STAT 525 FALL 2018

Chapter 7
General Linear Test and Multicollinearity

Professor Min Zhang

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

- 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 $X_1, X_2 \longrightarrow H_0 : \beta_3 = \beta_4 = 0$
  - X_1, X_2, X_4 vs $X_1 \longrightarrow H_0 : \beta_2 = \beta_4 = 0$
  - X_1, X_2, X_3, X_4 vs $X_1 \longrightarrow H_0 : \beta_2 = \beta_3 = \beta_4 = 0$
- Because SSM+SSE=SSTO, can also compare using explained SS (SSM)

Extra SS and Notation

- 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

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

- Recall SSM can also be used

$$\begin{aligned} \text{SSM}(X_2, X_4 | X_1, X_3) &= \text{SSM}(X_1, X_2, X_3, X_4) - \text{SSM}(X_1, X_3) \Longrightarrow \\ \text{SSM}(X_1, X_2, X_3, X_4) &= \text{SSM}(X_1, X_3) + \text{SSM}(X_2, X_4 | X_1, X_3) \end{aligned}$$

General Linear Test in Terms of Extra SS

- Can rewrite F test as

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

- 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 testing individual predictor X_i based on

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

  - These are related to SAS's indiv parameter t-tests

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

- Can decompose SSM variety of ways
  - Decomposition of SSM(X_1,X_2,X_3)

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

  - Stepwise sum of squares called Type I SS

Type I SS and Type II SS

- Type I and Type II are very different
  - Type I is sequential, so it depends on model statement
  - Type II is conditional on all others, so it does not depend on model statement
- For example,

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

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

Example: Body Fat (p.256)

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

- Investigate the model with all three predictors:

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

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

```
                      Analysis of Variance
                          Sum of          Mean
Source             DF     Squares        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
                     Parameter     Standard
Variable     DF       Estimate        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 individual parameters is 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 Using SS1 & SS2

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

- Investigate the model: `fat=skinfold`

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

```
                   Analysis of Variance
                          Sum of        Mean
Source             DF     Squares      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
                 Parameter   Standard
Variable    DF   Estimate       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.

- Does this variable alone do the job?

- Perform general linear test

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

```
     Test thimid Results for Dependent Variable fat

                                        Mean
Source                  DF            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 <u>additional</u> information in X_i that better explains Y

Example: Body Fat

```
proc reg data=a1;
    model fat=skinfold thigh midarm / pcorr2;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	1	117.08469	99.78240	1.17	0.2578	.
skinfold	1	4.33409	3.01551	1.44	0.1699	0.11435
thigh	1	-2.85685	2.58202	-1.11	0.2849	0.07108
midarm	1	-2.18606	1.59550	-1.37	0.1896	0.10501

- Squared partial correlation is also called coefficient of partial determination. Has similar interpretation to coefficient of multiple determination.
- In this case, variables only explain approximately 10% of the remaining variation after the other two variables are fit.

Standardized Regression Model

- Can reduce round-off errors in calculations
- Standardization

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

- Puts regression coefficients in common units
- A one SD change in X_i corresponds to $\tilde{\beta}_i$ SD increase in Y
- Can show

$$\beta_i = \left(\frac{s_Y}{s_{X_i}}\right)\tilde{\beta}_i$$

Example: Body Fat

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

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	117.08469	99.78240	1.17	0.2578	0
skinfold	1	4.33409	3.01551	1.44	0.1699	4.26370
thigh	1	-2.85685	2.58202	-1.11	0.2849	-2.92870
midarm	1	-2.18606	1.59550	-1.37	0.1896	-1.56142

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

Multicollinearity

- Numerical analysis problem is that the matrix $\mathbf{X}'\mathbf{X}$ is almost singular (linear dependent columns)
  - Makes it difficult to take the inverse
  - Generally handled with current algorithms
- Statistical problem: too much correlation among predictors
  - Difficult to determine regression coefficients $\longrightarrow$ Increased variance
- Want to refine model to remove redundancy in the predictors

Example

- Consider a two predictor model

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

- What is the estimate of β_1?
- Can show

$$b_1 = \frac{\tilde{b}_1 - \sqrt{\frac{s_Y^2}{s_{X_1}^2}} r_{12} r_{Y2}}{1 - r_{12}^2}$$

where $\tilde{b}_1$ is the estimate fitting Y vs X_1

Extreme Cases

- Consider X_1 and X_2 are uncorrelated
  - r_{12}=0
  - $b_1 = \tilde{b}_1$ (fitting Y vs X_1)
  - Estimator b_1 does not depend on X_2
  - Type I SS and Type II SS are the same
  - In other words, the contribution of each predictor is the same regardless of whether or not the other predictor is in the model

- Consider $X_1 = a + bX_2$
  - $r_{12} = \pm 1$
  - Estimator b_1 does not exist
  - Type II SS are zero
  - In other words, there is no contribution of the predictor if the other predictor is already in the model

Extreme Case in SAS

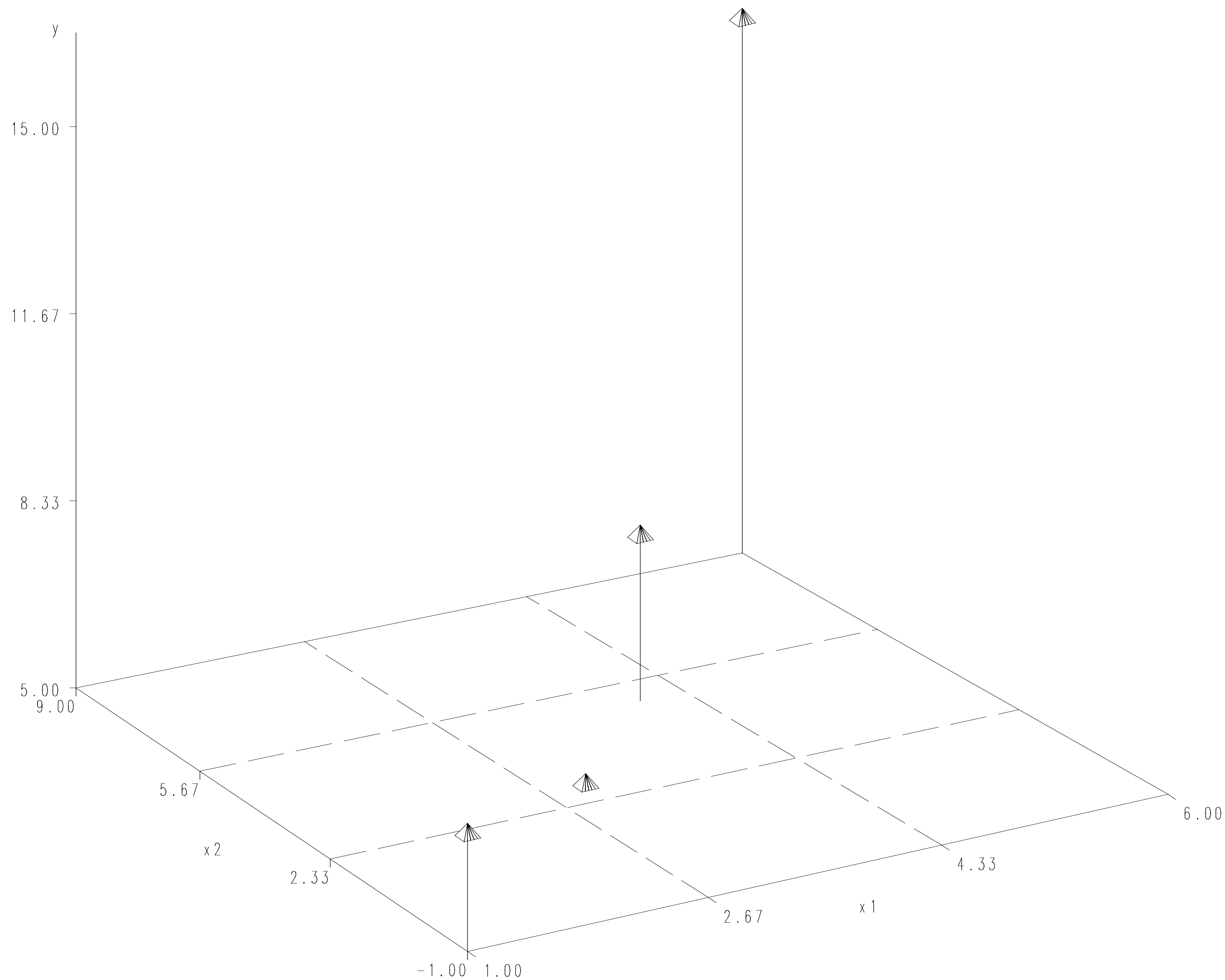
- Consider the following data set

```
data a1;
    input case x1 x2 y @@@@;
    cards;
    1  3  3  5
    2  4  5  8
    3  1 -1  7
    4  6  9 15
;
```

- Notice $x_2 = 2x_1 - 3$

- Will generate 3-D plot and run regression

```
/* Generate 3-D Scatterplot */
proc g3d data=a1;
    scatter x2*x1=y / rotate=30;
run;
```



```
proc reg data=a1;
    model y=x2 x1;
run; quit;
```

```
                      Analysis of Variance
                        Sum of       Mean
Source           DF    Squares     Square  F Value  Pr > F
Model             1   55.59211   55.59211    96.02  0.0103
Error             2    1.15789    0.57895
Corrected Total   3   56.75000

Root MSE              0.76089    R-Square     0.9796
Dependent Mean        8.75000    Adj R-Sq     0.9694
Coeff Var             8.69584

NOTE: Model is not full rank. Least-squares solutions for the parameters are
      not unique. Some statistics will be misleading. A reported DF of 0 or B
      means that the estimate is biased.
NOTE: The following parameters have been set to 0, since the variables are a
      linear combination of other variables as shown.

x1 =  1.5 * Intercept + 0.5 * x2

                 Parameter Estimates
              Parameter   Standard
Variable   DF  Estimate      Error  t Value   Pr > |t|
Intercept   B  -0.65789    1.03271    -0.64     0.5893
x2          B   1.71053    0.17456     9.80     0.0103
x1          0         0          .        .          .
```

- In this example, no inverse exists so X_1 dropped
- In practice, we are concerned with less extremal cases
- General results still hold
  - Regression coefficients are not well estimated
  - Regression coefficients may be meaningless
  - Type I SS and II SS will differ substantially
  - R^2 and predicted values are usually ok

Pairwise Correlations

- Assesses "pairwise collinearity" but not complicated multi-collinearity
- Consider our body fat example

```
proc reg data=a1 corr;
    model midarm = skinfold thigh;
run; quit;
```

Correlation

Variable	skinfold	thigh	midarm	fat
skinfold	1.0000	0.9238	0.4578	0.8433
thigh	0.9238	1.0000	0.0847	0.8781
midarm	0.4578	0.0847	1.0000	0.1424
fat	0.8433	0.8781	0.1424	1.0000

  – None of these are too troublesome

- "`MODEL midarm = skinfold thigh`" reported $R^2 = 0.9904$
  – All three $\rightarrow r = \sqrt{0.9904} = .995$
  – Should not use model with all three predictors

Coefficient Estimation

- Page 284 summarizes coefficients

Variables in Model	b_1	b_2
skinfold	0.8572	-
thigh	-	0.8565
skinfold, thigh	0.2224	0.6594
skinfold, thigh, midarm	4.3340	-2.857

- `skinfold` and `thigh` similar info
- Coeffs change when both are included (sum ≈ 0.86)
- Very dramatic change when `midarm` is in
- Reflected in std errors too

Chapter Review

- Extra Sums of Squares
- Partial correlations
- Standardized regression coefficients
- Multicollinearity
  - Effects
  - Remedies