Lecture 13
Extra Sums of Squares

STAT 512
Spring 2011

Background Reading
KNNL: 7.1-7.4
Topic Overview

- Extra Sums of Squares (Defined)
- Using and Interpreting $R^2$ and Partial-$R^2$
- Getting ESS and Partial-$R^2$ from SAS
- General Linear Test (Review Section 2.8)
- Testing single $\beta_k = 0$
- Testing several $\beta_k = 0$
- Other General Linear Tests
Extra Sums of Squares

- ESS measure the *marginal* reduction in the error sum of squares from the addition of a group of predictor variables to the model.

- Examples
  - $SSR(X_1, X_2, X_3)$ is the total variation explained by $X_1$, $X_2$, and $X_3$ in a model
  - $SSR(X_1 \mid X_2)$ is the additional variation explained by $X_1$ when added to a model already containing $X_2$
  - $SSR(X_1, X_4 \mid X_2, X_3)$ is the additional variation explained by $X_1$ and $X_4$ when added to a model already containing $X_2$ and $X_3$
Extra Sums of Squares (2)

• Can also view in terms of SSE’s
• ESS represents the part of the SSE that is explained by an added group of variables that was not previously explained by the rest.
• Examples
  \[ SSR(X_1 | X_2) = SSE(X_2) - SSE(X_1, X_2) \]
  \[ SSR(X_1, X_4 | X_2, X_3) = SSE(X_2, X_3) - SSE(X_1, X_2, X_3, X_4) \]
Extra Sums of Squares (3)

$SS_{TOT}$

$SSR(X_2,X_3)$

$SSR(X_1,X_4|X_2,X_3)$
Decomposition of SSR (TYPE I)

• Regression SS can be partitioned into pieces (in any order):

\[ SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2) + SSR(X_4 \mid X_1, X_2, X_3) \]

• This particular breakdown is called TYPE I sums of squares (variables added in order).
Extended ANOVA Table

- Row for “Model” or “Regression” becomes $p - 1$ rows, in terms of Type I SS and MS.

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>Sum of Sq</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td>SSR(X1)</td>
<td>MSR(X1)</td>
</tr>
<tr>
<td>X2</td>
<td>1</td>
<td>SSR(X2</td>
<td>X1)</td>
</tr>
<tr>
<td>X3</td>
<td>1</td>
<td>SSR(X3</td>
<td>X1,X2)</td>
</tr>
<tr>
<td>ERROR</td>
<td>n-4</td>
<td>SSE(X1,X2,X3)</td>
<td>MSE(X1,X2,X3)</td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SST</td>
<td></td>
</tr>
</tbody>
</table>

- Decomposition can be obtained in SAS
Type III SS

- Type III sums of squares refers to variables added last. These do NOT add to the SSR.

\[ SSR(X_1 \mid X_2, X_3, X_4) \]
\[ SSR(X_2 \mid X_1, X_3, X_4) \]
\[ SSR(X_3 \mid X_1, X_2, X_4) \]
\[ SSR(X_4 \mid X_1, X_2, X_3) \]

- Also can be obtained from SAS; Type III SS leads to variable-added-last tests.
Getting ESS from SAS

• New Procedure: GLM (stands for general linear model)
• GLM is quite similar to REG, but can handle ANOVA when we get there
• Computer Science Example

```proc glm data=cs;
  model gpa = hsm hss hse satm satv /
    clparm alpha=0.05;
```

• Note: Output gives way more decimals than needed. OK to cut to reasonable.
GLM Output

Source  DF   SS     MS  F Value  Pr > F
Model    5   28.64  5.73   11.69   <.0001
Error   218  106.82  0.49
Total   223  135.46

R-Square  Coeff Var  Root MSE  gpa Mean
0.2114     26.56      0.7000    2.635

- Standard output that we are used to. F-test is for the overall model – answers question of whether any important variables are involved.
### GLM Output (2)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>MS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsm</td>
<td>1</td>
<td>25.810</td>
<td>25.810</td>
<td>52.67</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>hss</td>
<td>1</td>
<td>1.237</td>
<td>1.237</td>
<td>2.52</td>
<td>0.1135</td>
</tr>
<tr>
<td>hse</td>
<td>1</td>
<td>0.665</td>
<td>0.665</td>
<td>1.36</td>
<td>0.2452</td>
</tr>
<tr>
<td>satm</td>
<td>1</td>
<td>0.699</td>
<td>0.699</td>
<td>1.43</td>
<td>0.2337</td>
</tr>
<tr>
<td>satv</td>
<td>1</td>
<td>0.233</td>
<td>0.233</td>
<td>0.47</td>
<td>0.4915</td>
</tr>
</tbody>
</table>

- Type I – Variables Added In Order; SS add to SSR on previous slide.
- F-tests are testing each variable *given* previous variables already in model.
GLM Output (3)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>MS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsm</td>
<td>1</td>
<td>6.772</td>
<td>6.772</td>
<td>13.82</td>
<td>0.0003</td>
</tr>
<tr>
<td>hss</td>
<td>1</td>
<td>0.442</td>
<td>0.442</td>
<td>0.90</td>
<td>0.3432</td>
</tr>
<tr>
<td>hse</td>
<td>1</td>
<td>0.957</td>
<td>0.957</td>
<td>1.95</td>
<td>0.1637</td>
</tr>
<tr>
<td>satm</td>
<td>1</td>
<td>0.928</td>
<td>0.928</td>
<td>1.89</td>
<td>0.1702</td>
</tr>
<tr>
<td>satv</td>
<td>1</td>
<td>0.233</td>
<td>0.233</td>
<td>0.47</td>
<td>0.4915</td>
</tr>
</tbody>
</table>

- Type III – Variables Added Last
- F-tests are testing variables *given* that all of the other variables already in model
Coefficients of Partial Determination

• Recall: $R^2$ is the coefficient of determination, and may be interpreted as the percentage of the total variation that has been explained by the model.

• Example: $R^2 = 0.87$ means 87% of the Total SS has been explained by the regression model (of however many variables)

• Can also consider the amount of remaining variation explained by a variable given other variables already in the model – this is called partial determination.
Coef. of Partial Determination (2)

- Notation: $R_{Y_{1|23}}^2$ represents the percentage of the leftover variation in $Y$ (after regressing on $X_2$ and $X_3$) that is explained by $X_1$.

- Mathematically,

$$R_{Y_{1|23}}^2 = \frac{\text{SSE}(X_2, X_3) - \text{SSE}(X_1, X_2, X_3)}{\text{SSE}(X_2, X_3)}$$

$$= \frac{\text{SSR}(X_1 | X_2, X_3)}{\text{SSE}(X_2, X_3)}$$

- Subscripts after bar (|) represent variables already in model.
**Example**

- Suppose total sums of squares is 100, and $X_1$ explains 60.
- Of the remaining 40, $X_2$ then explains 20, and of the remaining 20, $X_3$ explains 5.
- Then

\[
R^2_{Y|2|1} = \frac{20}{40} = 0.50
\]

\[
R^2_{Y|3|2} = \frac{5}{20} = 0.25
\]
Coefficient of Partial Correlation

- Square Root of the coefficient of partial determination
- Given plus/minus sign according to the corresponding regression coefficient
- Can be useful in model selection (Chapter 9); but no clear interpretation like $R^2$.
- Notation: $r_{y1|23}$
Getting Partial $R^2$ from SAS

- PROC REG can produce these, along with sums of squares (in REG, the TYPE III SS are actually denoted as TYPE II – there is no difference between the two types for normal regression, but there is for ANOVA so we’ll discuss this later)

- CS Example

```sas
proc reg data=cs;
   model gpa = hsm hss hse satm satv 
      /ss1 ss2 pcorr1 pcorr2;
```
### REG Output

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>SS(I)</th>
<th>SS(II)</th>
<th>Corr(I)</th>
<th>Corr(II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1555</td>
<td>0.327</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>hsm</td>
<td>1</td>
<td>25.8</td>
<td>6.772</td>
<td>0.19053</td>
<td>0.05962</td>
</tr>
<tr>
<td>hss</td>
<td>1</td>
<td>1.2</td>
<td>0.442</td>
<td>0.01128</td>
<td>0.00412</td>
</tr>
<tr>
<td>hse</td>
<td>1</td>
<td>0.7</td>
<td>0.957</td>
<td>0.00614</td>
<td>0.00888</td>
</tr>
<tr>
<td>satm</td>
<td>1</td>
<td>0.7</td>
<td>0.928</td>
<td>0.00648</td>
<td>0.00861</td>
</tr>
<tr>
<td>satv</td>
<td>1</td>
<td>0.2</td>
<td>0.233</td>
<td>0.00217</td>
<td>0.00217</td>
</tr>
</tbody>
</table>

- Example: HSE explains 0.6% of remaining variation after HSM and HSS in model
REG Output (2)

• Can get any partial coefficient of determination that we want, but may have to rearrange model to do it.

• Example: If we want HSE given HSM, we would need to list variables HSM and HSE as the first and second in the model.

• Can get any desired Type I SS in the same way.
### REG Output (3)

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>SS(I)</th>
<th>SS(II)</th>
<th>Corr(I)</th>
<th>Corr(II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1555</td>
<td>0.327</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>hsm</td>
<td>1</td>
<td>25.8</td>
<td>6.772</td>
<td>0.19053</td>
<td>0.05962</td>
</tr>
<tr>
<td>hse</td>
<td>1</td>
<td>1.5</td>
<td>0.957</td>
<td>0.01362</td>
<td>0.00412</td>
</tr>
</tbody>
</table>

- Interpretation: Once HSM is in the model, of the remaining variation (SSE=109) HSE explains only 1.36% of it.
General Linear Test

• Compare two models:
  - Full Model: All variables / parameters
  - Reduced Model: Apply NULL hypothesis to full model.

• Example: 4 variables, $H_0 : \beta_2 = \beta_3 = 0$
  FULL: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$
  REDUCED: $Y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$

• F-statistic is

$$F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$$
General Linear Test (2)

- Numerator of F-test is the difference in SSE’s, or the EXTRA SS associated to the “added” variables; divided by number of variables being “added” (d.f.)

- Denominator is MSE for full model.

- For the example, test statistic will be

\[
F = \frac{SSR(X_2, X_3 | X_1, X_4)/2}{MSE(X_1, X_2, X_3, X_4)}
\]

- Compare to F-distribution on 2 and \( n - 5 \) degrees of freedom.
Alternative Hypotheses

• Alternative is simply that the null is false.
• Most of the time, the alternative will be that at least one of the variables in the null group is important.
• Often looking to “fail to reject” when performing a test like this – our goal is to eliminate unnecessary variables.
• This means POWER / sample size must be a consideration! If our sample size is too small, we may incorrectly remove variables.
CS Example (ess.sas)

- Test whether HSS, SATM, SATV as a group are important when added to model containing HSM and HSE.
- SSE for HSM/HSE model is 108.16 on 221 degrees of freedom
- SSE for full model is 106.82 on 218 degrees of freedom; MSE is 0.49
- F statistic is

\[ F = \frac{(108.16 - 106.82)/3}{0.49} = 0.91 \]
CS Example (2)

- F < 1 so no need to even look up a value; fail to reject.
- With 224 data points, we likely have the power required to conclude that the three variables are not useful in the model that already contains HSM and HSE.
- Can obtain this test in SAS using a test statement:

```sas
proc reg data=cs;
    model gpa = hsm hss hse satm satv;
    TEST1: test hss=0, satm=0, satv=0;
```
TEST output

Test TEST1 Results

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>3</td>
<td>0.44672</td>
<td>0.91</td>
<td>0.4361</td>
</tr>
<tr>
<td>Denominator</td>
<td>218</td>
<td>0.49000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- P-value is 0.4361 (as long as it's > 0.1 and the sample size is reasonably large, we can discard the additional variables)
CS Example (3)

- Can also obtain the numbers we need from TYPE I / III Sums of Squares.
- How would we test...
  - Importance of HSS in addition to rest.
  - Importance of SAT’s added to HS’s
  - Importance of HSE after HSM/HSS
  - Importance of HSE after HSM
- Can obtain the numbers you need for any partial-F test by arranging the variables correctly.
Upcoming in Lecture 14….

- Diagnostics and Remedial Measures