Topic 11 - General Linear Test

STAT 525 - Fall 2013

Outline

- Extra Sums of Squares
- Partial correlations
- Standardized regression coefficients

General Linear Test

- Comparison of a **full** model and **reduced** model that involves a subset of full model predictors (i.e., hierarchical structure)
- Involves a comparison of unexplained SS
- Consider a full model with k predictors and reduced model with l predictors ($l < k$)
- Can show that
$$F^\star = \frac{(\text{SSE(R)} - \text{SSE(F)})/(k-l)}{\text{SSE(F)}/(n-k-1)}$$
- Degrees of freedom for F^* are the number of **extra** variables and the error degrees of freedom for the larger model

General Linear Test

- Testing the Null hypothesis that the regression coefficients for the **extra** variables are all zero.
- Examples:
  - X_1, X_2, X_3, X_4 vs $X_1, X_2 \longrightarrow \beta_3 = \beta_4 = 0$
  - X_1, X_2, X_4 vs $X_1 \longrightarrow \beta_2 = \beta_4 = 0$
  - X_1, X_2, X_3, X_4 vs $X_1 \longrightarrow \beta_2 = \beta_3 = \beta_4 = 0$
- Because SSM+SSE=SSTO, can also compare using explained SS (SSM)

Notation for Extra SS

- Consider $H_0 : X_1, X_3$ vs $H_a : X_1, X_2, X_3, X_4$
- Null can also be written $H_0 : \beta_2 = \beta_4 = 0$
- Write SSE(F) as SSE(X_1, X_2, X_3, X_4)
- Write SSE(R) as SSE(X_1, X_3)
- Difference in SSE's is the **extra SS**
- Write as

$$\mathrm{SSE(X_2, X_4|X_1, X_3) = SSE(X_1, X_3) - SSE(X_1, X_2, X_3, X_4)}$$

- Recall SSM can also be used

General Linear Test

- Can rewrite F test as

$$F^{\star} = \frac{\mathrm{SSE(X_2, X_4|X_1, X_3)/(4-2)}}{\mathrm{SSE(X_1, X_2, X_3, X_4)/(n-5)}}$$

- Under H_0 $F^* \sim F(2, n-5)$
- If reject, conclude either X_2 or X_4 or both contain additional useful information to predict Y in a linear model with X_1 and X_3
- Example: Consider predicting GPA with HS grades, do SAT scores add any useful information?

Special Cases

- Consider test based on

$$\mathrm{SSE(X_i|X_1, ..., X_{i-1}, X_{i+1}, ...X_{p-1})}$$

- These are SAS's indiv parameter t-tests

$$F(1, n-p) = t^2(n-p)$$

- Decomposition of SSM(X_1, X_2, X_3)

$$\begin{aligned} &= \mathrm{SSM(X_1) + SSM(X_2|X_1) + SSM(X_3|X_2, X_1)} \\ &= \mathrm{SSM(X_2) + SSM(X_1|X_2) + SSM(X_3|X_2, X_1)} \\ &= \mathrm{SSM(X_3) + SSM(X_2|X_3) + SSM(X_1|X_2, X_3)} \end{aligned}$$

- Can decompose SSM variety of ways
- Stepwise sum of squares called Type I SS

Example Page 256

- Twenty healthy female subjects
- Y is body fat via underwater weighing
- Underwater weighing expensive/difficult
- X_1 is triceps skinfold thickness
- X_2 is thigh circumference
- X_3 is midarm circumference

SAS code

```
options nocenter;
data a1;
  infile 'U:\Ch07ta01.txt';
  input skinfold thigh midarm fat;

proc reg data=a1;
  model fat=skinfold thigh midarm /ss1 ss2;
run;

proc reg data=a1;
  model fat=skinfold;
run;

proc reg data=a1;
   model fat=skinfold thigh midarm;
   thimid: test thigh, midarm;
run;
```


Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	117.08469	99.78240	1.17	0.2578
skinfold	1	4.33409	3.01551	1.44	0.1699
thigh	1	-2.85685	2.58202	-1.11	0.2849
midarm	1	-2.18606	1.59550	-1.37	0.1896

Conclusions

- Set of three variables helpful in predicting body fat ($P < 0.0001$)
- None of the indiv parameters significant
  - Addition of each predictor to a model containing the other two is not helpful
  - Example of **multicollinearity**
  - Will discuss more in next topic
- Will now focus on extra SS

Output

Parameter Estimates

Variable	DF	Parameter Estimate	Type I SS	Type II SS
Intercept	1	117.08469	8156.76050	8.46816
skinfold	1	4.33409	352.26980	12.70489
thigh	1	-2.85685	33.16891	7.52928
midarm	1	-2.18606	11.54590	11.54590

Interpretation

- Type I and Type II very different
- Type I depends on model statement
- In this example the SS are:

Type I	Type II
$SSM(X_1)$	$SSM(X_1 \mid X_2, X_3)$
$SSM(X_2 \mid X_1)$	$SSM(X_2 \mid X_1, X_3)$
$SSM(X_3 \mid X_1, X_2)$	$SSM(X_3 \mid X_1, X_2)$

- Could variables be explaining same SS and "canceling" each other out?
- Look at other models / general linear test

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

Root MSE	2.81977	R-Square	0.7111
Dependent Mean	20.19500	Adj R-Sq	0.6950
Coeff Var	13.96271		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.49610	3.31923	-0.45	0.6576
skinfold	1	0.85719	0.12878	6.66	<.0001

** Skinfold now helpful. Note the change in coefficient estimate and standard error compared to the full model

Output

- Does this variable alone do the job?
- Perform general linear test

Test thimid Results for Dependent Variable fat

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	22.35741	3.64	0.0500
Denominator	16	6.15031		

**Appears there is additional information in the variables. Perhaps the addition of one more variable would be helpful.

Partial Correlations

- Measures the strength of a linear relation between two variables taking into account other variables or after adjusting for other variables
- Procedure for X_i vs Y
  - Predict Y using other X's
  - Predict X_i using other X's
  - Find correlation between residuals
- Each residual represents what is not explained by the other variables
- Looking for additional information in X_i that better explains Y

SAS code and Output

```
proc reg data=a1;
   model fat=skinfold thigh midarm / pcorr2;
run;
                 Parameter Estimates
                                  Squared
                 Parameter        Partial
Variable   DF     Estimate   Corr Type II
Intercept   1    117.08469              .
skinfold    1      4.33409        0.11435
thigh       1     -2.85685        0.07108
midarm      1     -2.18606        0.10501
```

** Partial squared correlation also called coefficient of partial determination. Has similar interpretation.

In this case, variables only explain approximately 10% of the remaining variability after the other two variables are fit

Standardized Regression Model

- Can reduce round-off errors in calculations
- Standardization

$$Y_i' = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \overline{Y}}{s_Y}\right) \quad \text{and} \quad X_{ik}' = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \overline{X}_i}{s_{X_i}}\right)$$

- Puts regression coefficients in common units
- A one SD change in X_i' corresponds to β_i' SD increase in Y
- Can show

$$\beta_i = \left(\frac{s_Y}{s_{X_i}}\right)\beta_i'$$

SAS code and Output

```
proc reg data=a1;
   model fat=skinfold thigh midarm / stb;
run;

                       Parameter Estimates
                Parameter  Standardized
Variable   DF    Estimate      Estimate
Intercept   1   117.08469             0
skinfold    1     4.33409       4.26370
thigh       1    -2.85685      -2.92870
midarm      1    -2.18606      -1.56142
```

**Skinfold has highest standardized coefficient. Midarm does not appear to be as important a predictor. Perhaps best model includes skinfold and thigh.

Background Reading

- KNNL Sections 7.1-7.5
- knnl256.sas
- KNNL Sections 7.6-7.7