Chapter 12 : Variable selection

An example: State SAT scores – In 1982 there was concern for scores of the Scholastic Aptitude Test (SAT) scores that varied greatly between states. Researchers decided to study the scores to determine what factors caused the differences between the states.

```
1          ********************************************************;
2          *** State average SAT scores                          ***;
3          ********************************************************;
4
5          dm'log;clear;output;clear';
6          options nodate nocenter nonumber ps=512 ls=99 nolabel;
7          ODS HTML style=minimal rs=none
7        ! body='C:\Geaghan\Current\EXST3201\Fall2005\SAS\StateSAT01.html' ;
NOTE: Writing HTML Body file: C:\Geaghan\Current\EXST3201\Fall2005\SAS\StateSAT01.html
8
9          Title1 'Chapter 11 : State SAT scores for 1982';
10         filename input1 'C:\Geaghan\Current\EXST3201\Datasets\ASCII\case1201.csv';
11
12         data StateSAT; length state $ 15;
13            infile input1 missover DSD dlm="," firstobs=2;
14            input STATE $ SAT TAKERS INCOME YEARS PUBLIC EXPEND RANK;
15              label SAT = 'Average SAT score (verbal+quantitative)'
16                    Takers = 'percent of total eligible students taking the test'
17                    Income = 'Median income of test takers'
18                    years = 'Ave of yrs of formal study in social sci, nat sci & humanities'
19                    public = 'percentage of takers attending public secondary schools'
20                    expend = 'total state expenditure on secondary schools'
21                    rank = 'median percentile ranking of test takers in their secondary school classes';
22         datalines;
NOTE: The infile INPUT1 is:
      File Name=C:\Geaghan\Current\EXST3201\Datasets\ASCII\case1201.csv,
      RECFM=V,LRECL=256
NOTE: 50 records were read from the infile INPUT1.
      The minimum record length was 64.
      The maximum record length was 99.
NOTE: The data set WORK.STATESAT has 50 observations and 8 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.02 seconds
23         run;
24
25         PROC PRINT DATA=StateSAT; TITLE2 'Data Listing'; RUN;
NOTE: There were 50 observations read from the data set WORK.STATESAT.
NOTE: The PROCEDURE PRINT printed page 1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.26 seconds
      cpu time            0.04 seconds
```

Chapter 11 : State SAT scores for 1982
Data Listing

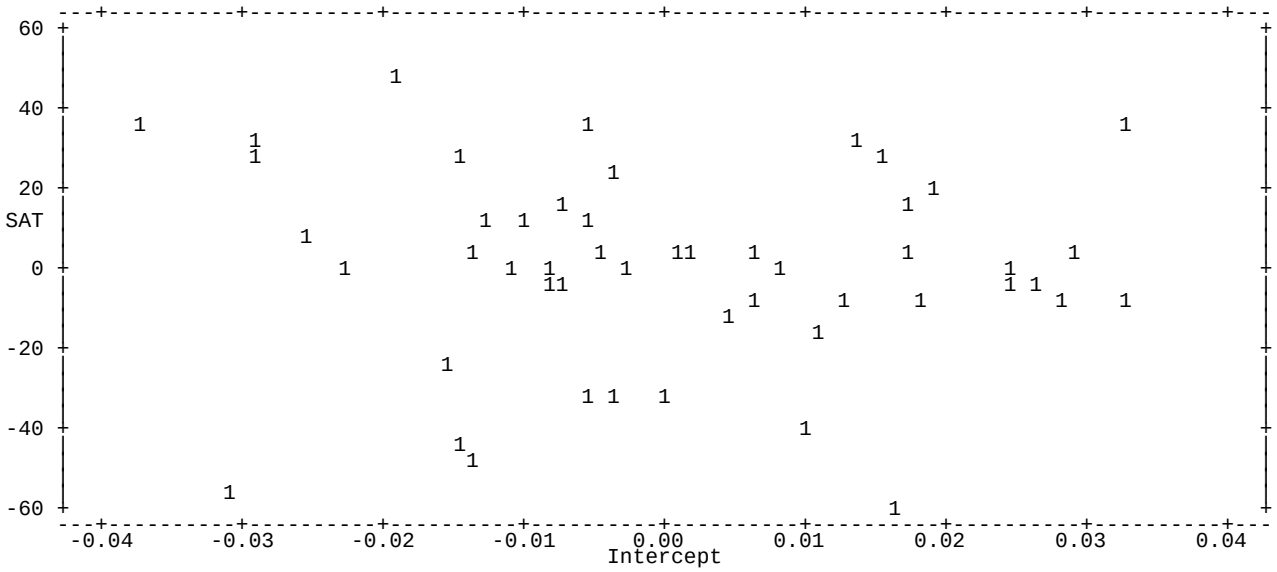
Obs	state	SAT	TAKERS	INCOME	YEARS	PUBLIC	EXPEND	RANK
1	Iowa	1088	3	326	16.7900	87.8000	25.6000	89.7000
2	SouthDakota	1075	2	264	16.0700	86.2000	19.9500	90.6000
3	NorthDakota	1068	3	317	16.5700	88.3000	20.6200	89.8000
4	Kansas	1045	5	338	16.3000	83.9000	27.1400	86.3000
5	Nebraska	1045	5	293	17.2500	83.6000	21.0500	88.5000
6	Montana	1033	8	263	15.9100	93.7000	29.4800	86.4000
7	Minnesota	1028	7	343	17.4100	78.3000	24.8400	83.4000
8	Utah	1022	4	333	16.5700	75.2000	17.4200	85.9000
9	Wyoming	1017	5	328	16.0100	97.0000	25.9600	87.5000
10	Wisconsin	1011	10	304	16.8500	77.3000	27.6900	84.2000
11	Oklahoma	1001	5	358	15.9500	74.2000	20.0700	85.6000
12	Arkansas	999	4	295	15.4900	86.4000	15.7100	89.2000
13	Tennessee	999	9	330	15.7200	61.2000	14.5800	83.4000
14	NewMexico	997	8	316	15.9200	79.5000	22.1900	83.7000
15	Idaho	995	7	285	16.1800	92.1000	17.8000	85.9000
16	Mississippi	988	3	315	16.7600	67.9000	15.3600	90.1000
17	Kentucky	985	6	330	16.6100	71.4000	15.6900	86.4000
18	Colorado	983	16	333	16.8300	88.3000	26.5600	81.8000

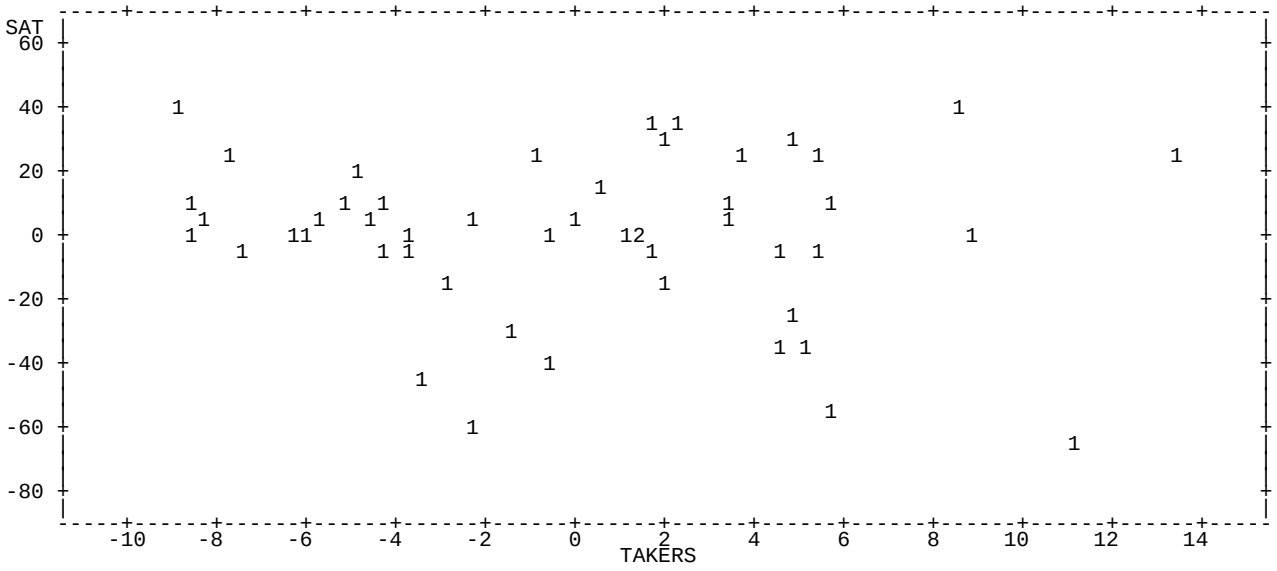
19	Washington	982	19	309	16.2300	87.5000	26.5300	83.2000
20	Arizona	981	11	314	15.9800	80.9000	19.1400	84.3000
21	Illinois	977	14	347	15.8000	74.6000	24.4100	78.7000
22	Louisiana	975	5	394	16.8500	44.8000	19.7200	82.9000
23	Missouri	975	10	322	16.4200	67.7000	20.7900	80.6000
24	Michigan	973	10	335	16.5000	80.7000	24.6100	81.8000
25	WestVirginia	968	7	292	17.0800	90.6000	18.1600	86.2000
26	Alabama	964	6	313	16.3700	69.6000	13.8400	83.9000
27	Ohio	958	16	306	16.5200	71.5000	21.4300	79.5000
28	NewHampshire	925	56	248	16.3500	78.1000	20.3300	73.6000
29	Alaska	923	31	401	15.3200	96.5000	50.1000	79.6000
30	Nevada	917	18	288	14.7300	89.1000	21.7900	81.1000
31	Oregon	908	40	261	14.4800	92.1000	30.4900	79.3000
32	Vermont	904	54	225	16.5000	84.2000	20.1700	75.8000
33	California	899	36	293	15.5200	83.0000	25.9400	77.5000
34	Delaware	897	42	277	16.9500	67.9000	27.8100	71.4000
35	Connecticut	896	69	287	16.7500	76.8000	26.9700	69.8000
36	NewYork	896	59	236	16.8600	80.4000	33.5800	70.5000
37	Maine	890	46	208	16.0500	85.7000	20.5500	74.6000
38	Florida	889	39	255	15.9100	80.5000	22.6200	74.6000
39	Maryland	889	50	312	16.9000	80.4000	25.4100	71.5000
40	Virginia	888	52	295	16.0800	88.8000	22.2300	72.4000
41	Massachusetts	888	65	246	16.7900	80.7000	31.7400	69.9000
42	Pennsylvania	885	50	241	17.2700	78.6000	27.9800	73.4000
43	RhodeIsland	877	59	228	16.6700	79.7000	25.5900	71.4000
44	NewJersey	869	64	269	16.3700	80.6000	27.9100	69.8000
45	Texas	868	32	303	14.9500	91.7000	19.5500	76.4000
46	Indiana	860	48	258	14.3900	90.2000	17.9300	74.1000
47	Hawaii	857	47	277	16.4000	67.6000	21.2100	69.9000
48	NorthCarolina	827	47	224	15.3100	92.8000	19.9200	75.3000
49	Georgia	823	51	250	15.5500	86.5000	16.5200	74.0000
50	SouthCarolina	790	48	214	15.4200	88.1000	15.6000	74.0000

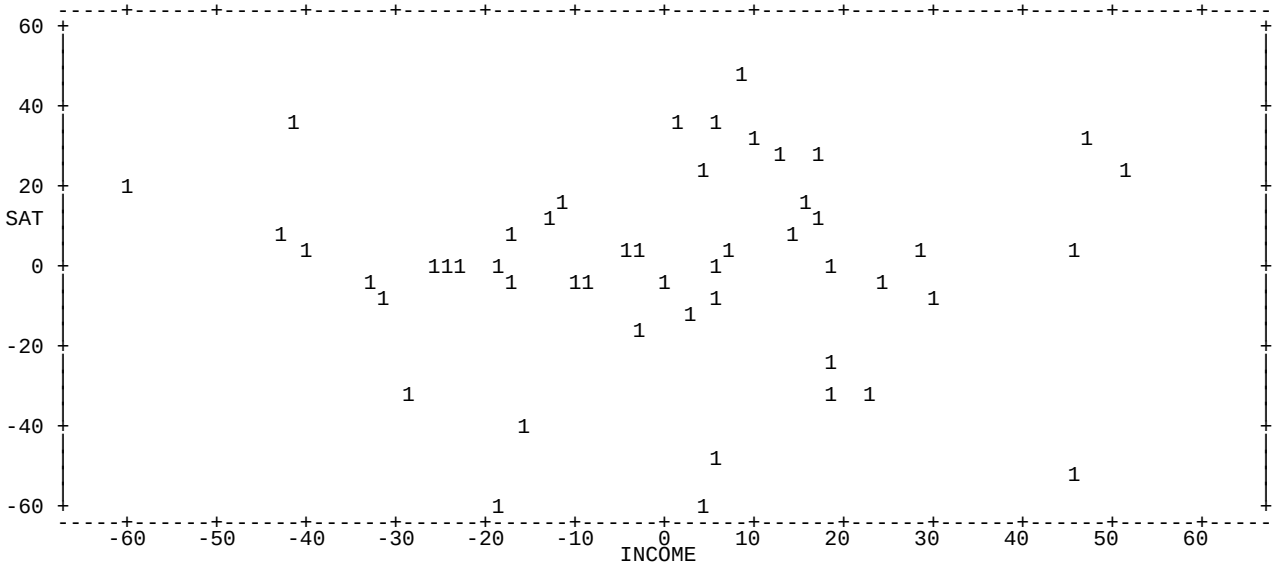
```
27         options ps=45 ls=99 nolabel;
28         PROC REG DATA=StateSAT; Title2 'Fit of SAT with REG';
29            MODEL SAT = TAKERS INCOME YEARS PUBLIC EXPEND RANK / VIF partial;
30         RUN;
NOTE: The PROCEDURE REG printed pages 2-9.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.28 seconds
      cpu time            0.07 seconds


Chapter 11 : State SAT scores for 1982
Fit of SAT with REG

The REG Procedure
Model: MODEL1
Dependent Variable: SAT

Number of Observations Read          50
Number of Observations Used          50

Analysis of Variance                Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F
Model                      6        216168          36028      51.91    <.0001
Error                     43         29842      694.00969
Corrected Total           49        246011

Parameter Estimates
                        Parameter       Standard                             Variance
Variable     DF          Estimate          Error    t Value    Pr > |t|      Inflation
Intercept     1         -94.65897      211.50956      -0.45      0.6567              0
TAKERS        1          -0.48008        0.69371      -0.69      0.4926        16.47863
INCOME        1          -0.00820        0.15236      -0.05      0.9574         3.12885
YEARS         1          22.61007        6.31458       3.58      0.0009         1.37941
PUBLIC        1          -0.46415        0.57910      -0.80      0.4272         2.28840
EXPEND        1           2.21201        0.84597       2.61      0.0123         1.90799
RANK          1           8.47622        2.10781       4.02      0.0002        13.34739
```

Chapter 11 : State SAT scores for 1982
Fit of SAT with REG

The REG Procedure
Model: MODEL1
Partial Regression Residual Plot

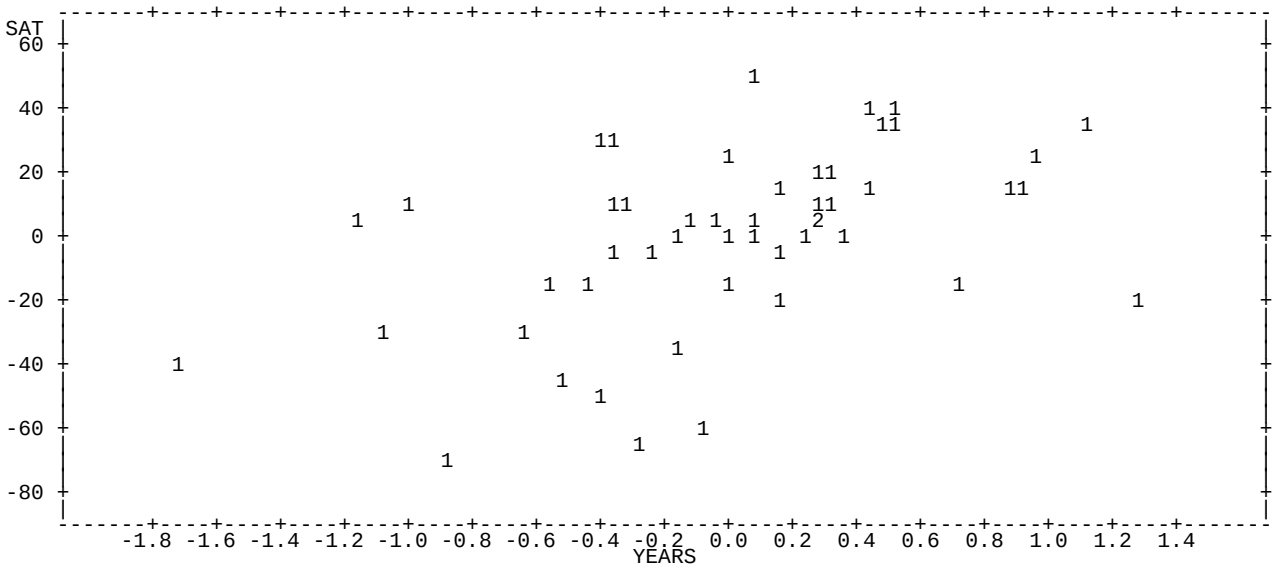


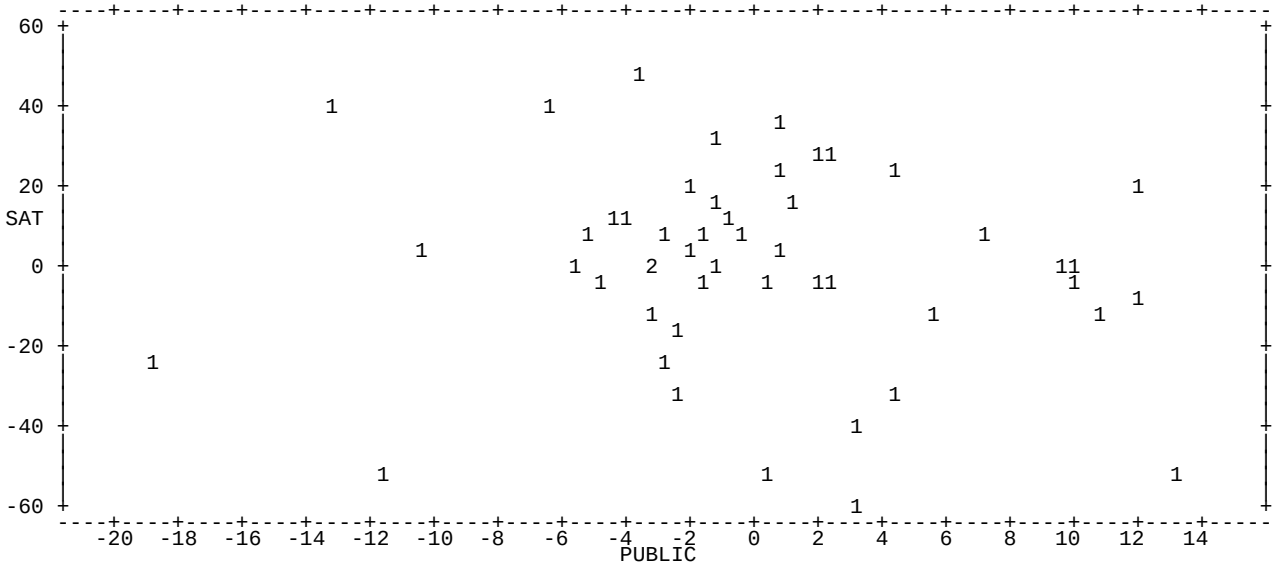


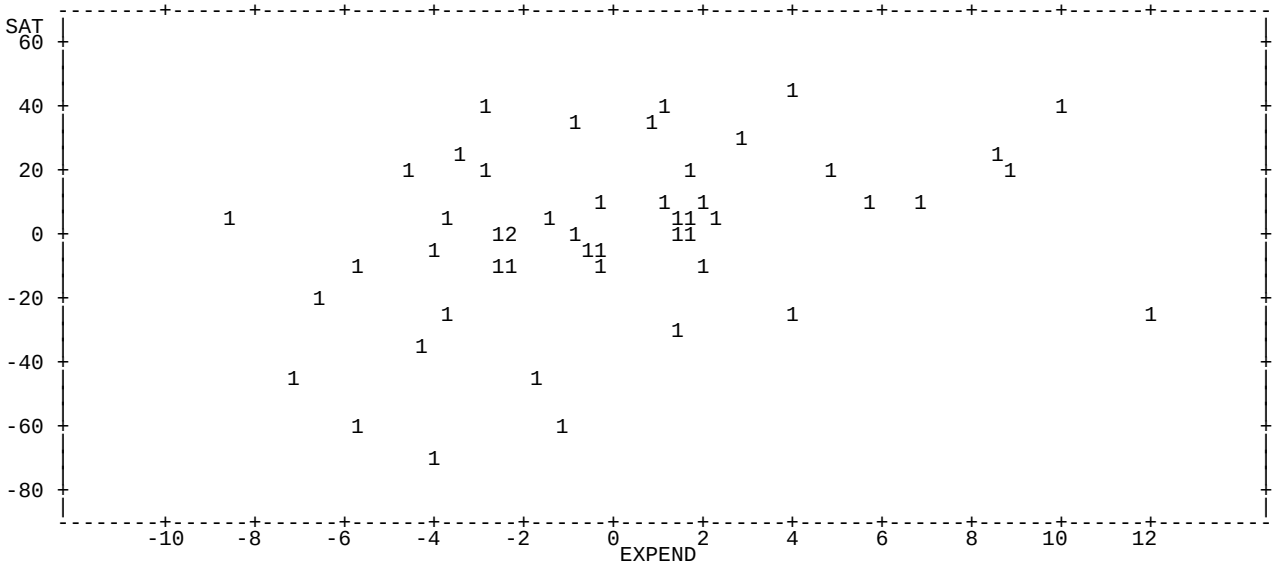


Chapter 11 : State SAT scores for 1982
Fit of SAT with REG

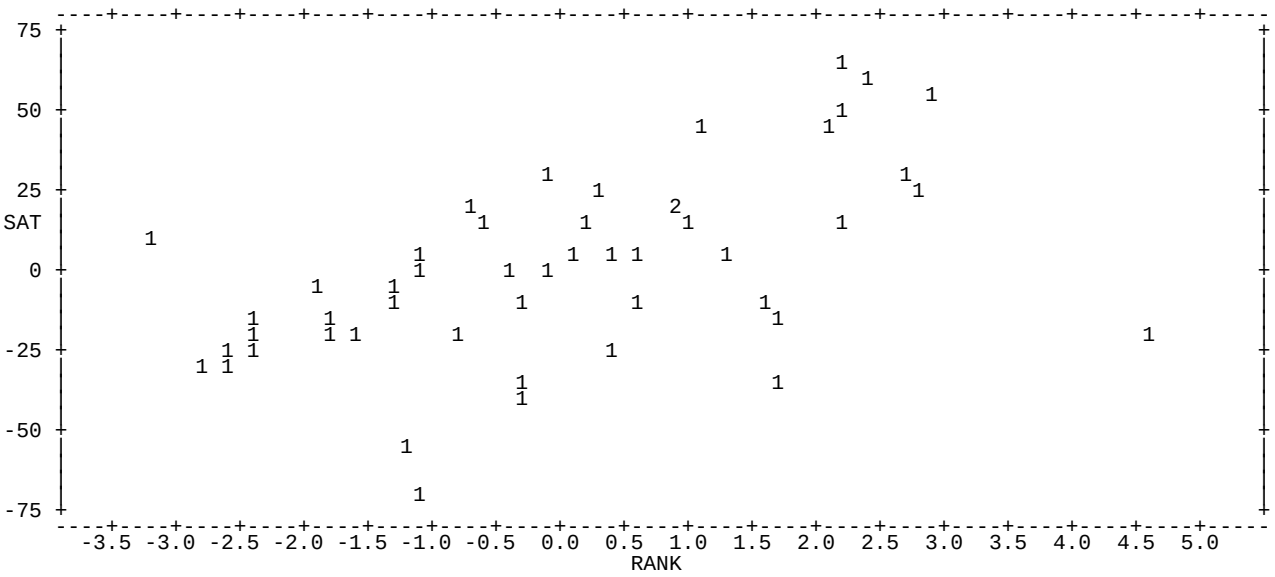
The REG Procedure
Model: MODEL1
Partial Regression Residual Plot







```
Chapter 11 : State SAT scores for 1982
Fit of SAT with REG

The REG Procedure
Model: MODEL1
Partial Regression Residual Plot
```



First, fit and examine the model. Note the VIF and numerous non-significant variables. We want to find a subset of significant variables that describe the variation in SAT scores. The approach we use will also tend to reduce the multicolinearity problem (though not necessarily).

The text suggests 3 specific issues related to this process of variable selection.

1) **Adjusting for a large set of explanatory variables**. There may be a specific hypothesis to be tested, and that variable should be included in the analysis (though not necessarily initially), but variance can be reduced by including other explanatory variables to refine the analysis.

2) **Fishing for an explanation.** Often there is not clear variable that must be included or any interest in testing a particular variable. The objective is to determine which variable may be correlated and may ultimately be useful explanatory varibles.

3) **Prediction.** The objective may not be interpretation, but rather prediction. The variable selection procedure can be used to find useful independent variables for future predictions.

Variable selection techniques.

Stepwise Regression : we know that we should not delete batches of variable from regressions when they are not significant, because we do not know what the effect of adjusting for each other has been (except when order dependent TYPE I SS has been used, as in a pure polynomial).

Generally, when we wish to examine the effect of adding and deleting variables, this should be done one variable at a time. There is an algorithm to do this automatically called stepwise regression.

Consider “p–1” candidate variables where we want to select the best set. One possibility is to fit and examine all possible combinations of the variables. This “all possible regressions” approach is used, but there are other more automated procedures.

1) **FORWARD stepwise regression**: Start by fitting all simple linear regressions between Y_i and the "p–1" different X_i variables. These would usually be evaluated with the F statistic (TYPE III) to determine which variable is the best predictor. So, out of the p–1 variables, the one with the best F statistic is the one that fits best. Now, that the BEST variable has been selected we want to determine if there is a second variable model which is also useful.

   There are p–2 variables remaining, so we try our "best" variable together with each of the p–2 remaining variable to determine if there is a second variable that is also significant. If there is a second best variable then it is included with the first, giving a "two factor model" as the best model.

   At this point there are p–3 variables remaining, so we try our "best two factors" together with each of the p–3 remaining variable to determine if there is a third best variable that is significant. If there is a third best variable then it is included with the first two, giving a "three factor model" as the best model. This process is repeated until no variable can be found that will enter the model at some predetermined criteria.

   With Forward stepwise regression, once a variable is entered into the model, it stays in the model. There are no provisions for removing variables once entered.

   This process is called FORWARD stepwise regression. Sometimes, no single variable will enter and be significant, but some group of 2 or 3, between them, will lower the MSE sufficiently for all to be significant. FORWARD will not catch this.

   Although the F statistic is the usual criteria, other criteria can be used, such as the R^2.

2) **BACKWARDS elimination**. This procedure starts with a fit the full model (p–1 variables). Examine the F statistic (TYPE II, fully adjusted for all other variables in the model). We first find the least significant of the p–1 variables, and eliminate this variable only. THIS IS DONE ONE VARIABLE AT A TIME. Now, refit the full model, minus the eliminated variable (p–2 variables). Again find the least significant and eliminate this variable. Repeat until all variables in the model are significant at some predetermine level of significance.

   Sometimes, a variable added to a model by FORWARD selection is not significant at some later point, after being adjusted for other variables in the model. Since no variable can be removed once entered, backwards elimination can avoid this problem.

3) **STEPWISE selection** (FORWARD selection with a backward glance). This selection proceeds as does FORWARD, but after each addition all variables in the model are examined to insure that they remain significant. If not significant, they are removed from the model.

Selection criteria

R_p^2, $AdjR_p^2$ and MSE_p can be used to graphically compare and evaluate models. The subscript p refers to the number of parameters in the model

Mallow's Cp criterion – Use of this statistic presumes no bias in the full model MSE, so the full model should be carefully chosen to have little or no multicolinearity. The Cp statistics will be approximately equal to p if there is no bias in the regression model

$$Cp = p + (n - p)\left(\hat{\sigma}^2 - \hat{\sigma}^2_{full} \Big/ \hat{\sigma}^2_{full} \right)$$

SAS application – Selection options: FORWARD, BACKWARD, STEPWISE, NONE (full model). These do the selection techniques discussed above. Additional options for working with these are;

SLENTRY = alpha value; level of entry for stepwise and forward

SLSTAY = alpha value; level of remaining for backward and stepwise

INCLUDE = n; automatically puts the first n variables of the list into the regression permanently (they will not be removed)

RSQUARE, ADJRSQ, CP : These are not sequential model development techniques. These are more like "all possible regressions" analysis, but they simply calculate the R^2, AdjR^2 and the C_p statistic for all models to find the best ones. Options for controlling these techniques are:

START = n

STOP = n (will stop all methods)

BEST = n

```
32         options ps=512 ls=111 nolabel;
33         PROC REG DATA=StateSAT; Title2 'Forward selection stepwise regression';
34            MODEL SAT = TAKERS INCOME YEARS PUBLIC EXPEND RANK / selection=forward;
35         RUN;
NOTE: The PROCEDURE REG printed page 10.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.17 seconds
      cpu time            0.10 seconds


Chapter 11 : State SAT scores for 1982
Forward selection stepwise regression

The REG Procedure
Model: MODEL1
Dependent Variable: SAT

Number of Observations Read          50
Number of Observations Used          50

Forward Selection: Step 1
Variable RANK Entered: R-Square = 0.7742 and C(p) = 34.0268

Analysis of Variance
                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       1      190471         190471     164.62    <.0001
Error                      48       55539     1157.07070
Corrected Total            49      246011

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F
Intercept    183.41869     59.78129        10892     9.41  0.0035
RANK           9.55795      0.74495       190471   164.62  <.0001

Bounds on condition number: 1, 1
--------------------------------------------------------------------------------
```

```
Forward Selection: Step 2
Variable YEARS Entered: R-Square = 0.8471 and C(p) = 10.2152

Analysis of Variance
                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       2      208385         104192     130.15    <.0001
Error                      47       37626      800.54988
Corrected Total            49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept   -243.93083    103.12221   4479.37428     5.60  0.0222
YEARS         27.38289      5.78872        17914    22.38  <.0001
RANK           9.35160      0.62118       181437   226.64  <.0001

Bounds on condition number: 1.005, 4.0198
--------------------------------------------------------------------------------

Forward Selection: Step 3
Variable EXPEND Entered: R-Square = 0.8711 and C(p) = 3.6884

Analysis of Variance
                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       3      214303          71434     103.63    <.0001
Error                      46       31708      689.30879
Corrected Total            49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept   -303.72425     97.84154   6642.41862     9.64  0.0033
YEARS         26.09522      5.38945        16160    23.44  <.0001
EXPEND         1.86087      0.63511   5917.64030     8.58  0.0053
RANK           9.82580      0.59870       185667   269.35  <.0001

Bounds on condition number: 1.0842, 9.5357
--------------------------------------------------------------------------------

Forward Selection: Step 4
Variable PUBLIC Entered: R-Square = 0.8771 and C(p) = 3.5812

Analysis of Variance
                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       4      215765          53941      80.25    <.0001
Error                      45       30246      672.12769
Corrected Total            49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept   -204.59819    117.68711   2031.40994     3.02  0.0890
YEARS         21.89047      6.03718   8836.76106    13.15  0.0007
PUBLIC        -0.66380      0.45001   1462.45814     2.18  0.1472
EXPEND         2.24164      0.67819   7343.09123    10.93  0.0019
RANK          10.00317      0.60329       184786   274.93  <.0001

Bounds on condition number: 1.4268, 20.496
--------------------------------------------------------------------------------
```

```
Forward Selection: Step 5
Variable TAKERS Entered: R-Square = 0.8787 and C(p) = 5.0029

Analysis of Variance
                                    Sum of           Mean
Source                      DF     Squares         Square    F Value    Pr > F
Model                        5      216166          43233      63.74    <.0001
Error                       44       29844      678.28238
Corrected Total             49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept   -100.47360    179.72555    211.97998     0.31  0.5790
TAKERS        -0.46208      0.60073    401.32149     0.59  0.4459
YEARS         22.66880      6.14859   9219.68008    13.59  0.0006
PUBLIC        -0.45226      0.52914    495.49979     0.73  0.3973
EXPEND         2.18591      0.68513   6904.41715    10.18  0.0026
RANK           8.49641      2.05047        11646    17.17  0.0002

Bounds on condition number: 12.924, 150.71
--------------------------------------------------------------------------------

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

                          Summary of Forward Selection

        Variable     Number     Partial      Model
Step    Entered      Vars In    R-Square    R-Square     C(p)      F Value    Pr > F
  1     RANK             1      0.7742      0.7742     34.0268     164.62    <.0001
  2     YEARS            2      0.0728      0.8471     10.2152      22.38    <.0001
  3     EXPEND           3      0.0241      0.8711      3.6884       8.58    0.0053
  4     PUBLIC           4      0.0059      0.8771      3.5812       2.18    0.1472
  5     TAKERS           5      0.0016      0.8787      5.0029       0.59    0.4459


36         PROC REG DATA=StateSAT; Title2 'Backward elimination stepwise regression';
37            MODEL SAT = TAKERS INCOME YEARS PUBLIC EXPEND RANK / selection=backward;
38         RUN;
NOTE: The PROCEDURE REG printed page 11.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.13 seconds
      cpu time            0.05 seconds


Chapter 11 : State SAT scores for 1982
Backward elimination stepwise regression

The REG Procedure
Model: MODEL1
Dependent Variable: SAT

Number of Observations Read          50
Number of Observations Used          50
```

```
Backward Elimination: Step 0
All Variables Entered: R-Square = 0.8787 and C(p) = 7.0000

Analysis of Variance
                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F
Model                      6        216168          36028      51.91    <.0001
Error                     43         29842      694.00969
Corrected Total           49        246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept     -94.65897    211.50956    139.00459     0.20  0.6567
TAKERS         -0.48008      0.69371    332.38174     0.48  0.4926
INCOME         -0.00820      0.15236      2.00790     0.00  0.9574
YEARS          22.61007      6.31458   8897.76835    12.82  0.0009
PUBLIC         -0.46415      0.57910    445.83415     0.64  0.4272
EXPEND          2.21201      0.84597   4744.88313     6.84  0.0123
RANK            8.47622      2.10781        11223    16.17  0.0002

Bounds on condition number: 16.479, 231.18
------------------------------------------------------------------------------

Backward Elimination: Step 1
Variable INCOME Removed: R-Square = 0.8787 and C(p) = 5.0029

Analysis of Variance
                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F
Model                      5        216166          43233      63.74    <.0001
Error                     44         29844      678.28238
Corrected Total           49        246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    -100.47360    179.72555    211.97998     0.31  0.5790
TAKERS         -0.46208      0.60073    401.32149     0.59  0.4459
YEARS          22.66880      6.14859   9219.68008    13.59  0.0006
PUBLIC         -0.45226      0.52914    495.49979     0.73  0.3973
EXPEND          2.18591      0.68513   6904.41715    10.18  0.0026
RANK            8.49641      2.05047        11646    17.17  0.0002

Bounds on condition number: 12.924, 150.71
------------------------------------------------------------------------------

Backward Elimination: Step 2
Variable TAKERS Removed: R-Square = 0.8771 and C(p) = 3.5812

Analysis of Variance
                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F
Model                      4        215765          53941      80.25    <.0001
Error                     45         30246      672.12769
Corrected Total           49        246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    -204.59819    117.68711   2031.40994     3.02  0.0890
YEARS          21.89047      6.03718   8836.76106    13.15  0.0007
PUBLIC         -0.66380      0.45001   1462.45814     2.18  0.1472
EXPEND          2.24164      0.67819   7343.09123    10.93  0.0019
RANK           10.00317      0.60329       184786   274.93  <.0001
```

```
Bounds on condition number: 1.4268, 20.496
------------------------------------------------------------------------------

Backward Elimination: Step 3
Variable PUBLIC Removed: R-Square = 0.8711 and C(p) = 3.6884

Analysis of Variance
                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       3      214303          71434     103.63    <.0001
Error                      46       31708      689.30879
Corrected Total            49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    -303.72425     97.84154   6642.41862     9.64  0.0033
YEARS          26.09522      5.38945        16160    23.44  <.0001
EXPEND          1.86087      0.63511   5917.64030     8.58  0.0053
RANK            9.82580      0.59870       185667   269.35  <.0001

Bounds on condition number: 1.0842, 9.5357
------------------------------------------------------------------------------

All variables left in the model are significant at the 0.1000 level.

                        Summary of Backward Elimination

        Variable     Number     Partial      Model
Step    Removed      Vars In    R-Square    R-Square     C(p)      F Value    Pr > F
  1     INCOME           5       0.0000      0.8787      5.0029       0.00    0.9574
  2     TAKERS           4       0.0016      0.8771      3.5812       0.59    0.4459
  3     PUBLIC           3       0.0059      0.8711      3.6884       2.18    0.1472
```

```
39         PROC REG DATA=StateSAT; Title2 'Stepwise regression';
40            MODEL SAT = TAKERS INCOME YEARS PUBLIC EXPEND RANK /
41                  selection=stepwise slentry=0.20 slstay=0.05;
42         RUN;
NOTE: The PROCEDURE REG printed page 12.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.12 seconds
      cpu time            0.04 seconds
```

```
Chapter 11 : State SAT scores for 1982
Stepwise regression

The REG Procedure
Model: MODEL1
Dependent Variable: SAT

Number of Observations Read          50
Number of Observations Used          50
```

```
Stepwise Selection: Step 1
Variable RANK Entered: R-Square = 0.7742 and C(p) = 34.0268

Analysis of Variance
                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       1      190471         190471     164.62    <.0001
Error                      48       55539     1157.07070
Corrected Total            49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept     183.41869     59.78129        10892     9.41  0.0035
RANK            9.55795      0.74495       190471   164.62  <.0001

Bounds on condition number: 1, 1
--------------------------------------------------------------------------------

Stepwise Selection: Step 2
Variable YEARS Entered: R-Square = 0.8471 and C(p) = 10.2152

Analysis of Variance
                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       2      208385         104192     130.15    <.0001
Error                      47       37626      800.54988
Corrected Total            49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    -243.93083    103.12221   4479.37428     5.60  0.0222
YEARS          27.38289      5.78872        17914    22.38  <.0001
RANK            9.35160      0.62118       181437   226.64  <.0001

Bounds on condition number: 1.005, 4.0198
--------------------------------------------------------------------------------

Stepwise Selection: Step 3
Variable EXPEND Entered: R-Square = 0.8711 and C(p) = 3.6884

Analysis of Variance
                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       3      214303          71434     103.63    <.0001
Error                      46       31708      689.30879
Corrected Total            49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    -303.72425     97.84154   6642.41862     9.64  0.0033
YEARS          26.09522      5.38945        16160    23.44  <.0001
EXPEND          1.86087      0.63511   5917.64030     8.58  0.0053
RANK            9.82580      0.59870       185667   269.35  <.0001

Bounds on condition number: 1.0842, 9.5357
--------------------------------------------------------------------------------
```

```
Stepwise Selection: Step 4
Variable PUBLIC Entered: R-Square = 0.8771 and C(p) = 3.5812

Analysis of Variance
                                    Sum of           Mean
Source                      DF     Squares         Square    F Value    Pr > F
Model                        4      215765          53941      80.25    <.0001
Error                       45       30246      672.12769
Corrected Total             49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept   -204.59819    117.68711   2031.40994     3.02  0.0890
YEARS         21.89047      6.03718   8836.76106    13.15  0.0007
PUBLIC        -0.66380      0.45001   1462.45814     2.18  0.1472
EXPEND         2.24164      0.67819   7343.09123    10.93  0.0019
RANK          10.00317      0.60329       184786   274.93  <.0001

Bounds on condition number: 1.4268, 20.496
------------------------------------------------------------------------------

Stepwise Selection: Step 5
Variable PUBLIC Removed: R-Square = 0.8711 and C(p) = 3.6884

Analysis of Variance
                                    Sum of           Mean
Source                      DF     Squares         Square    F Value    Pr > F
Model                        3      214303          71434     103.63    <.0001
Error                       46       31708      689.30879
Corrected Total             49      246011

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept   -303.72425     97.84154   6642.41862     9.64  0.0033
YEARS         26.09522      5.38945        16160    23.44  <.0001
EXPEND         1.86087      0.63511   5917.64030     8.58  0.0053
RANK           9.82580      0.59870       185667   269.35  <.0001

Bounds on condition number: 1.0842, 9.5357
------------------------------------------------------------------------------

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
       Variable     Variable     Number     Partial      Model
Step   Entered      Removed      Vars In    R-Square    R-Square     C(p)      F Value    Pr > F
  1    RANK                         1      0.7742      0.7742     34.0268     164.62    <.0001
  2    YEARS                        2      0.0728      0.8471     10.2152      22.38    <.0001
  3    EXPEND                       3      0.0241      0.8711      3.6884       8.58    0.0053
  4    PUBLIC                       4      0.0059      0.8771      3.5812       2.18    0.1472
  5                 PUBLIC          3      0.0059      0.8711      3.6884       2.18    0.1472


43         PROC REG DATA=StateSAT; Title2 'Reduced model with diagnostics';
44            MODEL SAT = YEARS EXPEND RANK / VIF;
45            output out=next1 r=resid p=yhat lclm=lclm uclm=uclm lcl=lcli ucl=ucli
46                 student=student rstudent=rstudent cookd=cookd h=leverage dffits=dffits;
47         RUN;
NOTE: The data set WORK.NEXT1 has 50 observations and 19 variables.
NOTE: The PROCEDURE REG printed page 13.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.13 seconds
      cpu time            0.06 seconds
```

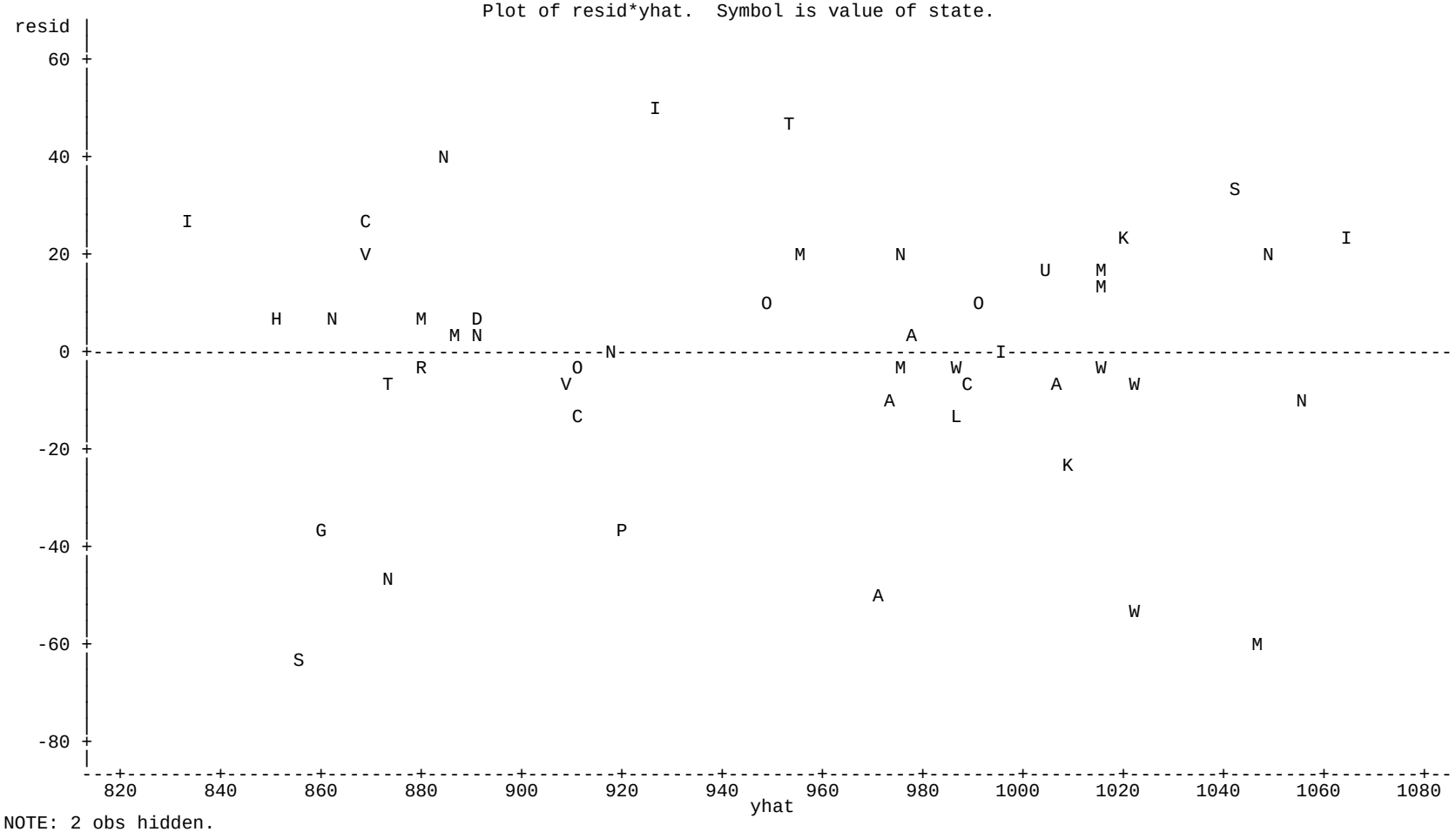
```
Chapter 11 : State SAT scores for 1982
Reduced model with diagnostics

The REG Procedure
Model: MODEL1
Dependent Variable: SAT

Number of Observations Read          50
Number of Observations Used          50

Analysis of Variance
                                  Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       3         214303          71434     103.63    <.0001
Error                      46          31708      689.30879
Corrected Total            49         246011

Root MSE              26.25469    R-Square     0.8711
Dependent Mean       947.94000    Adj R-Sq     0.8627
Coeff Var              2.76966

Parameter Estimates
                       Parameter       Standard                                 Variance
Variable     DF         Estimate          Error    t Value    Pr > |t|      Inflation
Intercept     1      -303.72425       97.84154      -3.10      0.0033              0
YEARS         1        26.09522        5.38945       4.84      <.0001        1.01168
EXPEND        1         1.86087        0.63511       2.93      0.0053        1.08271
RANK          1         9.82580        0.59870      16.41      <.0001        1.08418


49         proc print data=next1;
50            var state SAT yhat resid student rstudent cookd leverage dffits;
51         RUN;
NOTE: There were 50 observations read from the data set WORK.NEXT1.
NOTE: The PROCEDURE PRINT printed page 14.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.07 seconds
      cpu time            0.03 seconds
52         options ps=60 ls=132;
53         proc plot data=next1;  TITLE2 'Various plot with state variable';
54            plot resid * yhat = state / vref=0;
55         RUN;
55       !      options ps=40 ls=132;
NOTE: There were 50 observations read from the data set WORK.NEXT1.
NOTE: The PROCEDURE PLOT printed page 15.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.06 seconds
      cpu time            0.00 seconds
56         proc plot data=next1;  TITLE2 'Various plot with state variable';
57            plot student * state / vref=2;
58            plot rstudent * state / vref=2;
59            plot leverage * state / vref=0.16; /* 2p/n = 2*4/50 = 0.16 */
60            plot cookd * state / vref=1;
61            plot dffits * state / vref=1;
62         RUN;
62       !      OPTIONS PS=256 ls=132;
NOTE: There were 50 observations read from the data set WORK.NEXT1.
NOTE: The PROCEDURE PLOT printed pages 16-20.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.06 seconds
      cpu time            0.00 seconds
```

Chapter 11 : State SAT scores for 1982
Reduced model with diagnostics

Obs	state	SAT	yhat	resid	student	rstudent	cookd	leverage	dffits
1	Iowa	1088	1063.43	24.5735	0.98062	0.98020	0.02348	0.08899	0.30636
2	SouthDakota	1075	1042.97	32.0328	1.26930	1.27800	0.03315	0.07605	0.36666
3	NorthDakota	1068	1049.40	18.5990	0.73443	0.73070	0.01009	0.06963	0.19989
4	Kansas	1045	1020.10	24.9021	0.97742	0.97693	0.01480	0.05833	0.24315
5	Nebraska	1045	1055.17	-10.1724	-0.40723	-0.40351	0.00434	0.09480	-0.13059
6	Montana	1033	1015.26	17.7423	0.70671	0.70281	0.01169	0.08563	0.21507
7	Minnesota	1028	1016.29	11.7112	0.46645	0.46245	0.00509	0.08552	0.14142
8	Utah	1022	1005.13	16.8744	0.66001	0.65591	0.00594	0.05171	0.15317
9	Wyoming	1017	1022.13	-5.1254	-0.20186	-0.19974	0.00070	0.06471	-0.05254
10	Wisconsin	1011	1014.84	-3.8395	-0.15090	-0.14929	0.00037	0.06081	-0.03799
11	Oklahoma	1001	990.93	10.0699	0.39143	0.38780	0.00159	0.03988	0.07903
12	Arkansas	999	1006.19	-7.1858	-0.28820	-0.28531	0.00226	0.09812	-0.09410
13	Tennessee	999	953.10	45.9046	1.80996	1.85754	0.05865	0.06683	0.49709
14	NewMexico	997	975.42	21.5767	0.83482	0.83202	0.00556	0.03090	0.14857
15	Idaho	995	995.66	-0.6557	-0.02555	-0.02527	0.00001	0.04474	-0.00547
16	Mississippi	988	1047.52	-59.5187	-2.38378	-2.51840	0.15016	0.09560	-0.81878
17	Kentucky	985	1007.86	-22.8631	-0.90078	-0.89890	0.01420	0.06542	-0.23783
18	Colorado	983	988.63	-5.6330	-0.21951	-0.21723	0.00056	0.04470	-0.04699
19	Washington	982	986.68	-4.6761	-0.18141	-0.17949	0.00031	0.03612	-0.03474
20	Arizona	981	977.21	3.7911	0.14704	0.14547	0.00020	0.03564	0.02797
21	Illinois	977	927.29	49.7060	1.92091	1.98103	0.02718	0.02862	0.34005
22	Louisiana	975	987.23	-12.2349	-0.47691	-0.47287	0.00269	0.04518	-0.10286
23	Missouri	975	955.41	19.5943	0.75571	0.75213	0.00362	0.02470	0.11970
24	Michigan	973	976.39	-3.3929	-0.13100	-0.12960	0.00012	0.02691	-0.02155
25	WestVirginia	968	1022.76	-54.7589	-2.16916	-2.26439	0.09605	0.07549	-0.64705
26	Alabama	964	973.59	-9.5931	-0.37844	-0.37489	0.00260	0.06779	-0.10110
27	Ohio	958	948.40	9.6021	0.37064	0.36714	0.00093	0.02631	0.06035
28	NewHampshire	925	883.94	41.0575	1.60630	1.63526	0.03552	0.05219	0.38374
29	Alaska	923	971.42	-48.4172	-2.60892	-2.79550	1.70402	0.50035	-2.79748
30	Nevada	917	918.08	-1.0786	-0.04363	-0.04315	0.00006	0.11323	-0.01542
31	Oregon	908	910.06	-2.0579	-0.08695	-0.08601	0.00044	0.18737	-0.04130
32	Vermont	904	909.18	-5.1758	-0.20146	-0.19935	0.00045	0.04243	-0.04196
33	California	899	911.04	-12.0436	-0.46976	-0.46574	0.00269	0.04643	-0.10277
34	Delaware	897	891.90	5.0978	0.20301	0.20088	0.00096	0.08522	0.06131
35	Connecticut	896	869.40	26.6012	1.06001	1.06147	0.02656	0.08637	0.32637
36	NewYork	896	891.45	4.5524	0.18485	0.18290	0.00117	0.12013	0.06758
37	Maine	890	886.35	3.6509	0.14210	0.14058	0.00022	0.04243	0.02959
38	Florida	889	886.55	2.4522	0.09523	0.09420	0.00009	0.03806	0.01874
39	Maryland	889	887.11	1.8861	0.07483	0.07402	0.00012	0.07842	0.02159
40	Virginia	888	868.64	19.3585	0.75718	0.75361	0.00782	0.05173	0.17601

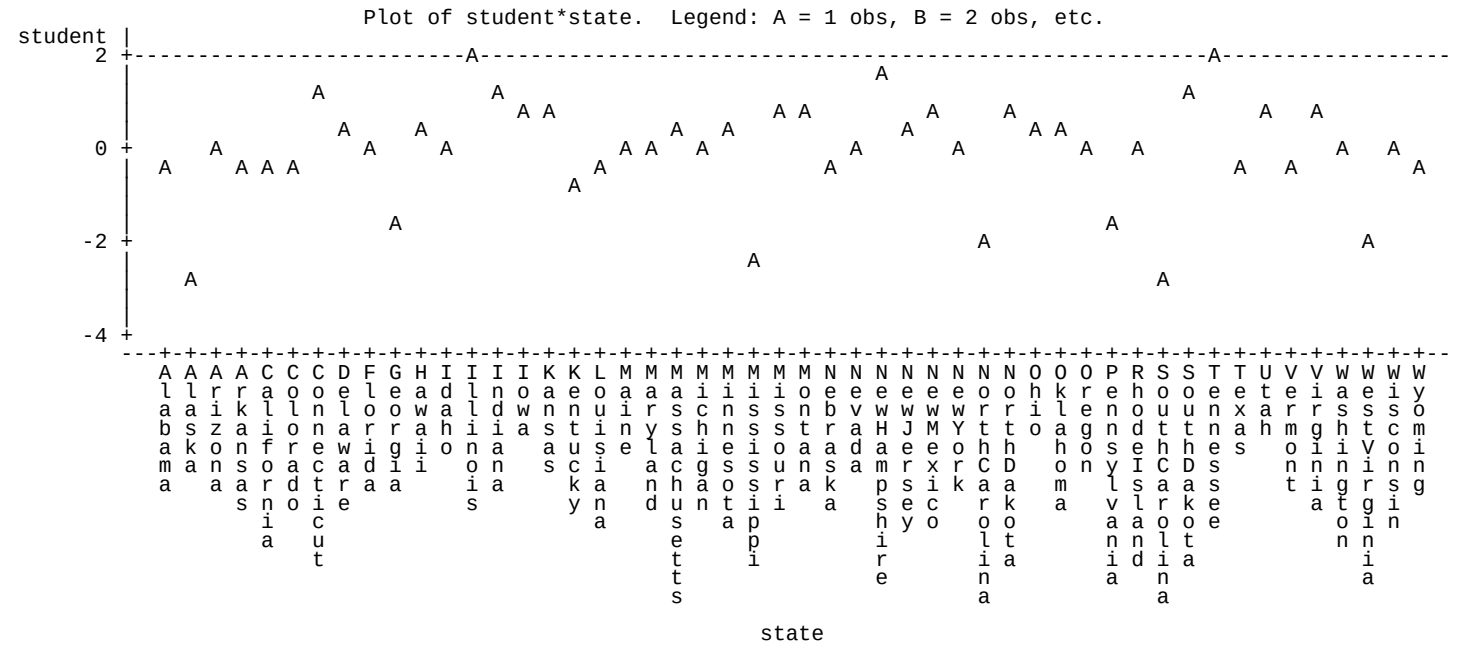
41	Massachusetts	888	880.30	7.6985	0.31022	0.30715	0.00287	0.10657	0.10608
42	Pennsylvania	885	920.22	-35.2206	-1.41124	-1.42705	0.05311	0.09639	-0.46609
43	RhodeIsland	877	880.46	-3.4644	-0.13661	-0.13514	0.00033	0.06694	-0.03620
44	NewJersey	869	861.23	7.7682	0.30765	0.30460	0.00192	0.07507	0.08678
45	Texas	868	873.47	-5.4700	-0.21921	-0.21693	0.00129	0.09672	-0.07099
46	Indiana	860	833.24	26.7573	1.12875	1.13220	0.07219	0.18478	0.53903
47	Hawaii	857	850.53	6.4706	0.25743	0.25480	0.00151	0.08342	0.07687
48	NorthCarolina	827	872.74	-45.7444	-1.80617	-1.85336	0.06086	0.06944	-0.50627
49	Georgia	823	859.91	-36.9068	-1.47039	-1.48975	0.05088	0.08603	-0.45707
50	SouthCarolina	790	854.80	-64.8024	-2.60336	-2.78852	0.19062	0.10113	-0.93531

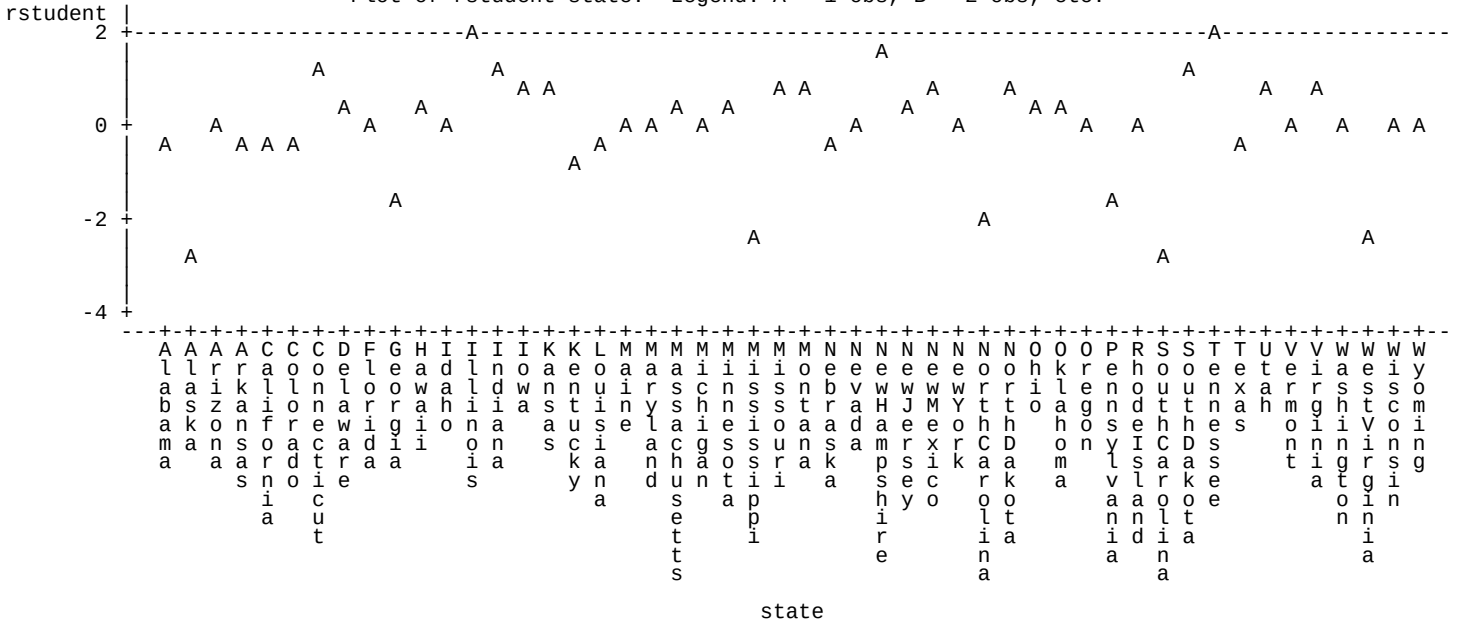
Chapter 11 : State SAT scores for 1982
Various plot with state variable



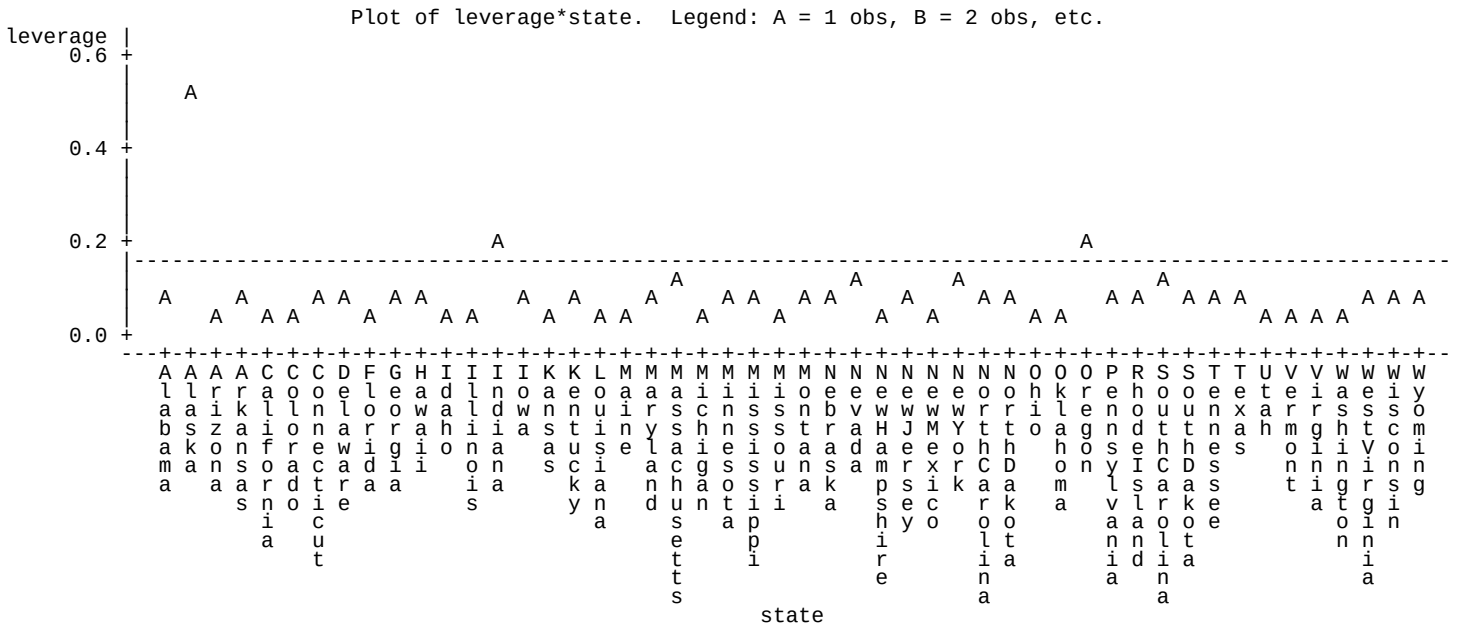
NOTE: 2 obs hidden.

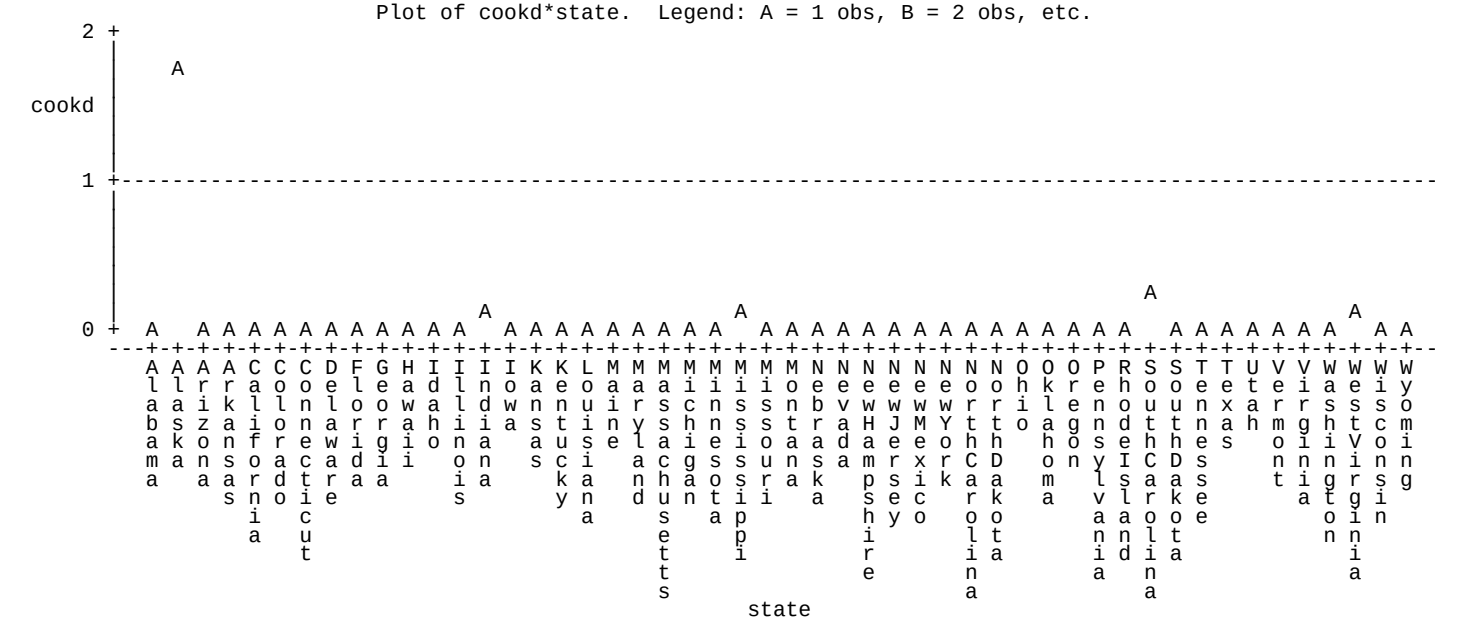
Chapter 11 : State SAT scores for 1982
Various plot with state variable



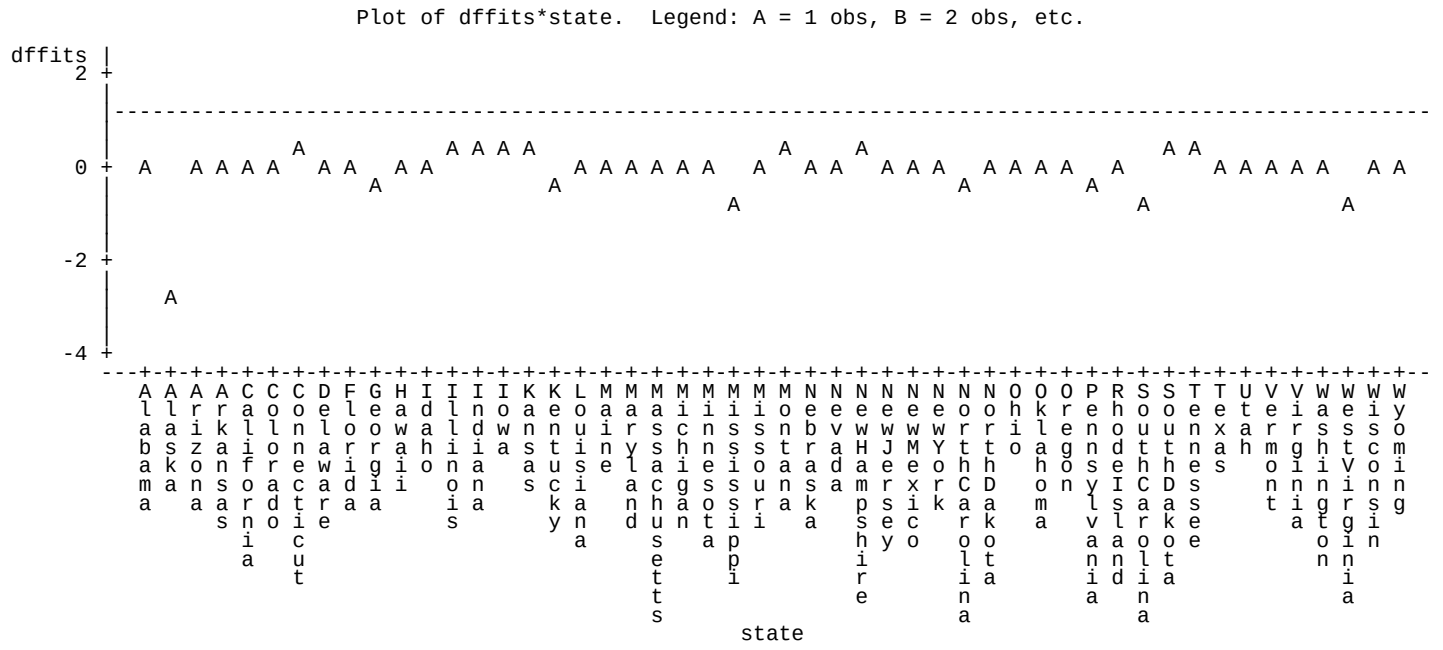


Chapter 11 : State SAT scores for 1982
Various plot with state variable





Chapter 11 : State SAT scores for 1982
Various plot with state variable




```
63         PROC UNIVARIATE DATA=NEXT1 NORMAL PLOT; VAR resid; RUN;
NOTE: The PROCEDURE UNIVARIATE printed page 21.
NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.09 seconds
      cpu time            0.04 seconds
64
```

Chapter 11 : State SAT scores for 1982
Various plot with state variable

The UNIVARIATE Procedure
Variable: resid

Moments

N	50	Sum Weights	50
Mean	0	Sum Observations	0
Std Deviation	25.4382824	Variance	647.10621
Skewness	-0.6876412	Kurtosis	0.6545135
Uncorrected SS	31708.2043	Corrected SS	31708.2043
Coeff Variation	.	Std Error Mean	3.59751639

Basic Statistical Measures

Location		Variability	
Mean	0.000000	Std Deviation	25.43828
Median	2.169153	Variance	647.10621
Mode	.	Range	114.50838
		Interquartile Range	24.92806

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	1	Pr >= |M|	0.8877
Signed Rank	S	57.5	Pr >= |S|	0.5840

Tests for Normality

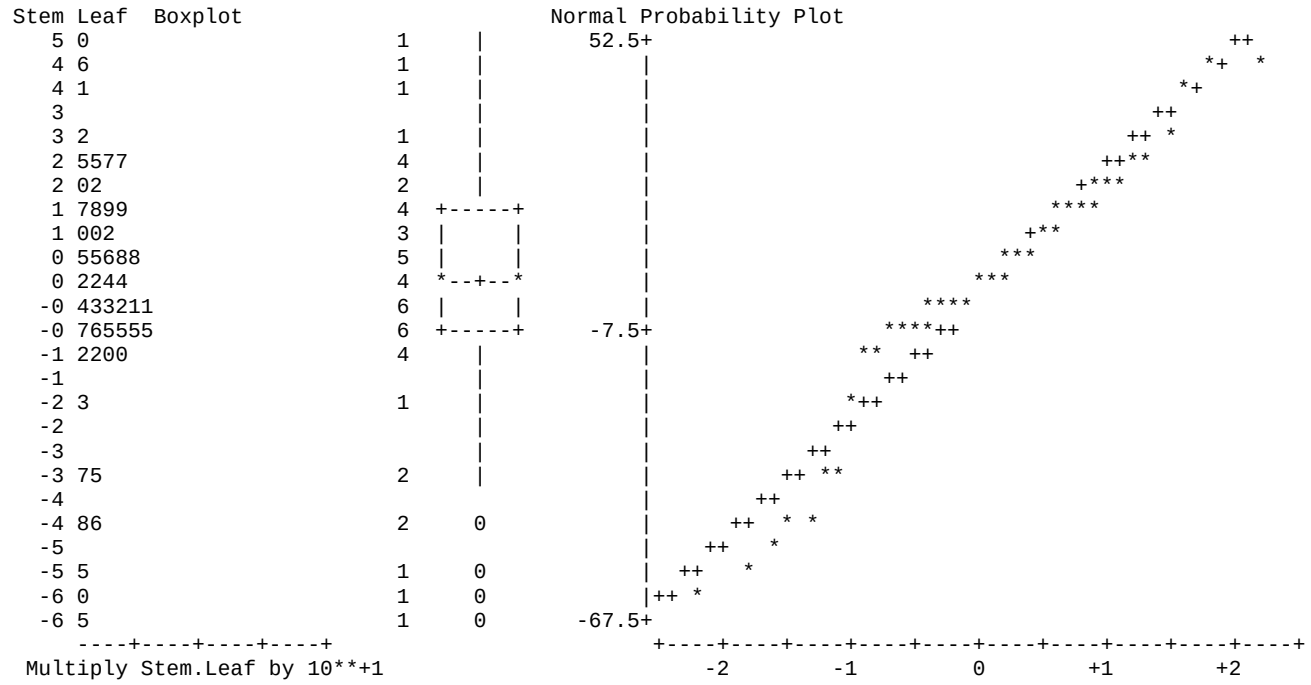
Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.941722	Pr < W	0.0158
Kolmogorov-Smirnov	D	0.155271	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.193852	Pr > W-Sq	0.0061
Anderson-Darling	A-Sq	1.131035	Pr > A-Sq	0.0054

Quantiles (Definition 5)

Quantile	Estimate
100% Max	49.70598
99%	49.70598
95%	41.05749
90%	26.67927
75% Q3	17.74226
50% Median	2.16915
25% Q1	-7.18581
10%	-41.32560
5%	-54.75894
1%	-64.80239
0% Min	-64.80239

Extreme Observations

------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
-64.8024	50	26.7573	46
-59.5187	16	32.0328	2
-54.7589	25	41.0575	28
-48.4172	29	45.9046	13
-45.7444	48	49.7060	21



Selection options: FORWARD, BACKWARD, STEPWISE, NONE (full model)

These do the selection techniques discussed. Additional options working with these are;

SLENTRY = alpha value, level of entry for stepwise and forward

SLSTAY = alpha value, level of remaining for backward and stepwise

INCLUDE = n, automatically puts the first n variables of the list into the regression (do not remove)

RSQUARE, ADJRSQ, CP : These are like all possible regressions. These will simply calculate the R^2, AdjR^2 and C_p statistic for all models to find the best ones. Other criteria include Akaike's information criterion (AIC), Schwartz's Bayesian information criterion (SBC), SSE, MSE and others.

Additional options

START = n

STOP = n (will stop all methods)

BEST = n

eg. with these options we could determine the best 5 models at each level, starting with 3 factors, ending with 6 factors

```
65          PROC REG DATA=StateSAT; Title2 'RSquare procedure';
66             MODEL SAT = TAKERS INCOME YEARS PUBLIC EXPEND RANK /
67                   selection=rsquare start=1 stop=4 best=5 cp adjrsq aic sbc mse;
68          RUN;
69
NOTE: The PROCEDURE REG printed page 22.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.12 seconds
      cpu time            0.06 seconds
```

Chapter 11 : State SAT scores for 1982
RSquare procedure

The REG Procedure
Model: MODEL1
Dependent Variable: SAT

R-Square Selection Method

Number of Observations Read 50
Number of Observations Used 50

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	BIC	MSE	SBC	SSE	Variables in Model
2	0.8471	0.8405	10.2152	337.1712	338.6926	800.54988	342.90724	37626	YEARS RANK
2	0.8054	0.7971	24.9737	349.2094	349.4076	1018.47722	354.94549	47868	EXPEND RANK
2	0.7959	0.7872	28.3393	351.5915	351.5479	1068.17341	357.32757	50204	TAKERS YEARS
2	0.7929	0.7841	29.3971	352.3174	352.2015	1083.79357	358.05344	50938	INCOME RANK
2	0.7899	0.7809	30.4827	353.0515	352.8632	1099.82344	358.78755	51692	PUBLIC RANK
2	0.7814	0.7721	33.4883	355.0295	354.6494	1144.20467	360.76555	53778	TAKERS RANK
3	0.8711	0.8627	3.6884	330.6154	333.3525	689.30879	338.26348	31708	YEARS EXPEND RANK
3	0.8584	0.8491	8.2059	335.3298	337.2970	757.46541	342.97791	34843	INCOME YEARS RANK
3	0.8502	0.8405	11.0931	338.1255	339.6526	801.02397	345.77356	36847	TAKERS YEARS RANK
3	0.8472	0.8372	12.1618	339.1220	340.4954	817.14864	346.77006	37589	YEARS PUBLIC RANK
3	0.8411	0.8308	14.3141	341.0704	342.1482	849.61972	348.71845	39083	PUBLIC EXPEND RANK
3	0.8235	0.8120	20.5602	346.3296	346.6428	943.85568	353.97768	43417	TAKERS YEARS EXPEND
4	0.8771	0.8661	3.5812	330.2544	333.6839	672.12769	339.81451	30246	YEARS PUBLIC EXPEND RANK
4	0.8767	0.8657	3.7169	330.4098	333.8058	674.22054	339.96995	30340	TAKERS YEARS EXPEND RANK
4	0.8746	0.8634	4.4659	331.2592	334.4728	685.77191	340.81934	30860	INCOME YEARS EXPEND RANK
4	0.8593	0.8468	9.8660	336.9901	339.0170	769.05519	346.55023	34607	INCOME YEARS PUBLIC RANK
4	0.8584	0.8458	10.2000	337.3239	339.2840	774.20684	346.88404	34839	TAKERS INCOME YEARS RANK
4	0.8506	0.8373	12.9515	339.9919	341.4283	816.64093	349.55206	36749	TAKERS YEARS PUBLIC RANK
5	0.8787	0.8649	5.0029	331.5865	335.4861	678.28238	343.05865	29844	TAKERS YEARS PUBLIC EXPEND RANK
5	0.8773	0.8634	5.4789	332.1370	335.8918	685.79087	343.60911	30175	INCOME YEARS PUBLIC EXPEND RANK
5	0.8769	0.8629	5.6424	332.3246	336.0303	688.36934	343.79675	30288	TAKERS INCOME YEARS EXPEND RANK
5	0.8594	0.8434	11.8369	338.9609	341.0002	786.07500	350.43308	34587	TAKERS INCOME YEARS PUBLIC RANK
5	0.8425	0.8246	17.8208	344.6305	345.3574	880.45875	356.10262	38740	TAKERS INCOME PUBLIC EXPEND RANK
5	0.8331	0.8141	21.1712	347.5449	347.6369	933.30439	359.01704	41065	TAKERS INCOME YEARS PUBLIC EXPEND

```
71          proc print data=parm1;
72          *   var ADJRSQ AIC BIC CP MSE RSQUARE SBC SSE;
73          run;
NOTE: There were 24 observations read from the data set WORK.PARM1.
NOTE: The PROCEDURE PRINT printed page 23.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.08 seconds
      cpu time            0.03 seconds
```

Chapter 11 : State SAT scores for 1982
RSquare procedure

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	TAKERS	INCOME	YEARS	PUBLIC	EXPEND	RANK
1	MODEL1	PARMS	SAT	28.2940	-243.931	.	.	27.3829	.	.	9.3516
2	MODEL1	PARMS	SAT	31.9136	92.764	.	.	.	.	2.11163	10.0850
3	MODEL1	PARMS	SAT	32.6829	613.897	-2.67946	.	24.9425	.	.	.
4	MODEL1	PARMS	SAT	32.9210	181.173	.	0.26206	.	.	.	8.6229
5	MODEL1	PARMS	SAT	33.1636	251.200	.	.	.	-0.90254	.	9.6268
6	MODEL1	PARMS	SAT	33.8261	412.855	-0.81700	.	.	.	.	6.9574
7	MODEL1	PARMS	SAT	26.2547	-303.724	.	.	26.0952	.	1.86087	9.8258
8	MODEL1	PARMS	SAT	27.5221	-226.136	.	0.20516	26.1300	.	.	8.6290
9	MODEL1	PARMS	SAT	28.3024	-80.997	-0.54627	.	26.7726	.	.	7.6174
10	MODEL1	PARMS	SAT	28.5858	-227.549	.	.	26.8031	-0.09765	.	9.3634
11	MODEL1	PARMS	SAT	29.1482	168.904	.	.	.	-1.43422	2.84704	10.3780
12	MODEL1	PARMS	SAT	30.7222	597.687	-2.84330	.	23.3656	.	2.00591	.
13	MODEL1	PARMS	SAT	25.9254	-204.598	.	.	21.8905	-0.66380	2.24164	10.0032
14	MODEL1	PARMS	SAT	25.9658	-89.636	-0.72893	.	25.2092	.	1.96437	7.5381
15	MODEL1	PARMS	SAT	26.1872	-285.461	.	0.11992	25.5321	.	1.61621	9.3411
16	MODEL1	PARMS	SAT	27.7318	-268.779	.	0.23001	27.5637	0.26704	.	8.5091
17	MODEL1	PARMS	SAT	27.8246	-239.955	0.04788	0.21049	26.1509	.	.	8.7622
18	MODEL1	PARMS	SAT	28.5769	-76.931	-0.66476	.	27.7471	0.18643	.	7.2186
19	MODEL1	PARMS	SAT	26.0439	-100.474	-0.46208	.	22.6688	-0.45226	2.18591	8.4964
20	MODEL1	PARMS	SAT	26.1876	-213.754	.	0.04267	22.3541	-0.55897	2.09446	9.8027
21	MODEL1	PARMS	SAT	26.2368	-118.235	-0.61163	0.03842	25.1714	.	1.86933	7.7509
22	MODEL1	PARMS	SAT	28.0370	-240.775	-0.11589	0.22027	27.6961	0.30112	.	8.1714
23	MODEL1	PARMS	SAT	29.6725	256.909	-0.33460	-0.10252	.	-1.51353	3.14508	9.7592
24	MODEL1	PARMS	SAT	30.5500	535.091	-3.01685	-0.11732	26.9268	0.53568	2.02386	.

Obs	SAT	_IN_	_P_	_EDF_	_SSE_	_MSE_	_RSQ_	_ADJRSQ_	_CP_	_AIC_	_BIC_	_SBC_
1	-1	2	3	47	37625.84	800.55	0.84706	0.84055	10.2152	337.171	338.693	342.907
2	-1	2	3	47	47868.43	1018.48	0.80542	0.79714	24.9737	349.209	349.408	354.945
3	-1	2	3	47	50204.15	1068.17	0.79593	0.78724	28.3393	351.591	351.548	357.328
4	-1	2	3	47	50938.30	1083.79	0.79294	0.78413	29.3971	352.317	352.201	358.053
5	-1	2	3	47	51691.70	1099.82	0.78988	0.78094	30.4827	353.051	352.863	358.788
6	-1	2	3	47	53777.62	1144.20	0.78140	0.77210	33.4883	355.029	354.649	360.766
7	-1	3	4	46	31708.20	689.31	0.87111	0.86270	3.6884	330.615	333.353	338.263
8	-1	3	4	46	34843.41	757.47	0.85837	0.84913	8.2059	335.330	337.297	342.978
9	-1	3	4	46	36847.10	801.02	0.85022	0.84045	11.0931	338.125	339.653	345.774
10	-1	3	4	46	37588.84	817.15	0.84721	0.83724	12.1618	339.122	340.495	346.770
11	-1	3	4	46	39082.51	849.62	0.84114	0.83077	14.3141	341.070	342.148	348.718
12	-1	3	4	46	43417.36	943.86	0.82351	0.81200	20.5602	346.330	346.643	353.978
13	-1	4	5	45	30245.75	672.13	0.87706	0.86613	3.5812	330.254	333.684	339.815
14	-1	4	5	45	30339.92	674.22	0.87667	0.86571	3.7169	330.410	333.806	339.970
15	-1	4	5	45	30859.74	685.77	0.87456	0.86341	4.4659	331.259	334.473	340.819
16	-1	4	5	45	34607.48	769.06	0.85933	0.84682	9.8660	336.990	339.017	346.550
17	-1	4	5	45	34839.31	774.21	0.85838	0.84579	10.2000	337.324	339.284	346.884
18	-1	4	5	45	36748.84	816.64	0.85062	0.83734	12.9515	339.992	341.428	349.552
19	-1	5	6	44	29844.42	678.28	0.87869	0.86490	5.0029	331.587	335.486	343.059
20	-1	5	6	44	30174.80	685.79	0.87734	0.86341	5.4789	332.137	335.892	343.609
21	-1	5	6	44	30288.25	688.37	0.87688	0.86289	5.6424	332.325	336.030	343.797
22	-1	5	6	44	34587.30	786.07	0.85941	0.84343	11.8369	338.961	341.000	350.433
23	-1	5	6	44	38740.19	880.46	0.84253	0.82463	17.8208	344.630	345.357	356.103
24	-1	5	6	44	41065.39	933.30	0.83307	0.81411	21.1712	347.545	347.637	359.017

```
74
75          PROC REG DATA=StateSAT; Title2 'Cp criterion';
76             MODEL SAT = TAKERS INCOME YEARS PUBLIC EXPEND RANK /
77                         selection=cp start=2 stop=5 best=10;
78          RUN;
79
80          ods html close;
81          quit;

NOTE: The PROCEDURE REG printed page 24.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.13 seconds
      cpu time            0.03 seconds
```

```
Chapter 11 : State SAT scores for 1982
Cp criterion

The REG Procedure
Model: MODEL1
Dependent Variable: SAT

C(p) Selection Method

Number of Observations Read          50
Number of Observations Used          50
```

Number in Model	C(p)	R-Square	Variables in Model
4	3.5812	0.8771	YEARS PUBLIC EXPEND RANK
3	3.6884	0.8711	YEARS EXPEND RANK
4	3.7169	0.8767	TAKERS YEARS EXPEND RANK
4	4.4659	0.8746	INCOME YEARS EXPEND RANK
5	5.0029	0.8787	TAKERS YEARS PUBLIC EXPEND RANK
5	5.4789	0.8773	INCOME YEARS PUBLIC EXPEND RANK
5	5.6424	0.8769	TAKERS INCOME YEARS EXPEND RANK
3	8.2059	0.8584	INCOME YEARS RANK
4	9.8660	0.8593	INCOME YEARS PUBLIC RANK
4	10.2000	0.8584	TAKERS INCOME YEARS RANK

Display 12.9 Cp plot for state SAT averages (showing only those models with Cp < 10); $t =$ log percentage of takers, $i =$ income, $y =$ years, $p =$ public, $e =$ expend, and $r =$ rank