Error</td>
<td>SSE(x_1, x_2, x_3)</td>
<td>$n-4$</td>
<td>MSE(x_1, x_2, x_3)</td>
</tr>
<tr>
<td>Total</td>
<td>SSTD</td>
<td>$n-1$</td>
<td></td>
</tr>
</tbody>
</table>

The order of these vars is arbitrary.

What we? 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

\[ H_0 : \beta_k = 0 \]
\[ H_1 : \beta_k \neq 0 \]

The test statistic for this is

\[ t^* = \frac{b_k}{s(b_k)} \]

this is one way.

- F-test way

Consider the full model

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \]

and testing the alternative

\[ H_0 : \beta_3 = 0 \]
\[ H_1 : \beta_3 \neq 0 \]

Recipe: fit full model and compute $\text{SSE}_{\text{full}}$, etc.

\[ \text{SSE}(F) = \text{SSE}(X_1, X_2, X_3) \]

\[ \text{df in full model} \quad \text{SSE is n - 4} \]

Reduced model where $H_0$ holds is

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad \text{(reduced model)} \]

\[ \text{SSE}(R) = \text{SSE}(X_1, X_2) \]

\[ \text{df in reduced model is n - 3} \]

The general F-test statistic is

\[ F^* = \frac{\text{SSE}(R) - \text{SSE}(F)}{\text{df}_R} \bigg/ \frac{\text{SSE}(F)}{\text{df}_F} \]

Here is

\[ \frac{\text{SSE}(X_1, X_2) - \text{SSE}(X_1, X_2, X_3)}{(n-3) - (n-4)} \bigg/ \frac{\text{SSE}(X_1, X_2, X_3)}{n-4} \]
But \( \text{SSE}(x_1, x_2) - \text{SSE}(x_1, x_2, x_3) = \text{SSR}(x_3 | x_1, x_2) \)

i.e.

\[
F^* = \frac{\text{SSR}(x_3 | x_1, x_2)}{\text{MSE}(x_1, x_2, x_3)}
\]

\[
= \frac{\text{MSR}(x_3 | x_1, x_2)}{\text{MSE}(x_1, x_2, x_3)}
\]

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 \( p_k = 0 \).

Review tests in 7.3.

---

**Multi-collinearity**

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

2) Correlated predictor vars lead to large sampling intervals for the estimated regression coeff.

Individual predictors might be deemed statistically insignificant even though there is 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.