STOR 455
STATISTICAL METHODS I

Jan Hannig
Guest Lecturer: Prof. Yufeng Liu

Regression Diagnostics

- Added-Variable plots
- Studentized deleted residuals (Y-outlier)
- Hat matrix leverage values (X-outlier)
- DFFITS, Cook's D, DFBETAS (Influential cases)
- Variance inflation factor (multicollinearity)

Life Insurance Example

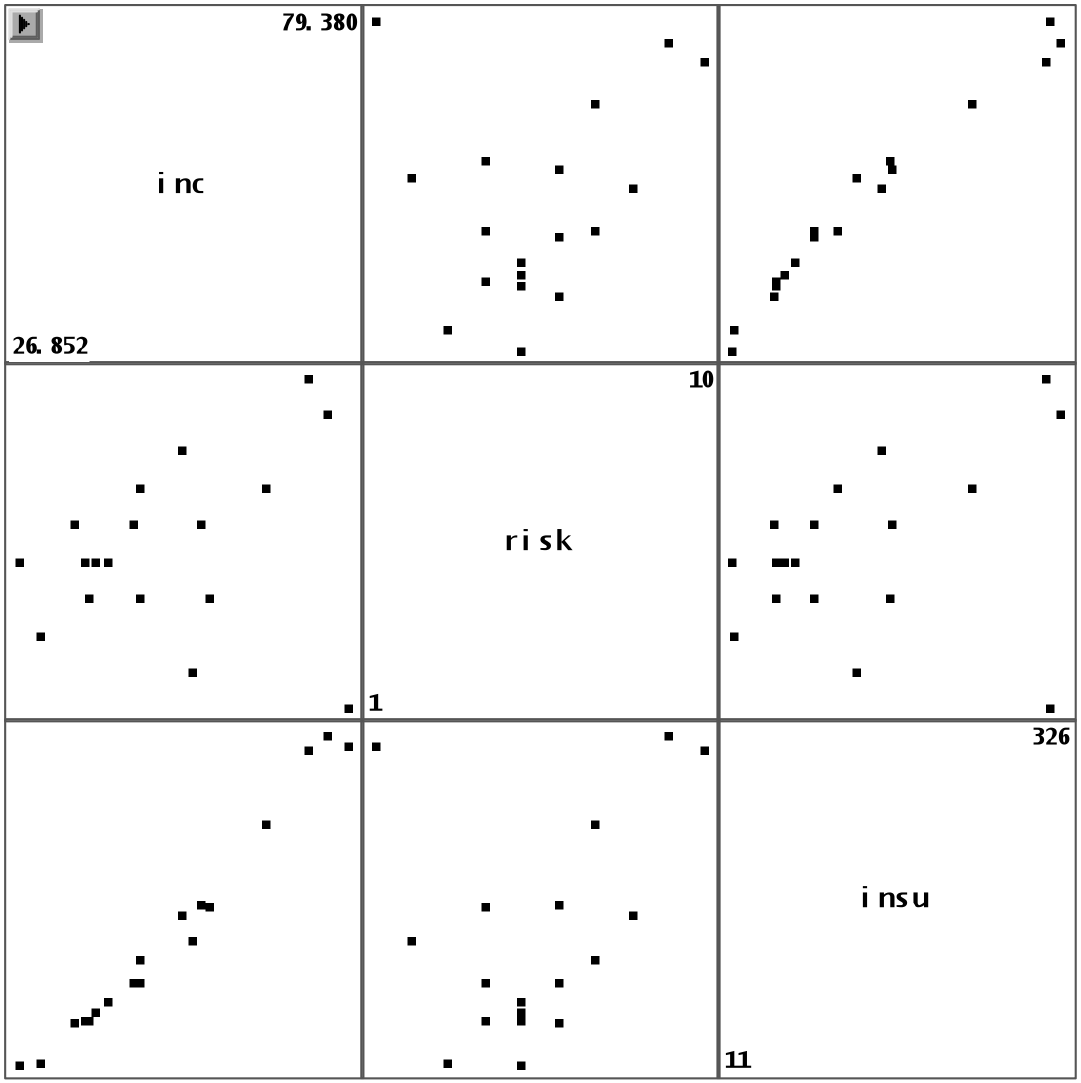
- Y: amount of life insurance
- X_1: average annual income
- X_2: a risk aversion score (higher means greater the degree of risk aversion)
- n = 18 managers

Do it in SAS

```
Data life;
 infile 'T:\…\life.txt';;
 input inc risk insu;
Proc print data=life;
run;
symbol1 v=dot h=.8 c=blue;

%include "C:\...
   \scatter.sas";
%scatter(data = life, var =
   inc risk insu);
```

Obs	inc	risk	insu
1	45.010	6	91
2	57.204	4	162
3	26.852	5	11
4	66.290	7	240
5	40.964	5	73
6	72.996	10	311
7	79.380	1	316
8	52.766	8	154
9	55.916	6	164
10	38.122	4	54
11	35.840	6	53
12	75.796	9	326
13	37.408	5	55
14	54.376	2	130
15	46.186	7	112
16	46.130	4	91
17	30.366	3	14
18	39.060	5	63

Do it in SAS



Added Variable Plots

- Partial regression for X_1
  - Use the other X's to predict Y
  - Use the other X's to predict X_1
  - Plot the residuals from the first regression vs the residuals from the second regression
- Can find multiple regression function from partial regressions

Added Variable Plots

- Also called partial regression plots or adjusted variable plots
- These plots can detect
  - Linear/Nonlinear relationships
  - Outliers

Do it in SAS

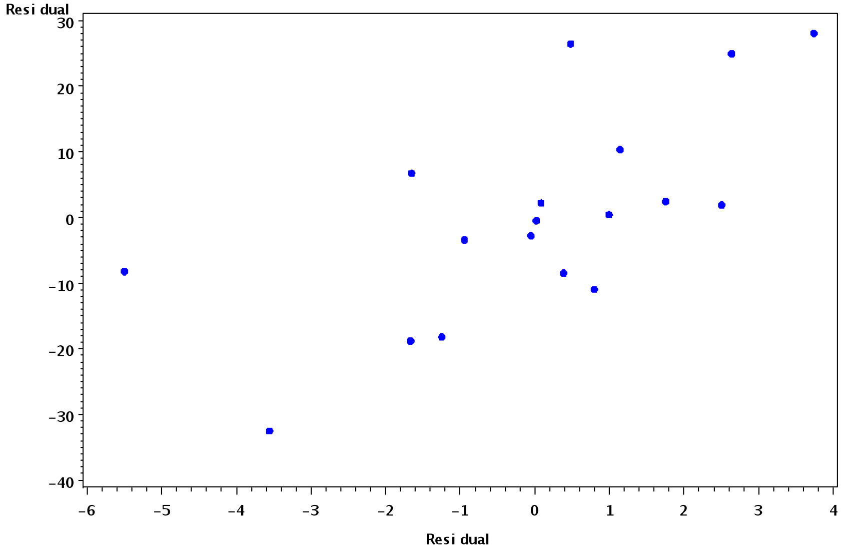
- The /partial option generates graphs in the output window
- OK for some purposes but can do better using proc gplot
- Need to generate residuals for gplot

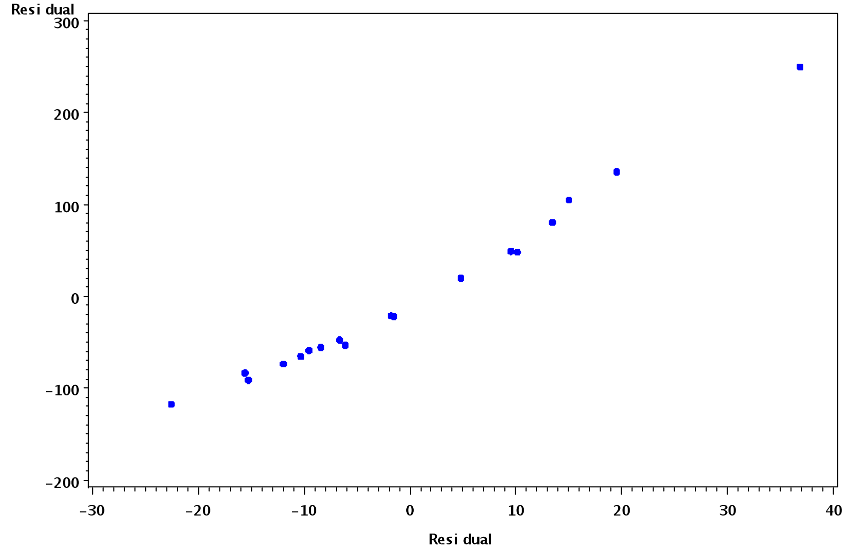
Do it in SAS

```
* added variable plot;
proc reg data = life ;
  model insu = inc risk /
   partial;
run;

* better looking added
   variable plot;
proc reg data=life;
  model insu risk = inc;
  output out=l2 r=resins
   resris;

proc reg data=l2;
   model resins=resris;
run;
```

```
proc gplot data=l2;
   plot resins*resris;
run;

proc reg data=life;
  model insu inc = risk;
  output out=l2 r=resins
   resinc;

proc gplot data=l2;
   plot resins*resinc;
run;

proc reg data=l2;
   model resins=resinc;
run;
```

Do it in SAS


Residual
30
20
10
0
−10
−20
−30
−40
−6
−5
−4
−3
−2
−1
0
1
2
3
4
Residual


Residual
300
200
100
0
−100
−200
−30
−20
−10
0
10
20
30
40
Residual

Do it in SAS: Parameter Estimators

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-205.71866	11.39268	-18.06	<.0001
inc	1	6.28803	0.20415	30.80	<.0001
risk	1	4.73760	1.37808	3.44	0.0037

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-8.4396E-15	2.88985	-0.00	1.0000
resris	Residual	1	4.73760	1.33432	3.55	0.0027

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.2561E-15	2.88985	0.00	1.0000
resinc	Residual	1	6.28803	0.19767	31.81	<.0001

Identifying Outliers

- Residuals $e_i = Y_i - Y_i$ (hat)
- semistudentized residuals ei/sqrt(MSE)
- Studentized residuals

  $r_i = e_i$ /sqrt(MSE(1 - h_{ii}))

Studentized Deleted Residuals

- We use the notation (i) to indicate that case i has been deleted from the computations
- $d_i = Y_i - Y_{(i)}$ (hat) is the deleted residual
- $MSE_{(i)}$ is the MSE with case i deleted
- The studentized deleted residual is

  $t_i = d_i$ / sqrt($MSE_{(i)}$ (1 - h_{ii}))

Use of Residuals

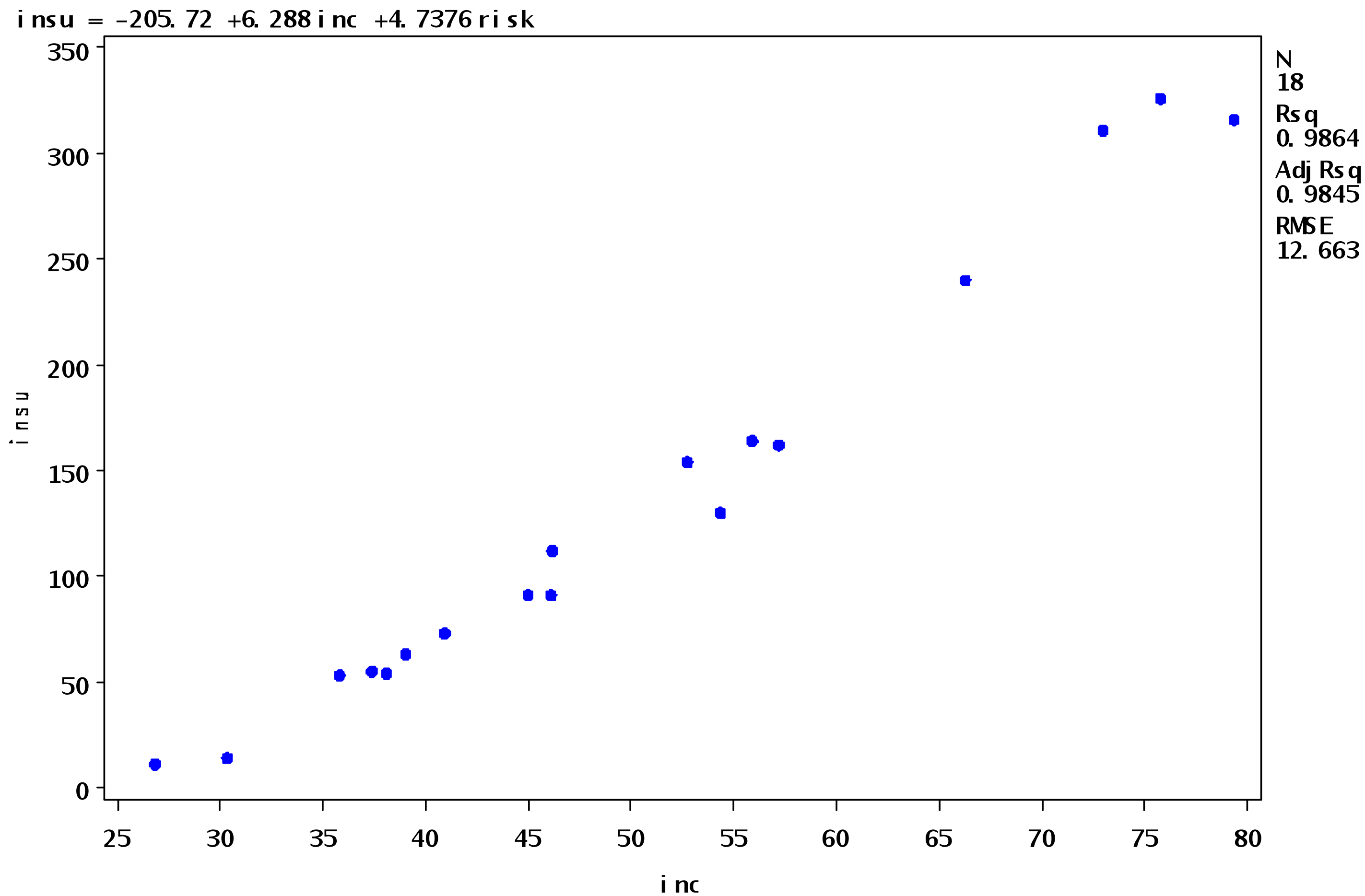
- We are looking for
  - Outliers (Bonferroni t-test)
  - Constant variance
  - Uncorrelated error
  - Normal error distributions

Do it in SAS: life insurance example

```
*obtain residuals;
proc reg data = life
  noprint ;
  model insu = inc
  risk  ;
  output out=l3
  r=residual h=hat
  student=rstudent;
  plot insu * inc (r.
  rstudent.) * p.
  rstudent.*(inc risk);
run;
proc print data = l3;
 var insu inc risk
  residual hat rstudent;
run;
```

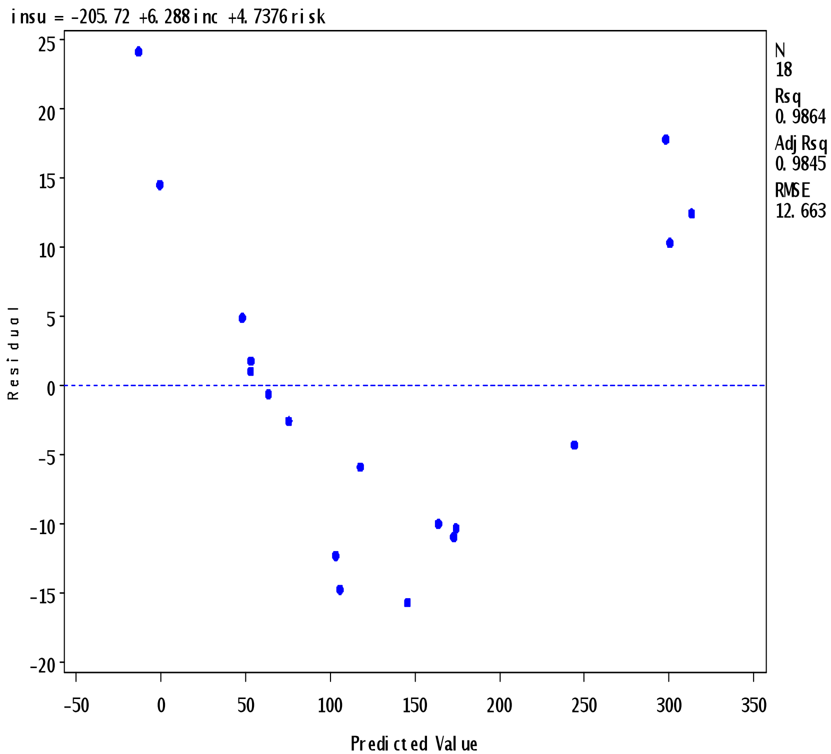
```
* add quadratic term;
data l4;
  set life;
  inc2=inc*inc;
proc print data=l4;
run;

proc reg data=l4;
  model insu = inc risk
  inc2/r;
  plot (rstudent.) * (p.
  inc risk nqq.);
run;
```

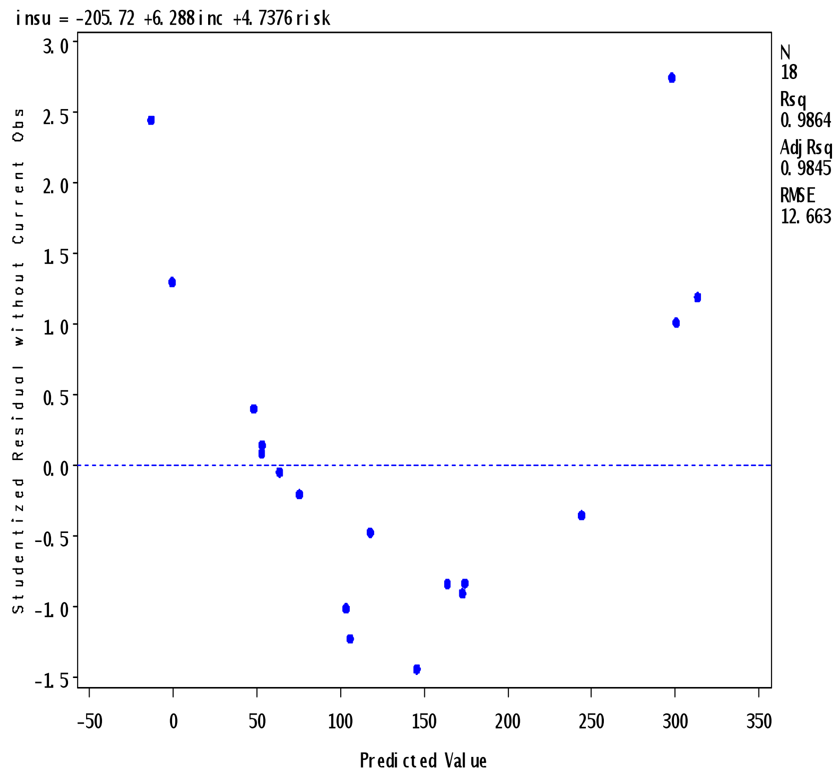
Do it in SAS



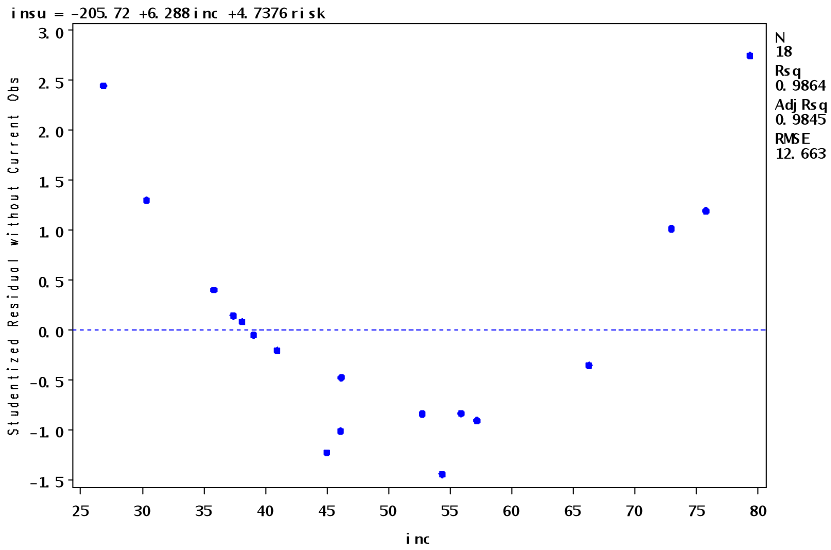
insu = -205.72 +6.288inc +4.7376risk
350
300
250
200
150
100
50
0
insu
25
30
35
40
45
50
55
60
65
70
75
80
inc
N
18
Rsq
0.9864
AdjRsq
0.9845
RMSE
12.663

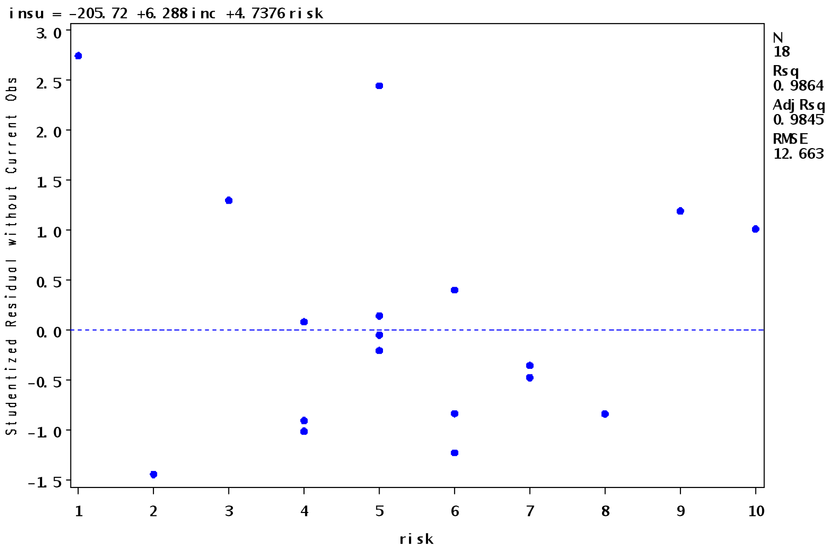
Residual vs Studentized Deleted Residual





Residual vs Explanatary Varaible





Do it in SAS

Obs	insu	inc	risk	residual	hat	rstudent
1	91	45.010	6	-14.7311	0.06929	-1.22593
2	162	57.204	4	-10.9321	0.10064	-0.90485
3	11	26.852	5	24.1845	0.18901	2.44867
4	240	66.290	7	-4.2780	0.13158	-0.35178
5	73	40.964	5	-2.5522	0.07559	-0.20282
6	311	72.996	10	10.3417	0.34986	1.01383
7	316	79.380	1	17.8373	0.62251	2.74827
8	154	52.766	8	-9.9763	0.13188	-0.83710
9	164	55.916	6	-10.3084	0.06575	-0.83363
10	54	38.122	4	1.0560	0.10052	0.08497
11	53	35.840	6	4.9301	0.12011	0.40331
12	326	75.796	9	12.4728	0.29940	1.19332
13	55	37.408	5	1.8081	0.09442	0.14507
14	130	54.376	2	-15.6744	0.20960	-1.44149
15	112	46.186	7	-5.8634	0.09569	-0.47419
16	91	46.130	4	-12.2985	0.07752	-1.01205
17	14	30.366	3	14.5636	0.18176	1.30042
18	63	39.060	5	-0.5798	0.08485	-0.04624

Add inc2

Analysis of Variance

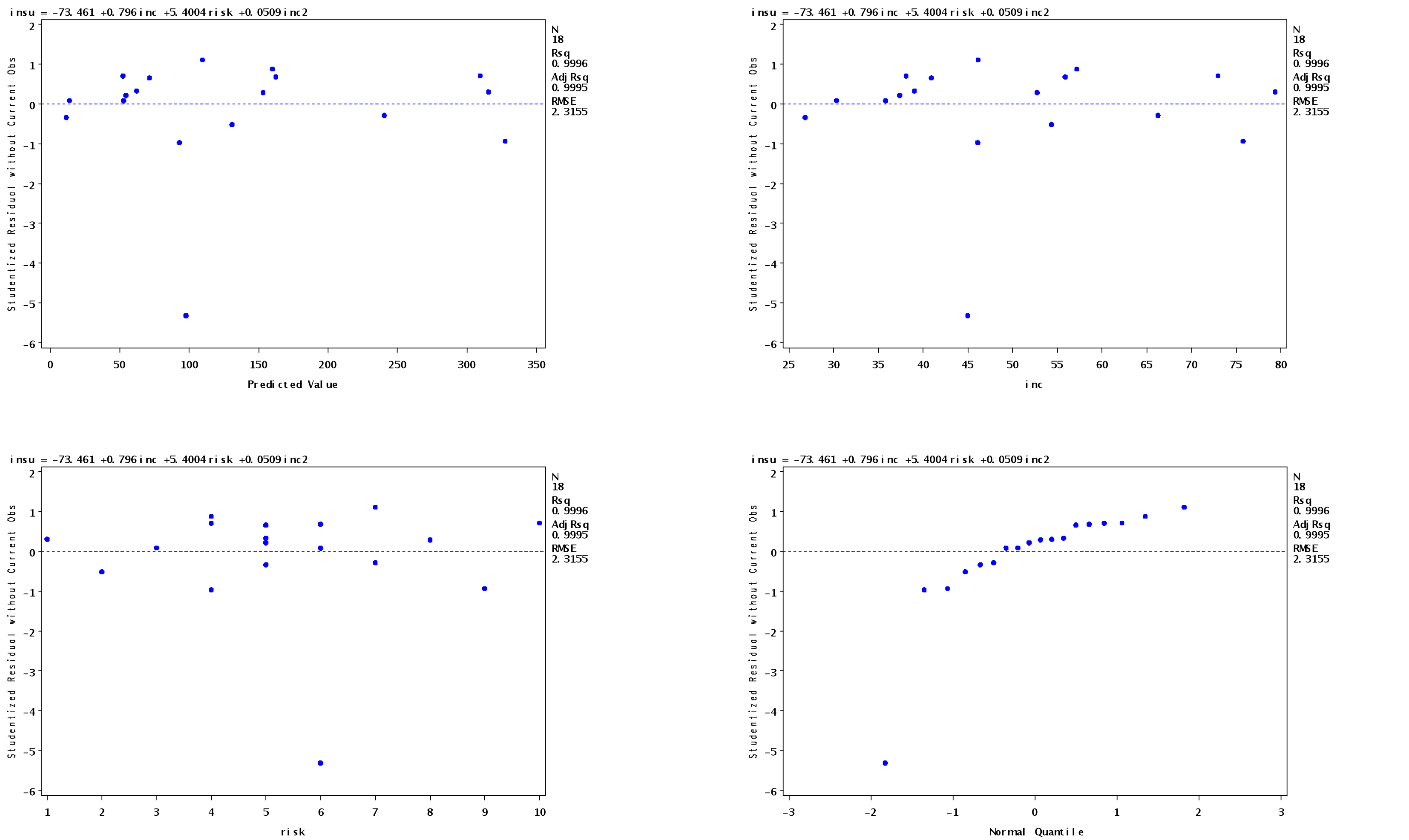
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	176249	58750	10958.0	<.0001
Error	14	75.05895	5.36135		
Corrected Total	17	176324			

Root MSE	2.31546	R-Square	0.9996
Dependent Mean	134.44444	Adj R-Sq	0.9995
Coeff Var	1.72224		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-73.46051	6.67743	-11.00	<.0001
inc	1	0.79596	0.26607	2.99	0.0097
risk	1	5.40039	0.25399	21.26	<.0001
inc2	1	0.05087	0.00244	20.85	<.0001

Do it in SAS


insu = -73.461 +0.796 inc +5.4004 risk +0.0509 inc2
Studentized Residual without Current Obs
Predicted Value
inc
risk
Normal Quantile
N
18
Rsq
0.9996
Adj Rsq
0.9995
RMSE
2.3155

Do it in SAS: output of /r

Output Statistics

```
     Dependent Predicted   Std Error          Std Error  Student                Cook's
Obs  Variable      Value Mean Predict  Residual  Residual Residual  -2-1 0 1 2        D

  1   91.0000   97.8164     0.7181  -6.8164     2.201  -3.097 |******|     |   0.255
  2  162.0000  160.1201     0.9577   1.8799     2.108   0.892 |      |*    |   0.041
  3   11.0000   11.5901     1.5574  -0.5901     1.713  -0.344 |      |     |   0.025
  4  240.0000  240.6278     0.8580  -0.6278     2.151  -0.292 |      |     |   0.003
  5   73.0000   71.5019     0.6656   1.4981     2.218   0.675 |      |*    |   0.010
  6  311.0000  309.6777     1.4363   1.3223     1.816   0.728 |      |*    |   0.083
  7  316.0000  315.6359     2.0100   0.3641     1.150   0.317 |      |     |   0.077
  8  154.0000  153.3645     0.9829   0.6355     2.096   0.303 |      |     |   0.005
  9  164.0000  162.4847     0.8211   1.5153     2.165   0.700 |      |*    |   0.018
 10   54.0000   52.4068     0.7346   1.5932     2.196   0.726 |      |*    |   0.015
 11   53.0000   52.8060     0.8340   0.1940     2.160  0.0898 |      |     |   0.000
 12  326.0000  327.6975     1.4378  -1.6975     1.815  -0.935 |     *|     |   0.137
 13   55.0000   54.4957     0.7142   0.5043     2.203   0.229 |      |     |   0.001
 14  130.0000  131.0179     1.2720  -1.0179     1.935  -0.526 |     *|     |   0.030
 15  112.0000  109.6080     0.8185   2.3920     2.166   1.104 |      |**   |   0.044
 16   91.0000   93.0992     0.8093  -2.0992     2.169  -0.968 |     *|     |   0.033
 17   14.0000   13.8135     1.2042   0.1865     1.978  0.0943 |      |     |   0.001
 18   63.0000   62.2363     0.6776   0.7637     2.214   0.345 |      |     |   0.003
```

Hat matrix diagonals

- h_{ii} is the leverage of the i^{th} observation
- $0 <= h_{ii} <= 1$; $Sum(h_{ii}) = p$
- The average value is p/n
- h_{ii} for new observation similarly defined
- We would like h_{ii} to be small; large value (>0.5 or 2p/n) indicates outlier/ extrapolation in X_i
- h_{ii} is also a measure of how much Y_i is contributing to the prediction Y_i(hat):

$Y_1(\text{hat}) = h_{11}Y_1 + h_{12} Y_2 + h_{13}Y_3 + \ \ldots$

Influential Cases: DFFITS

- A measure of the influence of case i on Y_i(hat)
- Standardized version of the difference between Y_i(hat) computed with and without case I
- Closely related to h_{ii}
- Large value (>1 or >2sqrt(p/n))indicate influential cases

Cook's Distance

- A measure of the influence of case i on all of the Y_i(hat)'s
- Standardized version of the sum of squares of the differences between the predicted values computed with and without case I
- The ith observation is influential if c_i >1 or >F (p, n-p, 0.5)

DFBETAS

- A measure of the influence of case i on each of the regression coefficients
- It is a standardized version of the difference between the regression coefficient computed with and without case i.
- Influential if >1 or >2/sqrt(n).

Do it in SAS

```
*more diagonostics;
proc reg data=l4;
  model insu = inc risk inc2/r influence;
  output out=lifeout cookd=ckd p=yhat
  rstudent=resid;
run;
* Index plot of cookD;
data lifeout;
  set lifeout;
  id = _n_;
run;
symbol1 v=circle i=join h = .8;

proc gplot data=lifeout;
  plot ckd*id ;
run;
```

Do it in SAS: output of /r

Output Statistics

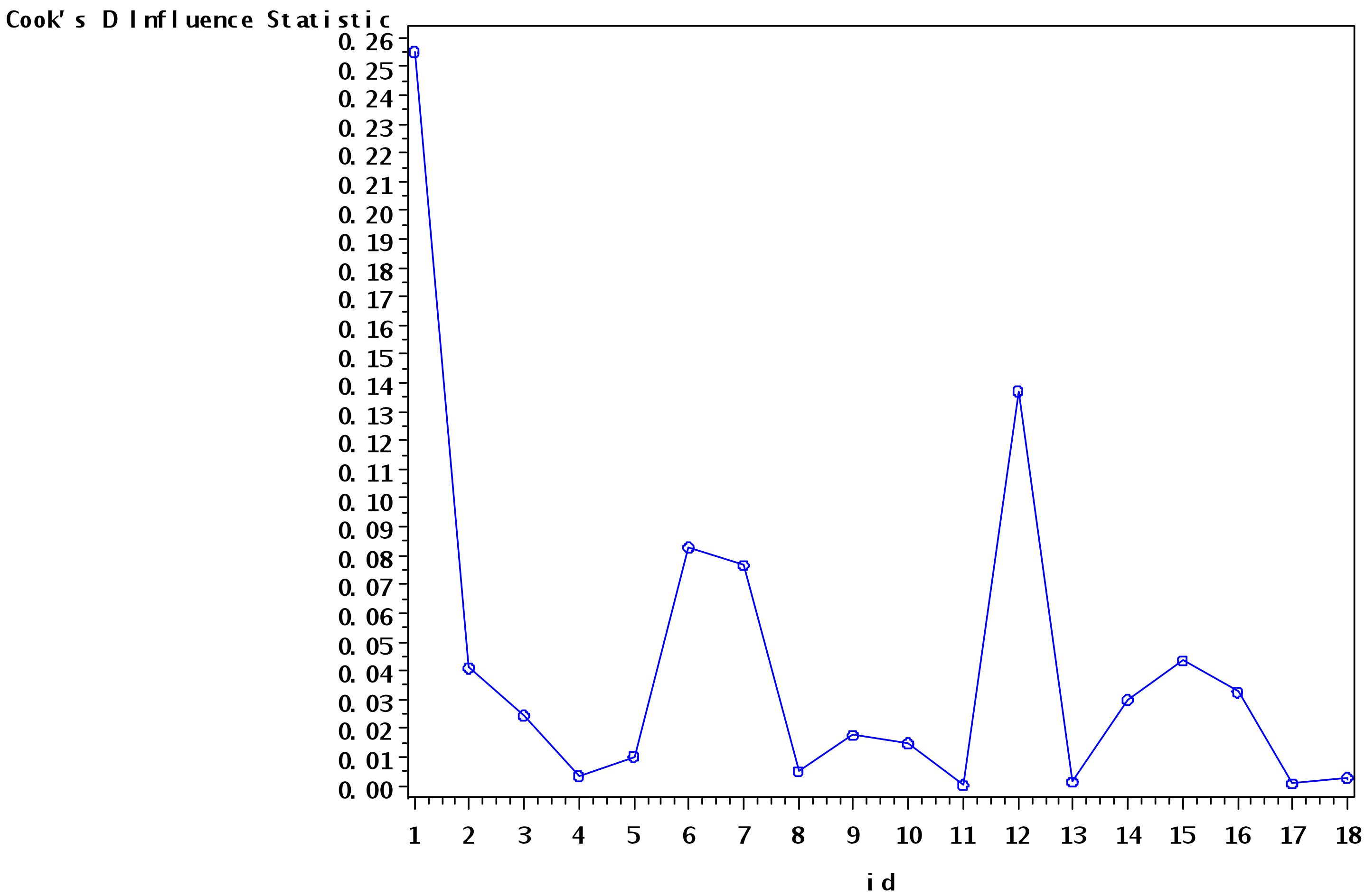
```
     Dependent Predicted    Std Error          Std Error  Student                  Cook's
 Obs  Variable     Value Mean Predict  Residual  Residual Residual   -2-1 0 1 2        D

   1   91.0000   97.8164     0.7181   -6.8164     2.201   -3.097 |******|     |    0.255
   2  162.0000  160.1201     0.9577    1.8799     2.108    0.892 |      |*    |    0.041
   3   11.0000   11.5901     1.5574   -0.5901     1.713   -0.344 |      |     |    0.025
   4  240.0000  240.6278     0.8580   -0.6278     2.151   -0.292 |      |     |    0.003
   5   73.0000   71.5019     0.6656    1.4981     2.218    0.675 |      |*    |    0.010
   6  311.0000  309.6777     1.4363    1.3223     1.816    0.728 |      |*    |    0.083
   7  316.0000  315.6359     2.0100    0.3641     1.150    0.317 |      |     |    0.077
   8  154.0000  153.3645     0.9829    0.6355     2.096    0.303 |      |     |    0.005
   9  164.0000  162.4847     0.8211    1.5153     2.165    0.700 |      |*    |    0.018
  10   54.0000   52.4068     0.7346    1.5932     2.196    0.726 |      |*    |    0.015
  11   53.0000   52.8060     0.8340    0.1940     2.160   0.0898 |      |     |    0.000
  12  326.0000  327.6975     1.4378   -1.6975     1.815   -0.935 |     *|     |    0.137
  13   55.0000   54.4957     0.7142    0.5043     2.203    0.229 |      |     |    0.001
  14  130.0000  131.0179     1.2720   -1.0179     1.935   -0.526 |     *|     |    0.030
  15  112.0000  109.6080     0.8185    2.3920     2.166    1.104 |      |**   |    0.044
  16   91.0000   93.0992     0.8093   -2.0992     2.169   -0.968 |     *|     |    0.033
  17   14.0000   13.8135     1.2042    0.1865     1.978   0.0943 |      |     |    0.001
  18   63.0000   62.2363     0.6776    0.7637     2.214    0.345 |      |     |    0.003
```

Do it in SAS: output of /influence

Output Statistics

Obs	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS inc	DFBETAS risk	DFBETAS inc2
1	-5.3155	0.0962	0.0147	-1.7339	0.7091	-0.8308	-0.3686	0.9168
2	0.8848	0.1711	1.2842	0.4020	-0.2325	0.2764	-0.2064	-0.2579
3	-0.3333	0.4524	2.3742	-0.3029	-0.2692	0.2518	-0.0525	-0.2312
4	-0.2822	0.1373	1.5215	-0.1126	0.0412	-0.0320	-0.0299	0.0230
5	0.6618	0.0826	1.2842	0.1986	-0.0149	0.0443	-0.0108	-0.0580
6	0.7153	0.3848	1.8735	0.5656	0.0420	-0.1377	0.3901	0.1704
7	0.3063	0.7535	5.3027	0.5356	0.1965	-0.1697	-0.3381	0.2233
8	0.2931	0.1802	1.5981	0.1374	-0.0768	0.0692	0.0788	-0.0712
9	0.6866	0.1258	1.3342	0.2604	-0.1791	0.1861	0.0084	-0.1799
10	0.7127	0.1006	1.2830	0.2384	0.0545	-0.0079	-0.0773	-0.0084
11	0.0866	0.1297	1.5420	0.0334	0.0145	-0.0122	0.0126	0.0091
12	-0.9308	0.3856	1.6912	-0.7373	-0.1800	0.2926	-0.3821	-0.3486
13	0.2210	0.0951	1.4643	0.0717	0.0225	-0.0125	0.0030	0.0063
14	-0.5120	0.3018	1.7786	-0.3366	0.1449	-0.1983	0.2583	0.1861
15	1.1138	0.1249	1.0675	0.4209	-0.1813	0.1838	0.2003	-0.2036
16	-0.9653	0.1222	1.1616	-0.3601	0.1516	-0.2120	0.1654	0.2177
17	0.0909	0.2705	1.8390	0.0553	0.0435	-0.0351	-0.0150	0.0317
18	0.3338	0.0856	1.4216	0.1022	0.0135	0.0015	-0.0003	-0.0097

Do it in SAS



Cook's D Influence Statistic
0.26
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
id

Multicollinearity Diagnostics

- Large correlation between explanatory variables
- t-test not significant for important explanatory variables
- Large change in estimated regression coefficients when add/remove var.
- The sign of estimated regression coefficient different from what we expected

Variance Inflation Factor

- $VIF = 1/(1 – R^2_k)$
- R^2_k is the squared multiple correlation obtained in a regression where all other explanatory variables are used to predict X_k
- One suggested rule: a value of 10 or more indicates excessive multicollinearity
- Tolerance: $TOL=1/VIF=(1 – R^2_k)$

Body fat example revisit

```
* check collinearity using VIF/TOL;
proc reg data = fat;
   model fat = skinfold thigh midarm /
  VIF TOL;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	117.08469	99.78240	1.17	0.2578	.	0
skinfold	1	4.33409	3.01551	1.44	0.1699	0.00141	708.84291
thigh	1	-2.85685	2.58202	-1.11	0.2849	0.00177	564.34339
midarm	1	-2.18606	1.59550	-1.37	0.1896	0.00956	104.60601

Regression Diagnostics Recommendations

- Plot the residuals versus fitted value, versus each of the X's, interactions, other variables (time, etc.)
- Examine the added variable plots
- Check normality of the residuals with a normal quantile plot

Regression Diagnostics Recommendations

- Examine
  - the studentized deleted residuals (RSTUDENT in the output)
  - The hat matrix diagonals
  - DFFITS, Cook's Distance, and the DFBETAS
- Check observations that are extreme on these measures relative to the other observations

Regression Diagnostics Recommendations

- Examine the tolerance/VIF for each X
- If there are variables with low tolerance, you need to do some model building
  - – Recode variables
  - – Variable selection

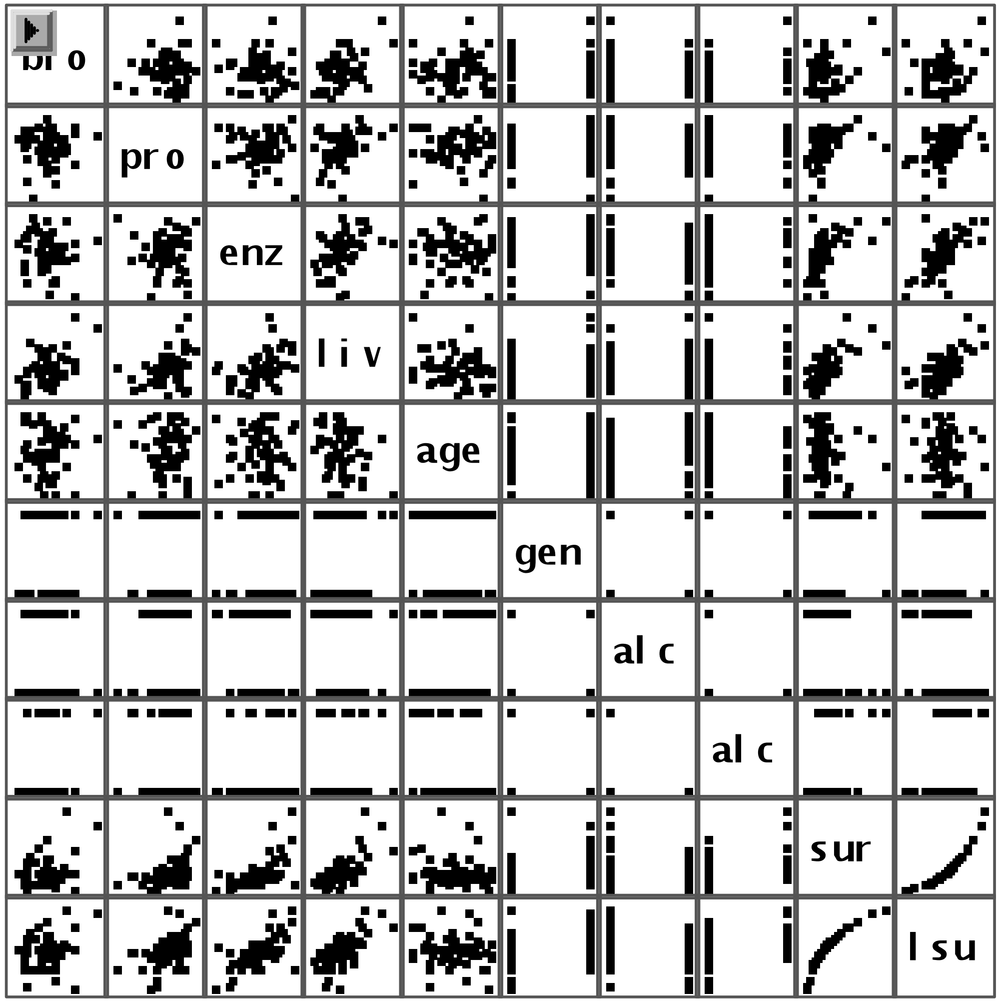
Surgical Unit Example

- Predicting survival after liver operation
- Y is survival time
- X's are
  - Blood clotting score
  - Prognostic index
  - Enzyme function test
  - Liver function test
  - Age
  - Gender
  - Alcohol use (three level)

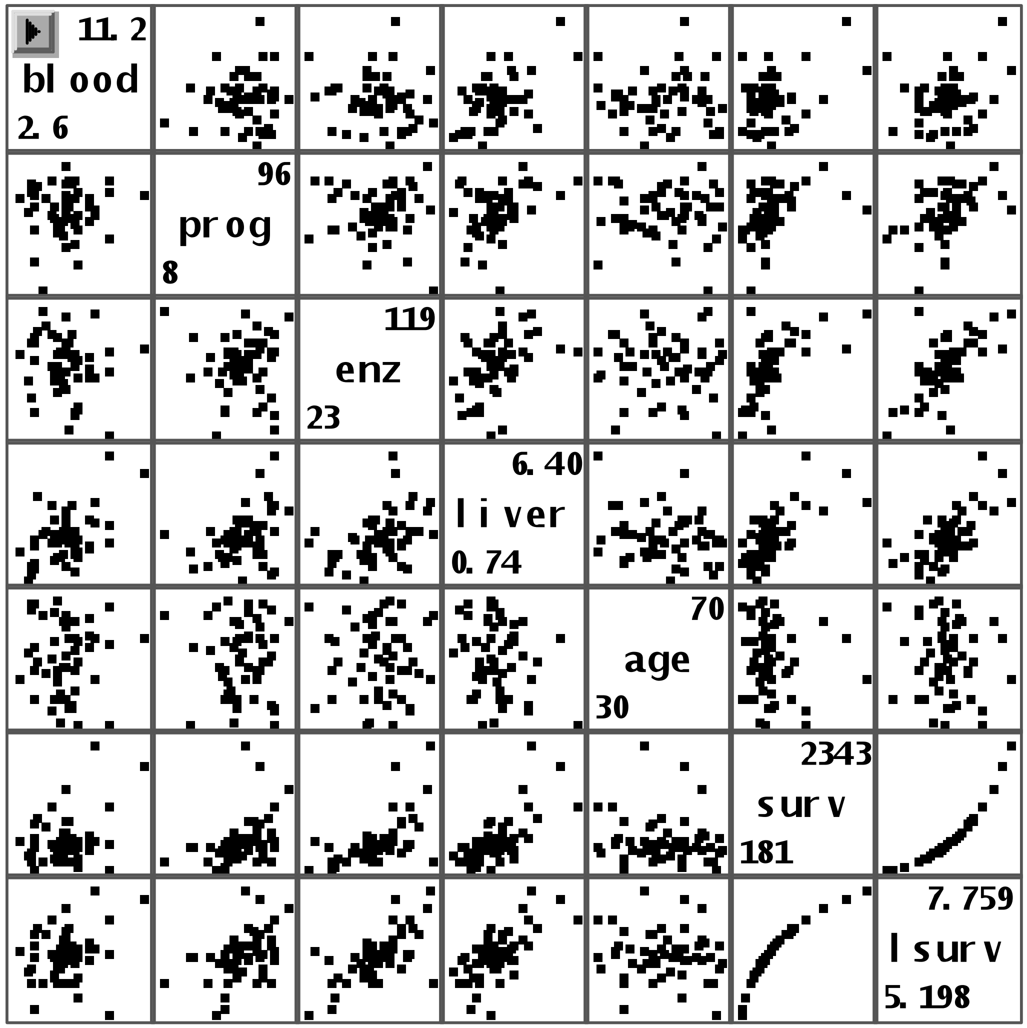
Do it in SAS

```
Data surg;
infile 'C:\...\surgical.txt' dlm='09'x;
input blood prog enz liver age gend alc1
  alc2 surv lsurv;
Proc print data=surg;
run;

%include "C:\...\scatter.sas";
%scatter(data = surg);
%scatter(data = surg, var = blood prog enz
  liver age surv lsurv);
```

Do it in SAS



Do it in SAS


11.2
blood
2.6
96
prog
8
119
enz
23
6.40
liver
0.74
70
age
30
2343
surv
181
7.759
lsurv
5.198

Surgical Unit Example (2)

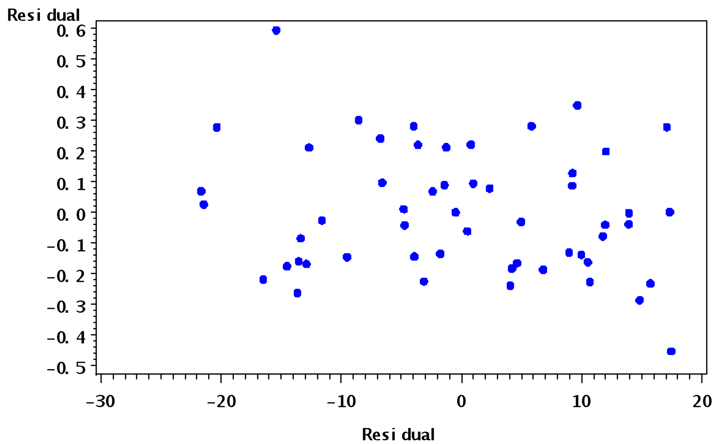
- n = 54 patients
- Diagnostics suggest that Y should be transformed with a log
- Focus on model using variables blood, prog, enz, and alc2 (reasons shown later)

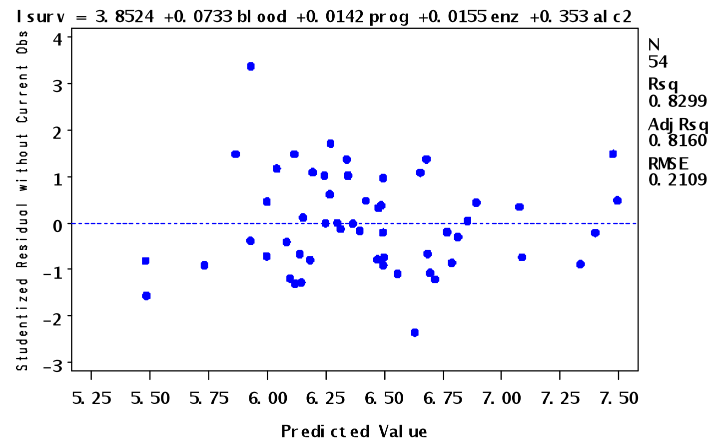
Do it in SAS

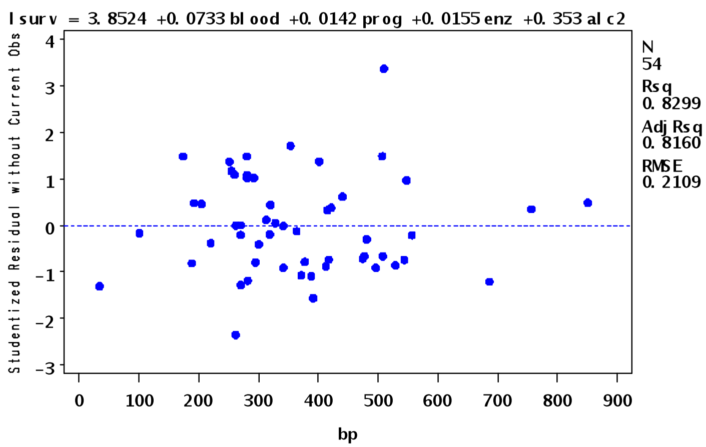
```
* Surgical unit example
   revisit;
Data surg;
infile 'C:\...\surgical.txt'
   dlm='09'x;
input blood prog enz liver age
   gend alc1 alc2 surv lsurv;

*added variable plot;
proc reg data = surg;
  model lsurv age = blood prog
   enz alc2;
  output out=s2 r=rsurv rage;
run;

symbol1 v=dot h=.8 c=blue;
proc gplot data=s2;
   plot rsurv*rage;
run;
```

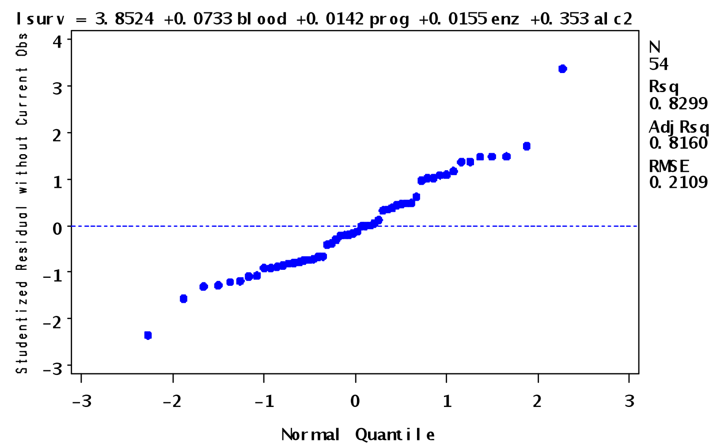
```
*nolinearity and interaction;
Data s2;
   set surg;
   b2=blood*blood;
   bp=blood*prog;
run;

proc reg data = s2;
  var b2 bp;
  model lsurv = blood prog enz
   alc2/r influence vif;
  plot rstudent.*(p. b2 bp
   nqq.);
run;
```

Do it in SAS


Residual
Residual
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-30
-20
-10
0
10
20


lsurv = 3.8524 +0.0733 blood +0.0142 prog +0.0155 enz +0.353 alc2
Studentized Residual without Current Obs
Predicted Value
5.25
5.50
5.75
6.00
6.25
6.50
6.75
7.00
7.25
7.50
N
54
Rsq
0.8299
Adj Rsq
0.8160
RMSE
0.2109


lsurv = 3.8524 +0.0733 blood +0.0142 prog +0.0155 enz +0.353 alc2
Studentized Residual without Current Obs
bp
0
100
200
300
400
500
600
700
800
900
N
54
Rsq
0.8299
Adj Rsq
0.8160
RMSE
0.2109


lsurv = 3.8524 +0.0733 blood +0.0142 prog +0.0155 enz +0.353 alc2
Studentized Residual without Current Obs
Normal Quantile
N
54
Rsq
0.8299
Adj Rsq
0.8160
RMSE
0.2109

Do it in SAS

```
*outlier and influential cases;
ods listing close;
proc reg data = s2;
  model lsurv = blood prog enz alc2/
   r influence;
  ods output  OutputStatistics=temp;
  output out=temp1 cookd = cooksd;
run;

ods listing;
data temp2;
  set temp;
  keep observation residual
   hatdiagonal rstudent dffits;
run;
data temp1;
  set temp1;
  observation = _n_;
  keep observation cooksd;
run;
```

```
data combined ;
  merge temp1 temp2;
  by observation;
run;
proc print data = combined;
 where observation=17 or
   observation=23 or observation=28
   or
       observation=32 or
   observation=38 or observation=42
   or observation=52;
 var residual hatdiagonal rstudent
   dffits cooksd;
run;
```

Do it in SAS

The SAS System   01:35 Monday, November 28, 2005 15

Obs	Residual	Hat Diagonal	RStudent	DFFITS	cooksd
17	0.5952	0.1499	3.3696	1.4151	0.33062
23	0.2788	0.1885	1.4854	0.7160	0.10006
28	0.0876	0.2914	0.4896	0.3140	0.02002
32	-0.2861	0.2202	-1.5585	-0.8283	0.13333
38	-0.2271	0.3059	-1.3016	-0.8641	0.14725
42	-0.0303	0.2262	-0.1620	-0.0876	0.00157
52	-0.1375	0.2221	-0.7358	-0.3931	0.03120