Topic 12 - Multicollinearity

STAT 525 - Fall 2013

Outline

- Multicollinearity
  - Effects
  - Remedies
- Polynomial Regression
- Interaction Models

Body fat Determination Page 256

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

Full model 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	Significant F test
Dependent Mean	20.19500	Adj R-Sq	0.7641	but no significant
Coeff Var	12.28017			t tests
				******************

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

Multicollinearity

- Numerical analysis problem: The matrix $\mathbf{X}'\mathbf{X}$ is almost singular (linear dependent columns - no inverse exists)
- Previously calculation of inverse was difficult
- Now generally handled well with current algorithms
- Statistical problem: Very high correlation among the explanatory variables
- While the inverse exists, regression coefficients very unstable
  - Increased uncertainty / variance
  - Spurious coefficient estimates

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

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

where b_1' is the estimate fitting Y vs X_1

Example continued

- Extreme case #1: Consider X_1 and X_2 are uncorrelated
  - r_{12}=0
  - $b_1 = b_1'$ (fitting Y vs X_1)
  - Estimator b_1 does not depend on X_2
  - Type I and Type II SS 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
- Extreme case #2: 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 #2 in SAS

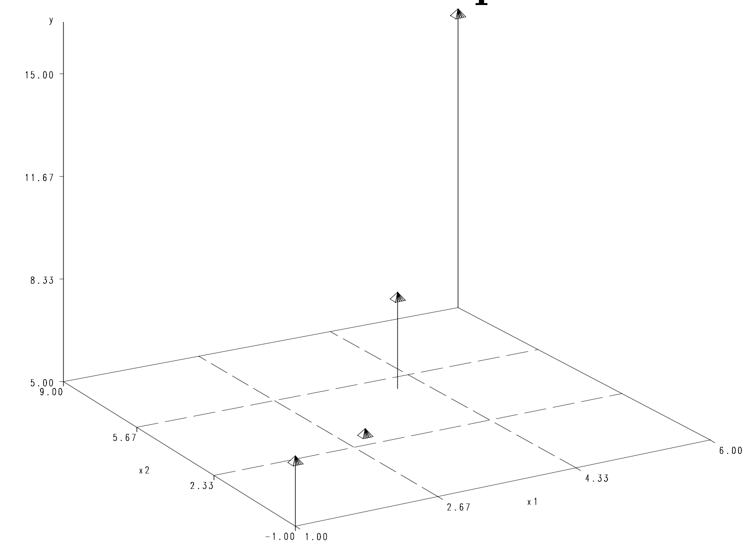
- Consider the following data set

Case	X_1	X_2	Y
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

```
proc g3d;
 scatter x2*x1=y / rotate=30;
run;

proc reg;
 model y=x2 x1;
run;
```

3-D Scatterplot



Regression output

Source	DF	Sum of Squares	Mean 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			

```
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

Variable	DF	Parameter Estimate	Standard 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	.	.	.

Summary of extreme case #2

- For this example, no inverse exists so SAS dropped X_1 in order to obtain estimates
- Could have as easily dropped X_2
- Not a unique solution...this is what is meant by "B" in the SAS output

- In practice, concerned with cases less extreme
- General results still hold
  - Regression coefficients not well estimated (imprecise)
  - Regression coefficients may be meaningless (spurious)
  - Type I and II SS will differ substantially
  - R^2 and predicted values usually ok

Pairwise Correlations

- Good to assess but don't always detect the issue
- Consider our body fat example

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

  - $r(\text{skinfold}, \text{thigh}) = 0.9218$
  - $r(\text{skinfold}, \text{midarm}) = 0.4578$
  - $r(\text{thigh}, \text{midarm}) = 0.0847$

- None of these are too troublesome
- Consider all three together $\rightarrow r = \sqrt{0.9904} = .995$
- Will describe alternative methods to detect issue soon

Example of issue on coefficient estimation

- Page 284 summarizes coefficients in this example

Model	b_1	b_2
X_1	0.8572	-
X_2	-	0.8565
X_1, X_2	0.2224	0.6594
X_1, X_2, X_3	4.3340	-2.857

- X_1 and X_2 contain similar information
- Coeffs change when both in but not too dramatically
- Very dramatic change when X_3 added (negative estimate for b_2)
- Dramatic change reflected in std errors too

Polynomial Regression

- Multiple regression using powers of X (e.g., X^2, X^3) as additional predictors
- Fitting of such models can often lead to a multicollinearity problem unless original variable is **centered**
- Centering

$$X'_i = X_i - \overline{X}$$

Example Page 300

- Response variable is the life (in cycles) of a power cell
- Explanatory variables are
  - Charge rate (3 levels)
  - Temperature (3 levels)
- This is a designed experiment
- Notice $\sum(X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2) = 0 \rightarrow r(X_1, X_2) = 0$
- Coded values are standardized ($x_{ij} = 0, \pm 1$)
  - Notice $\sum x_{i1}x_{i2} = 0 \rightarrow r(x_1, x_2) = 0$
  - Notice $\sum x_{i1}x_{i1}^2 = 0 \rightarrow r(x_1, x_1^2) = 0$
  - Notice $\sum x_{i2}x_{i2}^2 = 0 \rightarrow r(x_2, x_2^2) = 0$

SAS CODE

Creating New Variables in SAS

```
data a1;
   infile 'U:\.www\datasets525\Ch07ta09.txt';
   input cycles chrate temp;

data a1; set a1;
  chrate2=chrate*chrate;
  temp2=temp*temp;
  ct=chrate*temp;

proc reg data=a1;
   model cycles=chrate temp chrate2 temp2 ct;
run;

proc corr data=a1;
   var chrate temp chrate2 temp2 ct;
run;
```

Output

```
                    Analysis of Variance
                         Sum of          Mean
Source             DF   Squares        Square  F Value  Pr > F
Model               5     55366         11073    10.57  0.0109
Error               5  5240.43860  1048.08772
Corrected Total    10     60606
                                                          *******************
Root MSE           32.37418    R-Square     0.9135    Significant F test
Dependent Mean    172.00000    Adj R-Sq     0.8271    no significant
Coeff Var          18.82220                           t tests
                                                          *******************
                Parameter    Standard
Variable    DF   Estimate       Error  t Value  Pr > |t|
Intercept    1  337.72149   149.96163     2.25    0.0741
chrate       1 -539.51754   268.86033    -2.01    0.1011
temp         1    8.91711     9.18249     0.97    0.3761
chrate2      1  171.21711   127.12550     1.35    0.2359
temp2        1   -0.10605     0.20340    -0.52    0.6244
ct           1    2.87500     4.04677     0.71    0.5092
```


Correlation matrix

```
        Pearson Correlation Coefficients, N = 11
                Prob > |r| under H0: Rho=0
           chrate       temp    chrate2      temp2         ct
chrate    1.00000    0.00000    0.99103    0.00000    0.60532
                      1.0000     <.0001     1.0000     0.0485
temp      0.00000    1.00000    0.00000    0.98609    0.75665
           1.0000                1.0000     <.0001     0.0070
chrate2   0.99103    0.00000    1.00000    0.00592    0.59989
           <.0001     1.0000                0.9862     0.0511
temp2     0.00000    0.98609    0.00592    1.00000    0.74613
           1.0000     <.0001     0.9862                0.0084
ct        0.60532    0.75665    0.59989    0.74613    1.00000
           0.0485     0.0070     0.0511     0.0084
```

**As anticipated, high correlation between X_1 and X_1^2 as well as X_2 and X_2^2

SAS CODE

Standardizing (centering) in SAS

```
data a2; set a1;
   schrate=chrate; stemp=temp;
   keep cycles schrate stemp;

proc standard data=a2 out=a3 mean=0 std=1;
   var schrate stemp;

proc print data=a3;     ***Centering most important here***
run;

data a3; set a3;
  schrate2=schrate*schrate; stemp2=stemp*stemp; sct=schrate*stemp;

proc reg data=a3;
  model cycles=schrate stemp schrate2 stemp2 sct;
run;
```


Standardized Values

Obs	cycles	schrate	stemp
1	150	-1.29099	-1.29099
2	86	0.00000	-1.29099
3	49	1.29099	-1.29099
4	288	-1.29099	0.00000
5	157	0.00000	0.00000
6	131	0.00000	0.00000
7	184	0.00000	0.00000
8	109	1.29099	0.00000
9	279	-1.29099	1.29099
10	235	0.00000	1.29099
11	224	1.29099	1.29099

Output after Centering

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	55366	11073	10.57	0.0109
Error	5	5240.43860	1048.08772		
Corrected Total	10	60606			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	162.84211	16.60761	9.81	0.0002
schrate	1	-43.24831	10.23762	-4.22	0.0083
stemp	1	58.48205	10.23762	5.71	0.0023
schrate2	1	16.43684	12.20405	1.35	0.2359
stemp2	1	-6.36316	12.20405	-0.52	0.6244
sct	1	6.90000	9.71225	0.71	0.5092

***Notice that this is the same overall F test but now the two "main" effects are significant. A linear model appears reasonable. Could do a general linear test (test schrate,stemp2, sct;). Notice also that the P-values here are the same for the coded variable analysis but the coefficients are different.

Interaction Models

- With several explanatory variables, we need to consider the possibility that the effect of one variable depends on the value of another variable
- Model this relationship as the product of predictors
- Special Cases:
  - One binary (Y/N) and one continuous
  - Two continuous predictors

Special Case #1

- $X_1 = 0$ or 1 identifying two groups
- X_2 is a continuous variable

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

- When $X_1 = 0$ (Group 1)

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

- When $X_1 = 1$ (Group 2)

$$Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_{i2} + \varepsilon_i$$

Special Case #1

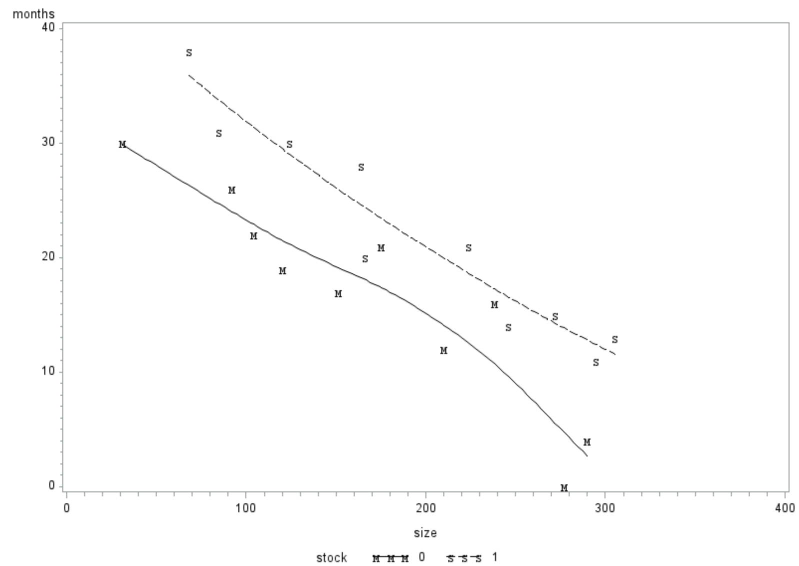
- Results in two regression lines
- β_2 is the slope for Group 1
- $\beta_2 + \beta_3$ is the slope for Group 2
- Similar relationship for the intercepts
- Three Hypotheses of Interest
  - $H_0 : \beta_1 = \beta_3 = 0 :$ regression lines are the same
  - $H_0 : \beta_1 = 0 :$ intercepts are the same
  - $H_0 : \beta_3 = 0 :$ slopes are the same

Example Page 314

- Y is the number of months for an insurance company to adopt an innovation
- X_1 is the size of the firm
- X_2 is the type of firm
  - $X_2 = 0 \rightarrow$ mutual fund firm
  - $X_2 = 1 \rightarrow$ stock firm
- Do stock firms adopt innovation faster? Is this true regardless of size?

SAS code

```
data a1;
   infile 'U:\.www\datasets525\Ch8ta02.txt';
   input months size stock;

symbol1 v=M i=sm70 c=black l=1;
symbol2 v=S i=sm70 c=black l=3;
proc sort data=a1; by stock size;
proc gplot data=a1;
   plot months*size=stock/frame;
run;

data a1; set a1;
   sizestoc=size*stock;

proc reg data=a1;
   model months=size stock sizestoc;
   test stock, sizestock;
run;
```

Output

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1504.41904	501.47301	45.49	<.0001
Error	16	176.38096	11.02381		
Corrected Total	19	1680.80000			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	
Intercept	1	33.83837	2.44065	13.86	<.0001	
size	1	-0.10153	0.01305	-7.78	<.0001	
stock	1	8.13125	3.65405	2.23	0.0408	**Diff intercept
sizestoc	1	-0.00041714	0.01833	-0.02	0.9821	**Same slope

Test 1 Results for Dependent Variable months

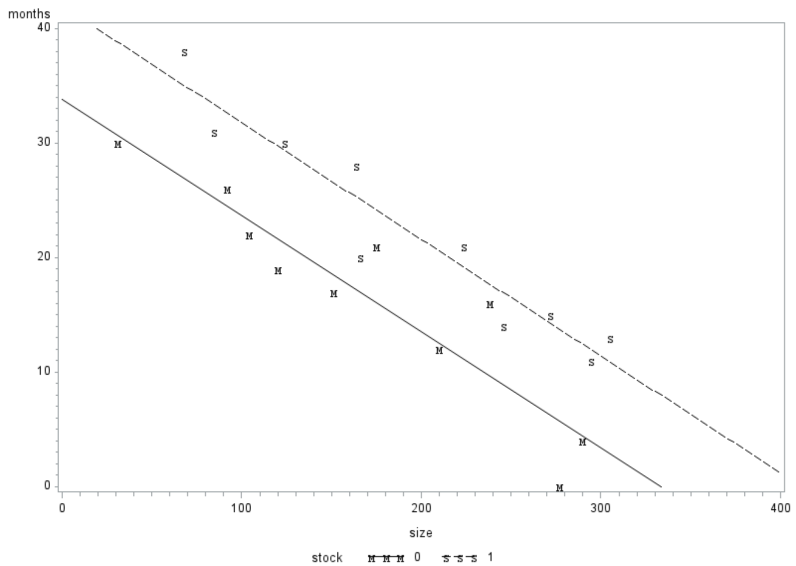
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	158.12584	14.34	0.0003 ***Not the same line
Denominator	16	11.02381		

Additional SAS code

```
proc reg data=a1;
   model months=size stock;     ***Same slope but different intercepts***
run;

symbol1 v=M i=rl c=black;
symbol2 v=S i=rl c=black;
proc gplot data=a1;
   plot months*size=stock/frame;
run;
```


Output

```
                    Analysis of Variance
                              Sum of        Mean
Source             DF        Squares      Square   F Value   Pr > F
Model               2     1504.41333   752.20667     72.50   <.0001
Error              17      176.38667    10.37569
Corrected Total    19     1680.80000

Root MSE              3.22113    R-Square     0.8951
Dependent Mean       19.40000    Adj R-Sq     0.8827
Coeff Var            16.60377

                   Parameter Estimates
                Parameter    Standard
Variable    DF   Estimate       Error  t Value    Pr > |t|
Intercept    1   33.87407     1.81386    18.68      <.0001
size         1   -0.10174     0.00889   -11.44      <.0001
stock        1    8.05547     1.45911     5.52      <.0001
```

Further Investigations

- When we fit both models together, we can allow for different slopes and/or intercepts but what do we assume is the same?

- Can use mixed model to assess if this is reasonable
- Compare fits of models where
  - Error variances are considered constant (OLS)
  - Error variances vary across stock type (Mixed)

Additional SAS code

```
***Standard model***;
proc mixed data=a1;
 class stock;
 model months = size stock / solution;
run;

***Two residual variance model;
proc mixed data=a1;
 class stock;
 model months = size stock / solution;
 repeated / group=stock;
run;
```


Output - Standard Model

```
Covariance Parameter Estimates
Cov Parm     Estimate                ****Variance estimate same as
Residual      10.3757                    with OLS. Parameter and t
                                         tests also the same ****
           Fit Statistics
-2 Res Log Likelihood           104.4
AIC (smaller is better)         106.4
AICC (smaller is better)        106.7
BIC (smaller is better)         107.2

                       Solution for Fixed Effects
                                  Standard
Effect       stock    Estimate       Error      DF    t Value    Pr > |t|
Intercept              41.9295      2.0101      17      20.86      <.0001
size                   -0.1017    0.008891      17     -11.44      <.0001
stock        0         -8.0555      1.4591      17      -5.52      <.0001
stock        1               0           .       .          .           .
```


Output - Different Variances Model

```
 Covariance Parameter Estimates
Cov Parm     Group      Estimate
Residual     stock 0     12.8735
Residual     stock 1      7.8006

          Fit Statistics
-2 Res Log Likelihood           103.8
AIC (smaller is better)         107.8
AICC (smaller is better)        108.7
BIC (smaller is better)         109.8   **Other model appears better

 Null Model Likelihood Ratio Test
   DF    Chi-Square      Pr > ChiSq  **Performs a likelihood test of
    1          0.56          0.4556     the two models

  Other model parameters are only slightly different in this case
```


Special Case #2

- X_1 and X_2 are continuous variables

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

- Can be written

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

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

- The coefficient of one explanatory variable depends on the value of the other explanatory variable
- Cannot discuss each predictor individually

Constrained Regression

- At times may want to put constraints on regression coefficients
  - $\beta_1 = 5$
  - $\beta_1 = \beta_2$
- Can do this in SAS by redefining explanatory variables
  - Page 268, wants to assess $\beta_1 = 5$ and $\beta_3 = 5$. Redefine so reduced model is Y' vs X_2
- Can also use restrict statement
  - Restrict X_1=5
  - Restrict $X_1 = X_2$

Background Reading

- KNNL Sections 7.4-7.9, Chapter 8
- knnl281.sas, knnl300.sas, knnl314.sas
- KNNL Chapter 9