Multicolinearity and singularity

With a perfect correlation an infinite number of models can be obtained. In practice, most software will bomb or detect the problem.

Sample problem (all perfectly correlated)

X1	X2	Y
1	1	2
2	2	4
3	3	6

this dataset could be fitted by

Y = 1*X1 + 1*X2
Y = 0*X1 + 2*X2
Y = 2*X1 + 0*X2
Y = 0.5*X1 + 1.5*X2
Y = 1.5*X1 + 0.5*X2
Y = 1.3*X1 + 0.7*X2
Y = 102*X1 - 100*X2
or any other model where $b_1 + b_2 = 2$

This results whenever two variables are perfectly correlated and there is a perfect fit with no error. It is clear that the regression coefficients cannot be interpreted.

What if the correlations are just high, not perfect?

1) We have no problem getting a good fit, but regressions coefficients will not be stable (they will vary widely from sample to sample). Also, the fact that the regression coefficients for each X are unstable makes prediction outside the range of that X variable untenable.

   As more variables are added to the model, the regression coefficients vary greatly, and the standard errors generally increase. However, even as the standard errors increase, the MSE decreases and the precision on the predicted value may be quite acceptable.

2) The whole idea of “holding one X constant" while varying another goes against the “high correlation" between variables. If we vary one, the other should vary in a predictable fashion as well.

   Suppose the variables “surface temperature" and “bottom temperature" are used to predict the abundance of shrimp. Since these vary together, how far can we realistically vary one while holding the other constant?

   Simple correlations may be a useful diagnostic for many situations, but this will not detect the most insidious multicolinearity problems. There are some more serious diagnostics.

The problem adversely affects estimates of regression coefficients their variances. Predicted values may be fine.

```
1          ************************************************************;
2          *** Applied Linear Statistical Models, 5th Edition, 2005      ***;
3          *** by Kutner, Nachtsheim, Neter, and Li, McGraw-Hill/Irwin   ***;
4          *** Table 7.1 - Bosy Fat Example                             ***;
5          ************************************************************;
6
7          dm'log;clear;output;clear';
8          options nodate nocenter nonumber ps=512 ls=99 nolabel;
9          ODS HTML style=minimal rs=none
9        ! body='C:\Geaghan\Current\EXST3201\Fall2005\SAS\BodyFat01.html' ;
NOTE: Writing HTML Body file:
C:\Geaghan\Current\EXST3201\Fall2005\SAS\BodyFat01.html
10
11         Title1 'Chapter 9 : Extra example of multicolinearity';
12
13         DATA BodyFat; INFILE CARDS MISSOVER;
14              TITLE1 'EXST7034 - Bodyfat Example, KNNL Table 7.1';
15              LABEL X1 = 'Triceps skinfold thickness';
16              LABEL X2 = 'Thigh circumference';
17              LABEL X3 = 'Midarm circumference';
18              LABEL Y = 'Body Fat';
19             INPUT SUBJECT X1 X2 X3 Y;
20         CARDS;
NOTE: The data set WORK.BODYFAT has 20 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.04 seconds
      cpu time            0.04 seconds
20       !        RUN;
41         ;
42
43         PROC SORT DATA=BodyFat; BY X1 X2 X3 Y; RUN;
NOTE: There were 20 observations read from the data set WORK.BODYFAT.
NOTE: The data set WORK.BODYFAT has 20 observations and 5 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.04 seconds
      cpu time            0.05 seconds
44         PROC PRINT DATA=BodyFat; VAR Y X1 X2 X3; TITLE2 'Raw Data Listing'; RUN;
NOTE: There were 20 observations read from the data set WORK.BODYFAT.
NOTE: The PROCEDURE PRINT printed page 1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.21 seconds
      cpu time            0.05 seconds
```

EXST7034 - Bodyfat Example, KNNL Table 7.1
Raw Data Listing

Obs	Y	X1	X2	X3		Obs	Y	X1	X2	X3
						10	19.3	25.5	53.5	24.8
						11	21.7	25.6	53.9	23.7
1	12.8	14.6	42.7	21.3		12	22.6	27.7	55.3	25.7
2	11.7	18.7	46.5	23.0		13	25.4	27.9	52.1	30.6
3	12.9	19.1	42.2	30.9		14	23.9	29.5	54.4	30.1
4	11.9	19.5	43.1	29.1		15	20.1	29.8	54.3	31.1
5	17.8	19.7	44.2	28.6		16	25.4	30.2	58.6	24.6
6	21.3	22.1	49.9	23.2		17	27.2	30.4	56.7	28.3
7	14.8	22.7	48.2	27.1		18	18.7	30.7	51.9	37.0
8	22.8	24.7	49.8	28.2		19	25.4	31.1	56.6	30.0
9	21.1	25.2	51.0	27.5		20	27.1	31.4	58.5	27.6

```
46         PROC REG DATA=BodyFat; TITLE2 'Simple Linear fits : Body Fat';
47              X1:MODEL  Y = X1;
48              X2:MODEL  Y = X2;
49              X3:MODEL  Y = X3; RUN;
NOTE: The PROCEDURE REG printed pages 2-4.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.17 seconds
      cpu time            0.10 seconds

EXST7034 - Bodyfat Example, KNNL Table 7.1
Simple Linear fits : Body Fat

The REG Procedure
Model: X1
Dependent Variable: Y

Number of Observations Read          20
Number of Observations Used          20

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

Parameter Estimates
                      Parameter       Standard
Variable     DF        Estimate          Error    t Value    Pr > |t|
Intercept     1        -1.49610        3.31923      -0.45      0.6576
X1            1         0.85719        0.12878       6.66      <.0001


Analysis of Variance
                                  Sum of           Mean
Source                     DF    Squares         Square    F Value    Pr > F
Model                       1  381.96582      381.96582      60.62    <.0001
Error                      18  113.42368        6.30132
Corrected Total            19  495.38950

Parameter Estimates
                      Parameter       Standard
Variable     DF        Estimate          Error    t Value    Pr > |t|
Intercept     1       -23.63449        5.65741      -4.18      0.0006
X2            1         0.85655        0.11002       7.79      <.0001


Analysis of Variance
                                  Sum of           Mean
Source                     DF    Squares         Square    F Value    Pr > F
Model                       1   10.05160       10.05160       0.37    0.5491
Error                      18  485.33790       26.96322
Corrected Total            19  495.38950

Parameter Estimates
                      Parameter       Standard
Variable     DF        Estimate          Error    t Value    Pr > |t|
Intercept     1        14.68678        9.09593       1.61      0.1238
X3            1         0.19943        0.32663       0.61      0.5491
```

```
55         PROC CORR DATA=BodyFat; TITLE2 'Correlation matrix';  var Y X1 X2 X3;
RUN;
NOTE: The PROCEDURE CORR printed page 6.
NOTE: PROCEDURE CORR used (Total process time):
      real time           0.13 seconds
      cpu time            0.04 seconds


EXST7034 - Bodyfat Example, KNNL Table 7.1
Correlation matrix

The CORR Procedure

   4  Variables:    Y        X1       X2       X3

Simple Statistics
Variable           N          Mean       Std Dev           Sum       Minimum       Maximum
Y                 20      20.19500       5.10619     403.90000      11.70000      27.20000
X1                20      25.30500       5.02326     506.10000      14.60000      31.40000
X2                20      51.17000       5.23461          1023      42.20000      58.60000
X3                20      27.62000       3.64715     552.40000      21.30000      37.00000


        Pearson Correlation Coefficients, N = 20
                Prob > |r| under H0: Rho=0

                Y            X1            X2            X3
Y         1.00000       0.84327       0.87809       0.14244
                         <.0001        <.0001        0.5491

X1        0.84327       1.00000       0.92384       0.45778
           <.0001                      <.0001        0.0424

X2        0.87809       0.92384       1.00000       0.08467
           <.0001        <.0001                      0.7227

X3        0.14244       0.45778       0.08467       1.00000
           0.5491        0.0424        0.7227



51         PROC REG DATA=BodyFat; TITLE2 'Multiple regression : Body Fat';
52            MODEL  Y = X1 X2 X3 / collin vif tol stb seqb;
53         RUN;
NOTE: The PROCEDURE REG printed page 5.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.16 seconds
      cpu time            0.09 seconds


EXST7034 - Bodyfat Example, KNNL Table 7.1
Multiple regression : Body Fat

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read          20
Number of Observations Used          20
```

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|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	117.08469	99.78240	1.17	0.2578	0	.	0
X1	1	4.33409	3.01551	1.44	0.1699	4.26370	0.00141	708.84291
X2	1	-2.85685	2.58202	-1.11	0.2849	-2.92870	0.00177	564.34339
X3	1	-2.18606	1.59550	-1.37	0.1896	-1.56142	0.00956	104.60601

SequentialParameter Estimates

Intercept	X1	X2	X3
20.195000	0	0	0
-1.496105	0.857187	0	0
-19.174246	0.222353	0.659422	0
117.084695	4.334092	-2.856848	-2.186060

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	X1	X2	X3
1	3.96796	1.00000	0.00000195	0.00000320	0.00000110	0.00000980
2	0.02052	13.90482	0.00037152	0.00132	0.00003262	0.00139
3	0.01151	18.56570	0.00059915	0.00021875	0.00032550	0.00693
4	0.00000865	677.37207	0.99903	0.99846	0.99964	0.99167

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error
X1	1	0.857187	0.12878079
X1 | X2	1	0.222353	0.30343892
X1 | X3	1	1.000585	0.12823209
X1 | X2,X3	1	4.334092	3.01551136
X2	1	0.856547	0.11001562
X2 | X1	1	0.659422	0.29118728
X2 | X3	1	0.850882	0.11244824
X2 | X1,X3	1	-2.856848	2.58201527
X3	1	0.199429	0.32662975
X3 | X1	1	-0.431442	0.17661556
X3 | X2	1	0.096029	0.16139267
X3 | X1,X2	1	-2.186060	1.59549900