STOR 455
STATISTICAL METHODS I

Jan Hannig

Regression Diagnostics (Chapter 5)

- Added variable plot
- Studentized deleted residuals (Y-outlier)
- Hat matrix leverage values (X-outlier)
- DFFITS, Cook's D, DFBETAS (Influential cases)

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
- The /partial option generates graphs in the output window

Identifying Outliers

- Residuals $e_i = Y_i - Y_i$ (hat)
- semistudentized residuals ei/sqrt(MSE)
- Studentized residuals
  $r_i = e_i / \text{sqrt}(MSE(1 - h_{ii}))$
- In proc reg use the handle student

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 $t_i = d_i / \text{sqrt}(MSE_{(i)} (1 - h_{ii}))$
- In proc reg use the handle rstudent

Use of Residuals

- We are looking for
  - Outliers (Bonferroni t-test)
  - Constant variance
  - Uncorrelated error
  - Normal error distributions
- /r in the model statement requests residual analysis

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} is also a measure of how much Y_i is contributing to the prediction $Y_i(hat)$:
  $Y_1(hat) = h_{11}Y_1 + h_{12}Y_2 + h_{13}Y_3 + \ldots$
- We would like h_{ii} to be small;
  - large value (>0.5 or 2p/n) indicates outlier/ extrapolation in X_i

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

$$\mathrm{DFFITS}_i = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

- 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

$$c_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}}r_i^2$$

- The ith observation is influential if c_i >1 or c_i>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).

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, also /r in the model statement)
  - The hat matrix diagonals
  - DFFITS, Cook's D, and the DFBETAS (/influence in the model statement)
- Check observations that are extreme on these measures relative to the other observations

Regression Diagnostics Recommendations

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

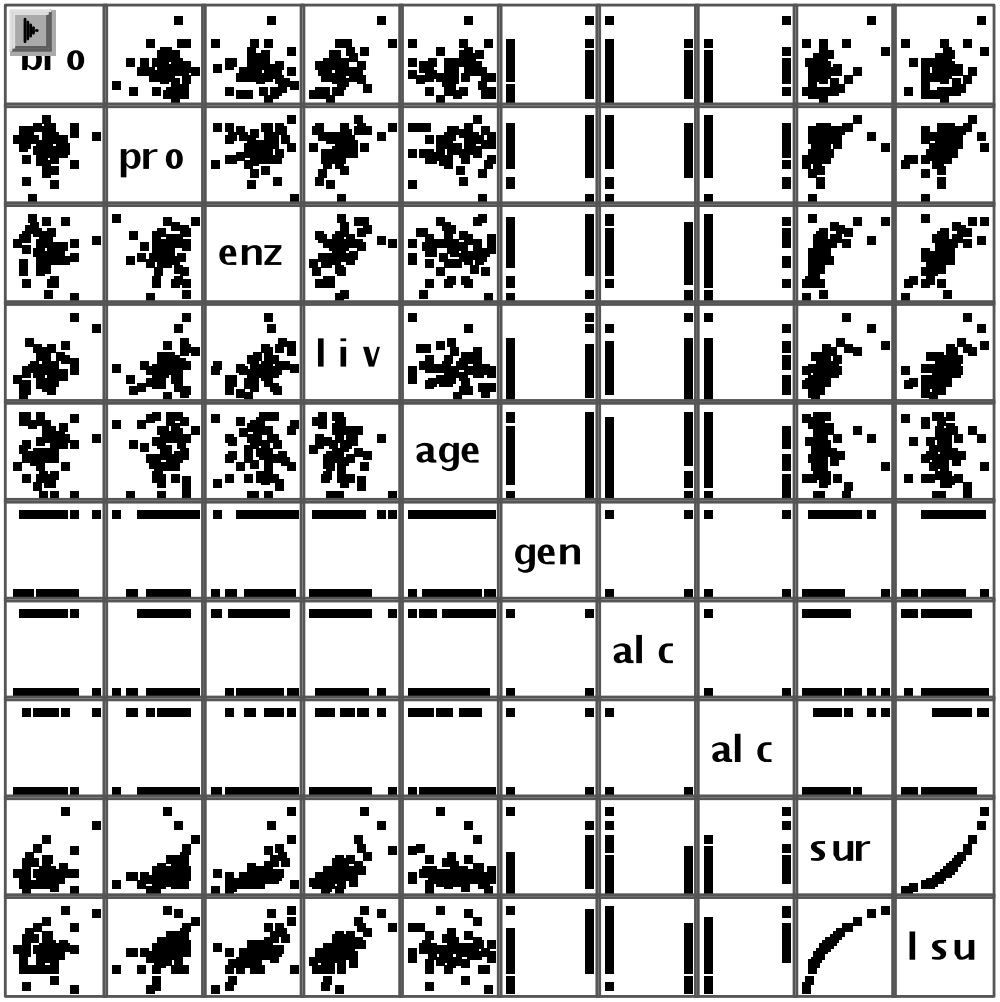
Surgical Unit Example

- 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)

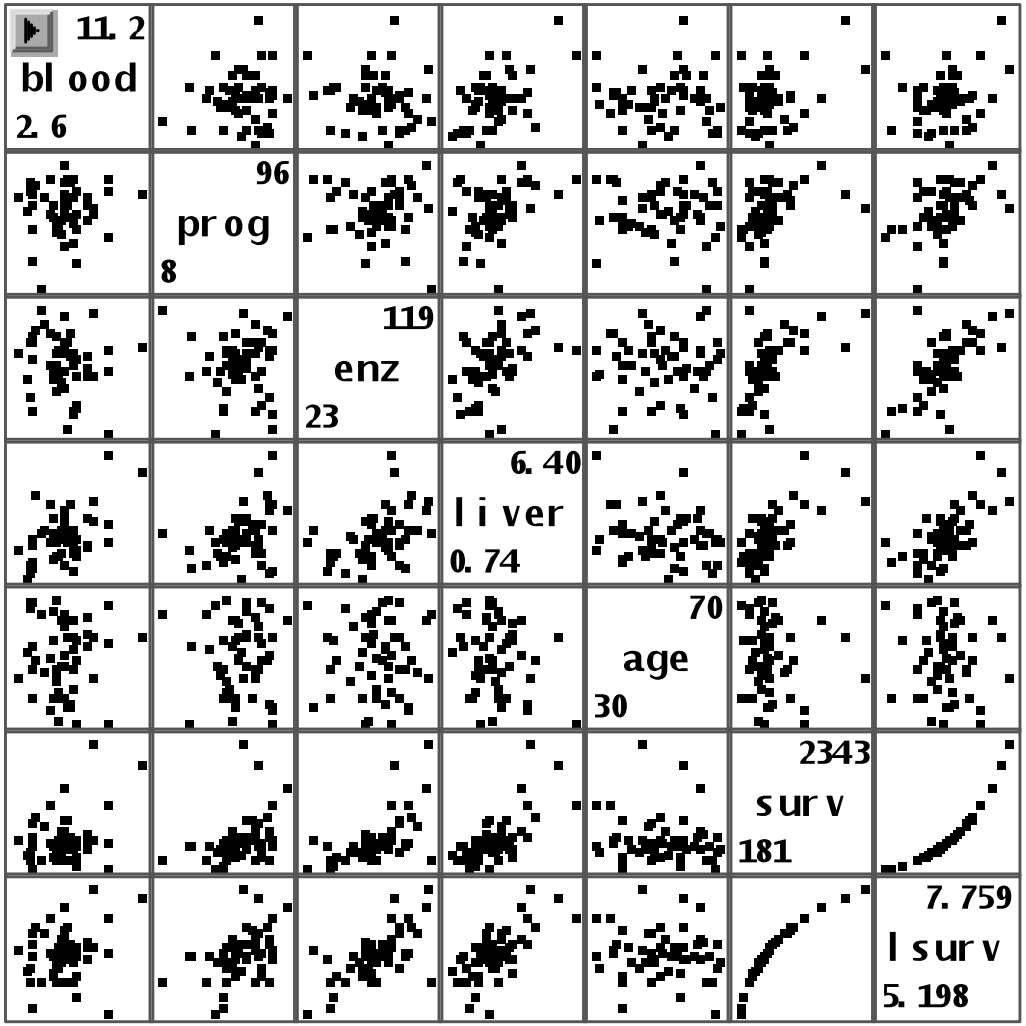
Do it in SAS

```
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)

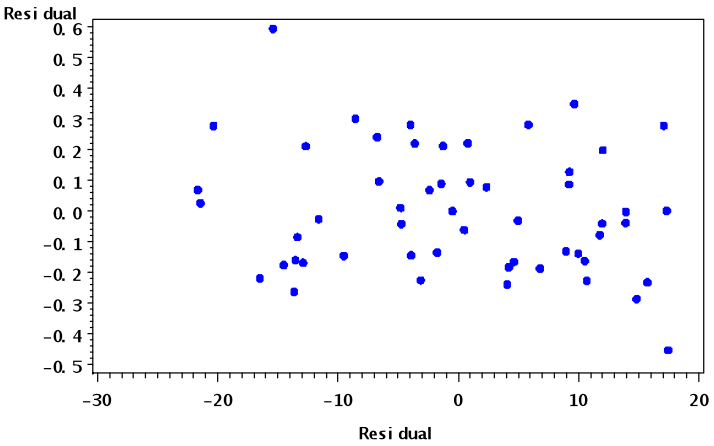
- n = 54 patients
- Diagnostics suggest that Y should be transformed with a log
- Focus on model using variables blood, prog, enz and alc2 (reasons shown later)

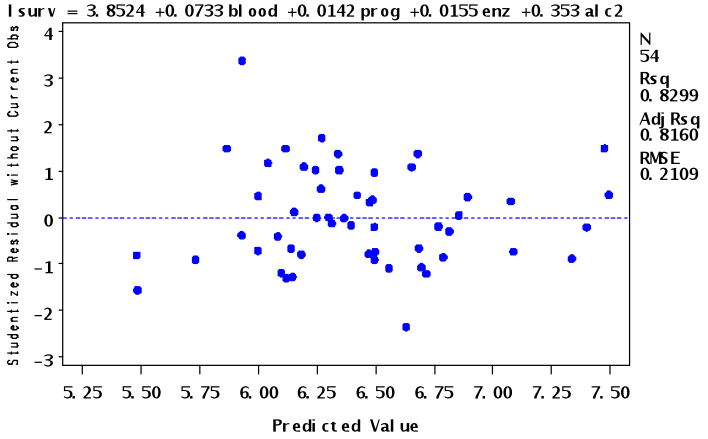
Do it in SAS

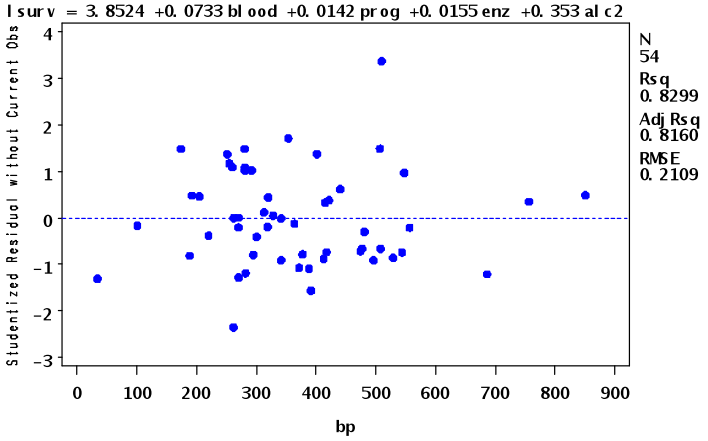
```
* Surgical unit example
   revisit;
Data surg;
infile 'C:\...\surgical.txt'
   dlm='09'x;
input blood prog enz liver age
   gend alc1 alc2 surv lsurv;

*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;
```

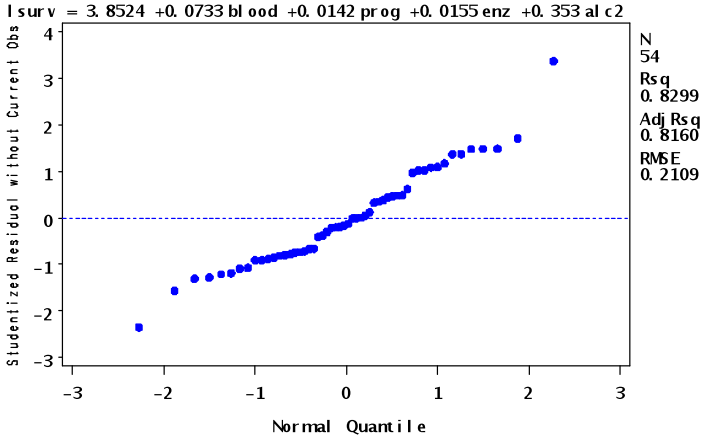
```
*nolinearity and interaction;
Data s2;
   set surg;
   b2=blood*blood;
   bp=blood*prog;
run;

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
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
Residual


lsurv = 3.8524 +0.0733 blood +0.0142 prog +0.0155 enz +0.353 alc2
Studentized Residual without Current Obs
4
3
2
1
0
-1
-2
-3
5.25
5.50
5.75
6.00
6.25
6.50
6.75
7.00
7.25
7.50
Predicted Value
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
4
3
2
1
0
-1
-2
-3
0
100
200
300
400
500
600
700
800
900
bp
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
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
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

Model Selection (Chapter 7)

- The original data set had 7 predictors
  (Blood clotting score, Prognostic index, Enzyme function test, Liver function test, Age, Gender, Alcohol use - three level)
- Why did we choose to use to use blood, prog, enz, and alc2?
- Model selection methods
  - All subset selection
  - Forward selection, backward elimination, stepwise regression

All-possible Procedure

- Consider all subset of the full model
- Select submodel based on certain criterion
- May have several “good models” instead of one “best model”
- Number of models to consider: 2^p
- Not possible when p is large

Criteria for Variable Selection

- Variance based: minimize root-MSE
- SSE based criterion: maximize R^2 =1-SSE(model)/SSTO
- MSE based criterion: maximize adjusted-R^2=1-MSE(model)/MSTO

  accounting for d.f. of the model
- C_p criterion: minimizing the bias of sub-model
- PRESS: minimizing prediction error

Mallow's C_p

- Compare subset models with the full model
  - A subset model is good if there is not substantial bias in the predicted values (relative to the full model)
- C_p is a measure of this bias

$$C_p = SSE_p / MSE_{Full} - (n-2p)$$

- See Section 7.3 for details

Use of C_p

- p is the number of regression coefficients including the intercept
- A model is good according to this criterion if C_p is approximately equal to p
- C_p < p is ok too

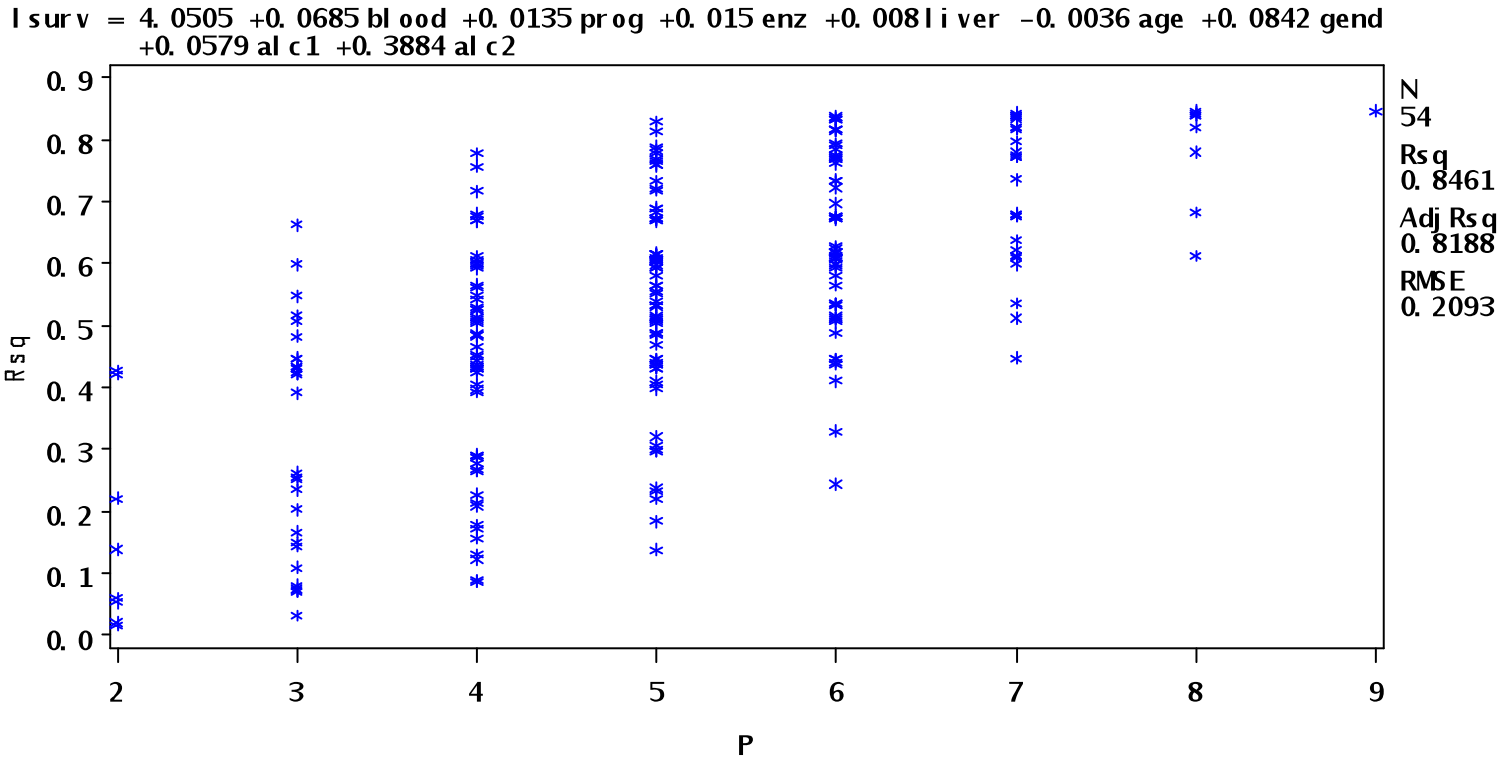
PRESS

- Prediction error sums of squares
- For each case i, delete the case and predict Y using a model based on the other n-1 cases
- PRESS= SS for observed minus predicted
- Small PRESS indicate better fit

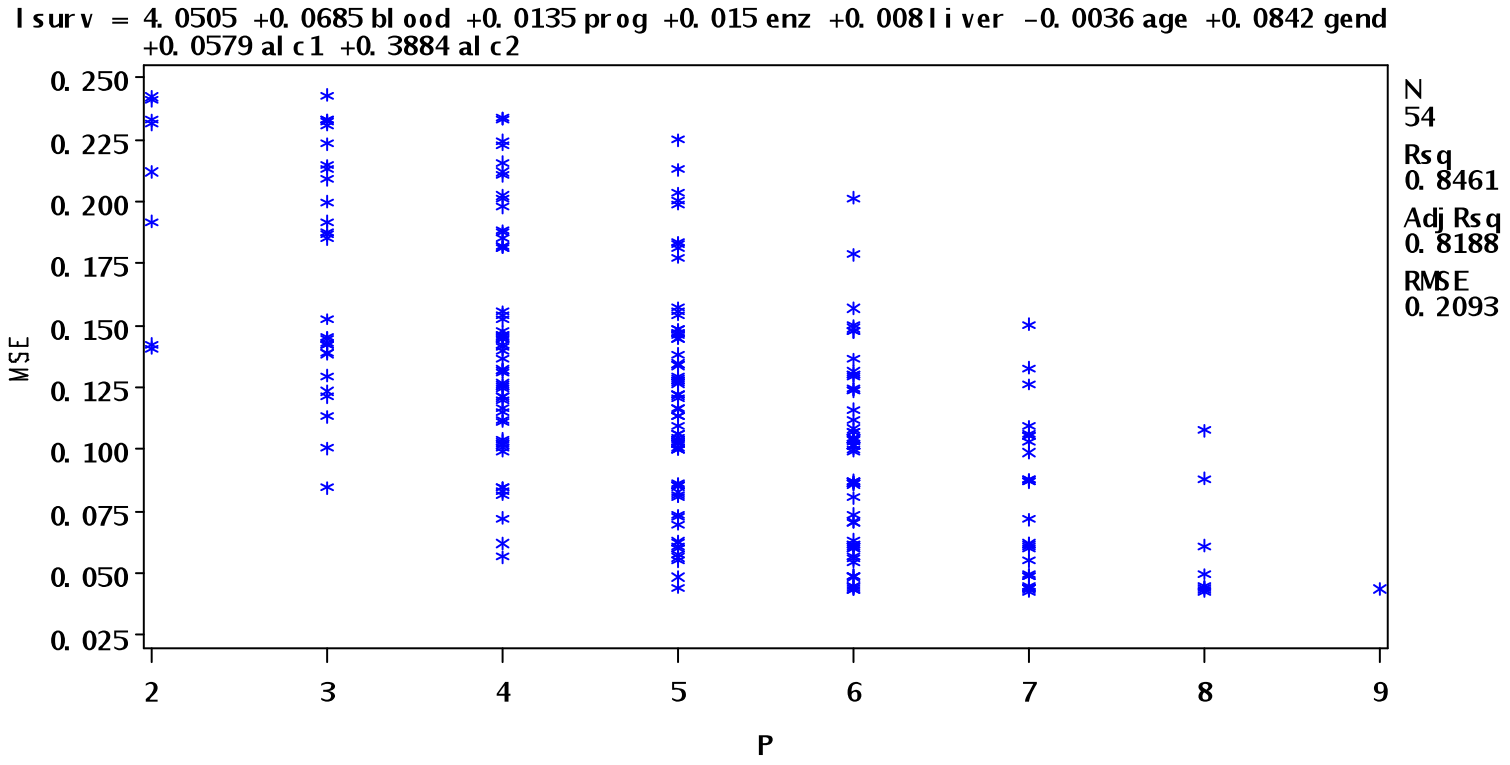
Do it in SAS

```
* All subset procedure, R^2, adj. R^2 and cp
  as criteria;
* Plot may not be informative;
goptions reset=all;
symbol1 v=star c=blue h = .8;
proc reg data = surg;
   model lsurv = blood prog enz liver age
   gend alc1 alc2/selection= rsquare adjrsq
   cp mse;
   plot rsq.*np.;
   plot mse.*np.;
   plot cp.*np. ;
run;
```

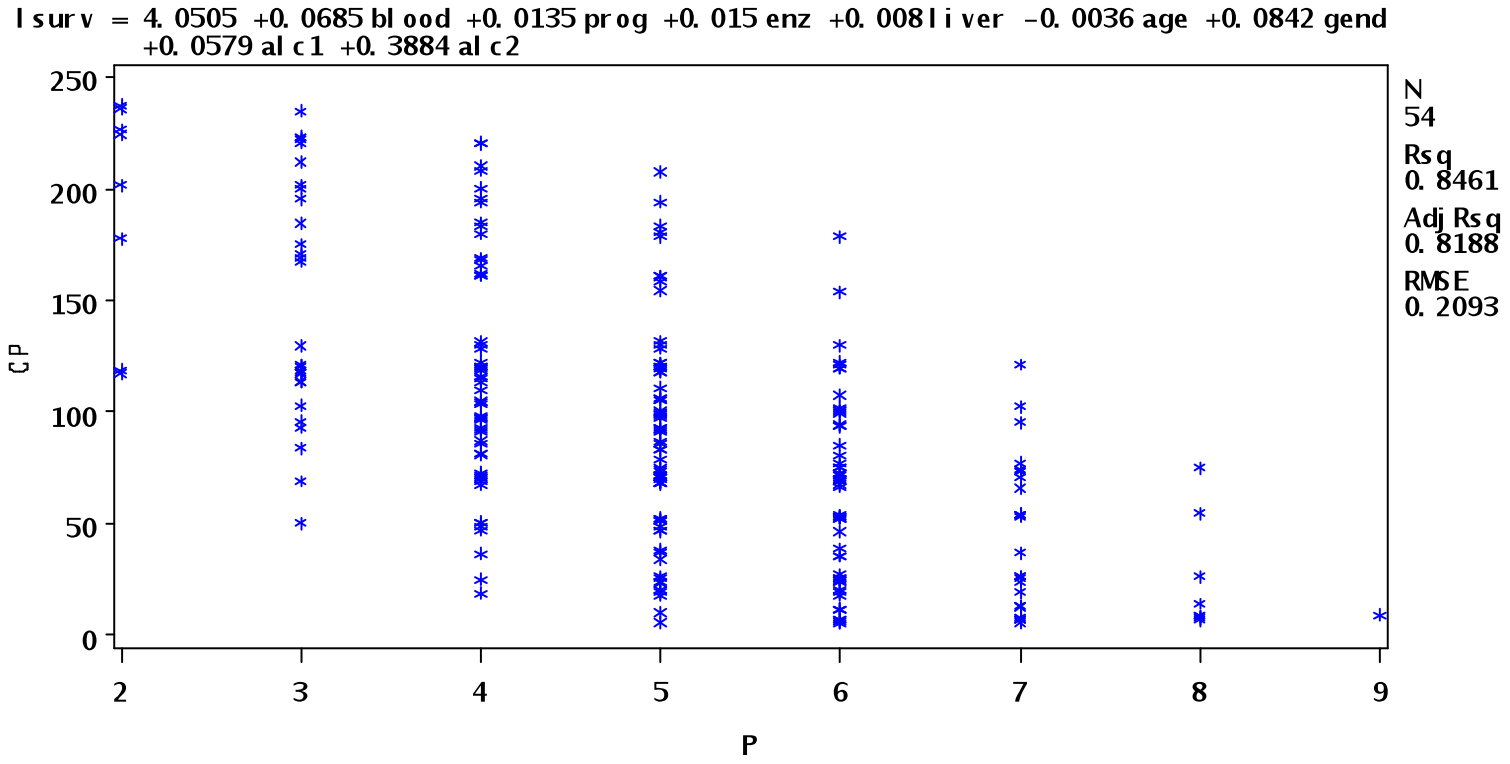
Do it in SAS


lsurv = 4.0505 +0.0685 blood +0.0135 prog +0.015 enz +0.008 liver -0.0036 age +0.0842 gend
+0.0579 alc1 +0.3884 alc2
Rsq
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2
3
4
5
6
7
8
9
P
N
54
Rsq
0.8461
Adj Rsq
0.8188
RMSE
0.2093

Do it in SAS


lsurv = 4.0505 +0.0685 blood +0.0135 prog +0.015 enz +0.008 liver -0.0036 age +0.0842 gend
+0.0579 alc1 +0.3884 alc2
MSE
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
2
3
4
5
6
7
8
9
P
N
54
Rsq
0.8461
Adj Rsq
0.8188
RMSE
0.2093

Do it in SAS


lsurv = 4.0505 +0.0685 blood +0.0135 prog +0.015 enz +0.008 liver -0.0036 age +0.0842 gend
+0.0579 alc1 +0.3884 alc2
250
200
150
100
50
0
CP
2
3
4
5
6
7
8
9
P
N
54
Rsq
0.8461
Adj Rsq
0.8188
RMSE
0.2093

Do it in SAS

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
3	0.7780	0.7647	18.9145	0.05686	prog enz alc2
3	0.7573	0.7427	24.9805	0.06217	blood prog enz
3	0.7178	0.7009	36.5252	0.07228	prog enz liver
3	0.6810	0.6618	47.3038	0.08172	prog enz alc1
4	0.8299	0.8160	5.7508	0.04447	blood prog enz alc2
4	0.8144	0.7993	10.2670	0.04850	prog enz liver alc2
4	0.7888	0.7715	17.7770	0.05521	prog enz gend alc2
4	0.7836	0.7659	19.2976	0.05657	prog enz age alc2
5	0.8374	0.8205	5.5406	0.04338	blood prog enz gend alc2
5	0.8358	0.8187	6.0182	0.04381	blood prog enz age alc2
5	0.8331	0.8158	6.7986	0.04452	blood prog enz liver alc2
5	0.8317	0.8141	7.2269	0.04491	blood prog enz alc1 alc2
5	0.8179	0.7989	11.2608	0.04859	prog enz liver gend alc2
5	0.8159	0.7968	11.8266	0.04911	prog enz liver alc1 alc2
5	0.8157	0.7965	11.9099	0.04919	prog enz liver age alc2
5	0.7944	0.7730	18.1324	0.05486	prog enz age gend alc2

Do it in SAS

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
6	0.8434	0.8234	5.7874	0.04266	blood prog enz age gend alc2
6	0.8392	0.8187	7.0295	0.04382	blood prog enz gend alc1 alc2
6	0.8387	0.8181	7.1662	0.04395	blood prog enz liver gend alc2
6	0.8384	0.8178	7.2462	0.04402	blood prog enz age alc1 alc2
6	0.8371	0.8164	7.6270	0.04438	blood prog enz liver age alc2
6	0.8348	0.8137	8.3146	0.04502	blood prog enz liver alc1 alc2
7	0.8460	0.8226	7.0295	0.04287	blood prog enz age gend alc1 alc2
7	0.8436	0.8198	7.7352	0.04354	blood prog enz liver age gend alc2
7	0.8404	0.8161	8.6793	0.04444	blood prog enz liver gend alc1 alc2
7	0.8396	0.8151	8.9214	0.04467	blood prog enz liver age alc1 alc2
7	0.8213	0.7941	14.2632	0.04976	prog enz liver age gend alc1 alc2
7	0.7801	0.7466	26.3115	0.06123	blood prog enz liver age gend alc1
7	0.6829	0.6347	54.7334	0.08829	blood enz liver age gend alc1 alc2
7	0.6126	0.5537	75.2932	0.10786	blood prog liver age gend alc1 alc2
8	0.8461	0.8188	9.0000	0.04379	blood prog enz liver age gend alc1 alc2

Best Models Using of C_p

- blood prog enz gend alc2, 5.54
- blood prog enz alc2, 5.75
- blood prog enz age gend alc2, 5.79
- blood prog enz age gend alc1 alc2, 7.03
- blood prog enz liver age gend alc2, 7.74

SAS MACRO allsubsreg

```
* Use macro to compute PRESS;
%include "T:\...\allsubsreg.sas";
%allsubsreg(data=surg, depvar=lsurv,
  indepvar=blood prog enz liver age
  gend alc1 alc2, sortvar=_PRESS_,
  printvar=_RMSE_ _RSQ_ _CP_ _PRESS_);
run;
```

Do it in SAS

---------- Number of regressors in model=4 ----------
(continued)

Obs	VarsInModel	_IN_	_RMSE_	_RSQ_	_CP_	_PRESS_	Intercept	blood
156	blood prog enz alc1	4	0.24705	0.76650	24.2885	3.88900	3.83207	0.09176
157	blood prog enz age	4	0.24579	0.76888	23.5924	3.86305	4.02810	0.09483
158	prog enz age alc2	4	0.23785	0.78356	19.2976	3.53115	4.47171	.
159	prog enz alc1 alc2	4	0.23982	0.77996	20.3519	3.52287	4.26344	.
160	prog enz gend alc2	4	0.23498	0.78876	17.7770	3.50513	4.29334	.
161	prog enz liver alc2	4	0.22023	0.81444	10.2670	3.02103	4.34067	.
162	blood prog enz alc2	4	0.21087	0.82988	5.7508	2.73777	3.85242	0.07332

---------- Number of regressors in model=5 ----------
(continued)

Obs	VarsInModel	_IN_	_RMSE_	_RSQ_	_CP_	_PRESS_	Intercept
209	prog enz age alc1 alc2	5	0.23877	0.78633	20.4874	3.56560	4.45344
210	prog enz age gend alc2	5	0.23423	0.79439	18.1324	3.55562	4.47592
211	prog enz gend alc1 alc2	5	0.23636	0.79063	19.2313	3.55539	4.26651
212	prog enz liver age alc2	5	0.22178	0.81566	11.9099	3.13206	4.42572
213	prog enz liver gend alc2	5	0.22044	0.81788	11.2608	3.12603	4.33834
214	prog enz liver alc1 alc2	5	0.22161	0.81595	11.8266	3.10027	4.31629
215	blood prog enz liver alc2	5	0.21100	0.83314	6.7986	2.82935	3.96517
216	blood prog enz alc1 alc2	5	0.21193	0.83168	7.2269	2.79565	3.82669
217	blood prog enz gend alc2	5	0.20827	0.83744	5.5406	2.78271	3.86710
218	blood prog enz age alc2	5	0.20931	0.83581	6.0182	2.73893	4.03812

Do it in SAS

---------- Number of regressors in model=6 ----------
(continued)

Obs	VarsInModel	_IN_	_RMSE_	_RSQ_	_CP_	_PRESS_	Intercept
239	prog enz liver gend alc1 alc2	6	0.22185	0.81939	12.8203	3.20786	4.31396
240	prog enz liver age alc1 alc2	6	0.22298	0.81754	13.3595	3.20039	4.41122
241	blood prog enz liver alc1 alc2	6	0.21218	0.83479	8.3146	2.89906	3.93809
242	blood prog enz liver gend alc2	6	0.20964	0.83872	7.1662	2.87494	3.93868
243	blood prog enz liver age alc2	6	0.21066	0.83715	7.6270	2.85668	4.08721
244	blood prog enz gend alc1 alc2	6	0.20934	0.83919	7.0295	2.83917	3.84163
245	blood prog enz age alc1 alc2	6	0.20982	0.83845	7.2462	2.77826	4.02082
246	blood prog enz age gend alc2	6	0.20655	0.84344	5.7874	2.77233	4.05397

---------- Number of regressors in model=7 ----------

Obs	VarsInModel	_IN_	_RMSE_	_RSQ_	_CP_	_PRESS_	Intercept
250	prog enz liver age gend alc1 alc2	7	0.22306	0.82129	14.2632	3.30254	4.41771
251	blood prog enz liver gend alc1 alc2	7	0.21081	0.84039	8.6793	2.94350	3.91149
252	blood prog enz liver age alc1 alc2	7	0.21136	0.83956	8.9214	2.90722	4.06637
253	blood prog enz liver age gend alc2	7	0.20867	0.84361	7.7352	2.88266	4.07191
254	blood prog enz age gend alc1 alc2	7	0.20705	0.84603	7.0295	2.80871	4.03678

---------- Number of regressors in model=8 ----------

Obs	VarsInModel	_IN_	_RMSE_	_RSQ_	_CP_	_PRESS_	Intercept
255	blood prog enz liver age gend alc1 alc2	8	0.20927	0.84613	9	2.93123	4.05052

Best Models Using PRESS

- blood prog enz alc2, 2.73777
- blood prog enz age alc2, 2.73893
- blood prog enz age alc1 alc2, 2.77826
- blood prog enz age gend alc2, 2.77233
- blood prog enz gend alc2 2.78271
- blood prog enz alc1 alc2, 2.79565

Automatic search procedures

- When p large, can't do all subset
- Stepwise type procedures
  - Forward selection (Step up)
  - Backward elimination (Step down)
  - Stepwise: combines F and B, allow removal of variables after adding new variables.
- Many other alternatives, but NONE guarantees the optimal solution (NP-hard problem)

Forward Selection

- Start with an intercept
- At each step add the "best" variable (using some criteria – R^2, adj R^2, C_p, partial correlation, SSE, …)
- Compare the p-value for the test whether the just added variable is 0 with some pre-selected value.
  - If larger– add the variable and repeat the procedure
  - If smaller – stop. You have arrived at the final model.
- This is purely exploratory – see the book for comments

Backwards Elimination

- Start with the full model
- At each step delete the "worst" variable (using some criteria – smallest increase in SSE, smallest p-value,…)
- Compare the p-value for the test whether the just deleted variable is 0 with some pre-selected value.
  - If smaller – delete the variable and repeat the procedure
  - If larger – stop. You have arrived at the final model.

Stepwise Regression

- Combines the forward and backward algorithm
- Starts at some model (empty or full is usual)
- First eliminates as many parameters as possible using backwards rule (using some alpha-delete)
- Than attempt to add one variable (using some alpha-add)
- If a variable is added repeat, if not stop (alpha-add>alpha-delete required for convergence).

Do it in SAS

```
* Forward selection;
proc reg data=surg;
   model lsurv=blood prog enz liver
    age gend alc1 alc2 /
    selection=FORWARD slentry=0.5;
run;

* backward elimination;
proc reg data=surg;
   model lsurv=blood prog enz liver
    age gend alc1 alc2 /selection=B
    slstay=0.05;
run;

* Forward stepwise;
proc reg data=surg;
   model lsurv=blood prog enz liver
    age gend alc1 alc2 /
    selection=STEPWISE slentry=0.50
    slstay=0.05;
run;
```

```
* Forward stepwise with first
   variable always in the model;
proc reg data=surg;
   model lsurv=blood prog enz liver
    age gend alc1 alc2 /
    selection=STEPWISE include=1;
run;

* Stepwise with at least two
   variables in the model;
proc reg data=surg;
   model lsurv=blood prog enz liver
    age gend alc1 alc2 /
    selection=STEPWISE start=2;
run;
```

Forward Selection and Backward Elimination

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	enz	1	0.4276	0.4276	117.409	38.84	<.0001
2	prog	2	0.2357	0.6633	50.4716	35.70	<.0001
3	alc2	3	0.1147	0.7780	18.9145	25.85	<.0001
4	blood	4	0.0519	0.8299	5.7508	14.93	0.0003
5	gend	5	0.0076	0.8374	5.5406	2.23	0.1418
6	age	6	0.0060	0.8434	5.7874	1.80	0.1862
7	alc1	7	0.0026	0.8460	7.0295	0.77	0.3835

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	liver	7	0.0001	0.8460	7.0295	0.03	0.8645
2	alc1	6	0.0026	0.8434	5.7874	0.77	0.3835
3	age	5	0.0060	0.8374	5.5406	1.80	0.1862
4	gend	4	0.0076	0.8299	5.7508	2.23	0.1418

Stepwise

All variables left in the model are significant at the 0.0500 level.

The stepwise method terminated because the next variable to be entered was just removed.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	enz		1	0.4276	0.4276	117.409	38.84	<.0001
2	prog		2	0.2357	0.6633	50.4716	35.70	<.0001
3	alc2		3	0.1147	0.7780	18.9145	25.85	<.0001
4	blood		4	0.0519	0.8299	5.7508	14.93	0.0003
5	gend		5	0.0076	0.8374	5.5406	2.23	0.1418
6		gend	4	0.0076	0.8299	5.7508	2.23	0.1418

All variables left in the model are required or significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	enz		2	0.4424	0.6633	50.4716	67.01	<.0001
2	alc2		3	0.1147	0.7780	18.9145	25.85	<.0001
3	blood		4	0.0519	0.8299	5.7508	14.93	0.0003
4	gend		5	0.0076	0.8374	5.5406	2.23	0.1418

Summary

- No method is the best for all model selection problems
- Consider more than one criterion
- "Best model" from automatic search procedures should be used as the starting point
- Apply knowledge of the subject matter to make a final selection – use your head!

Model validation

- Three approaches to checking the validity of the model
  - Collect new data, does it fit the model
  - Compare with theory, other data, simulation
  - Use some of the data for the basic analysis and some for validity check, compare SSE with PRESS, MSE with MSPE