

Lecture 10

1. Proc Reg regression example: Minnesota school math tests
2. Plots in Proc Reg
3. Variance Inflation Factor (VIF)
4. Subset selection using Mallows's C_p
5. Structuring SAS programs

1

ANOVA table and related statistics are the same as in Proc GLM.

New: **Adjusted R Squared**, which adjusts for the number of predictors in the model, because $R^2 = \text{SS}(\text{Model}) / \text{SS}(\text{Total})$ always increases as predictors are added whether they help or not.

Adjusted R Squared is

$$1 - \frac{(n-1)(1-R^2)}{n-p},$$

where n is the number of observations and p is the number of predictors in the model. What if $p = 1$?

In comparing regression models, many prefer adjusted R^2 , because adding a non-significant predictor may *reduce* adjusted R^2 .

2

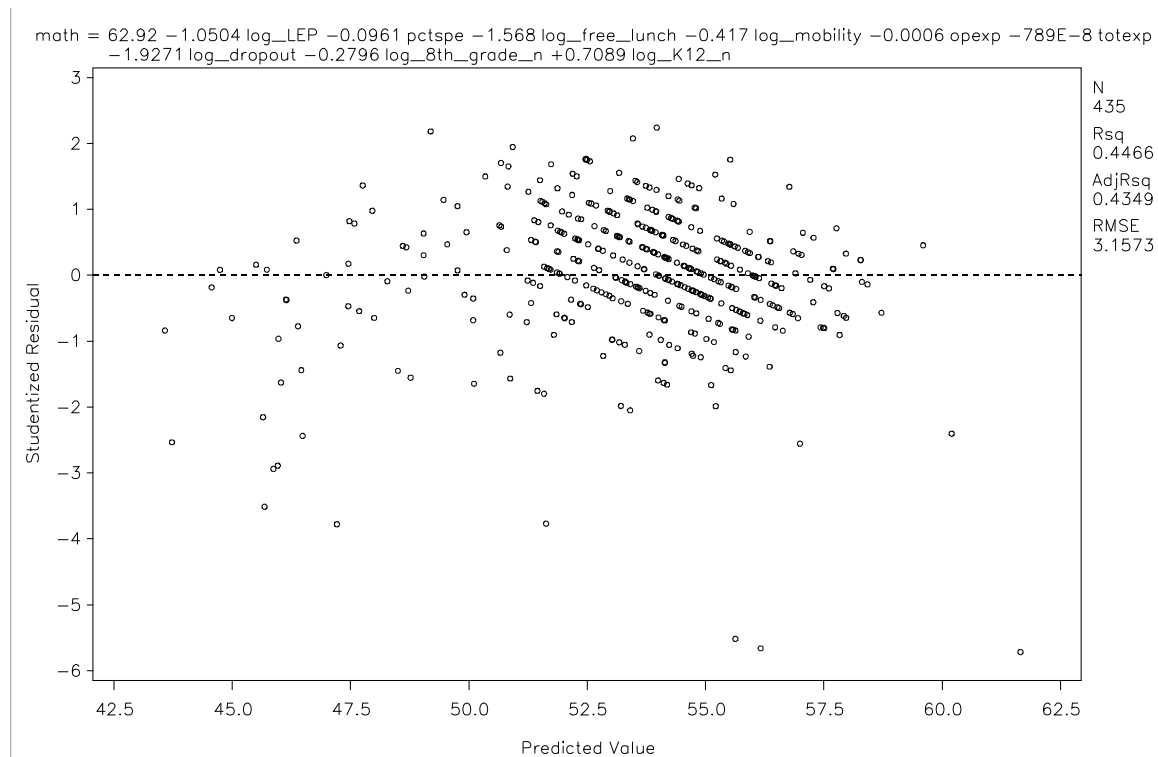
Plots from Proc REG

Proc REG can make several types of plots by itself.

```
Proc REG data = a;
  model math = log_LEP pctspe
            log_free_lunch log_mobility
            opexp totexp log_dropout
            log_8th_grade_n log_K12_n;
  plot student. * predicted. ; requests plot---names include the dot
  plot cookd. * obs. ;

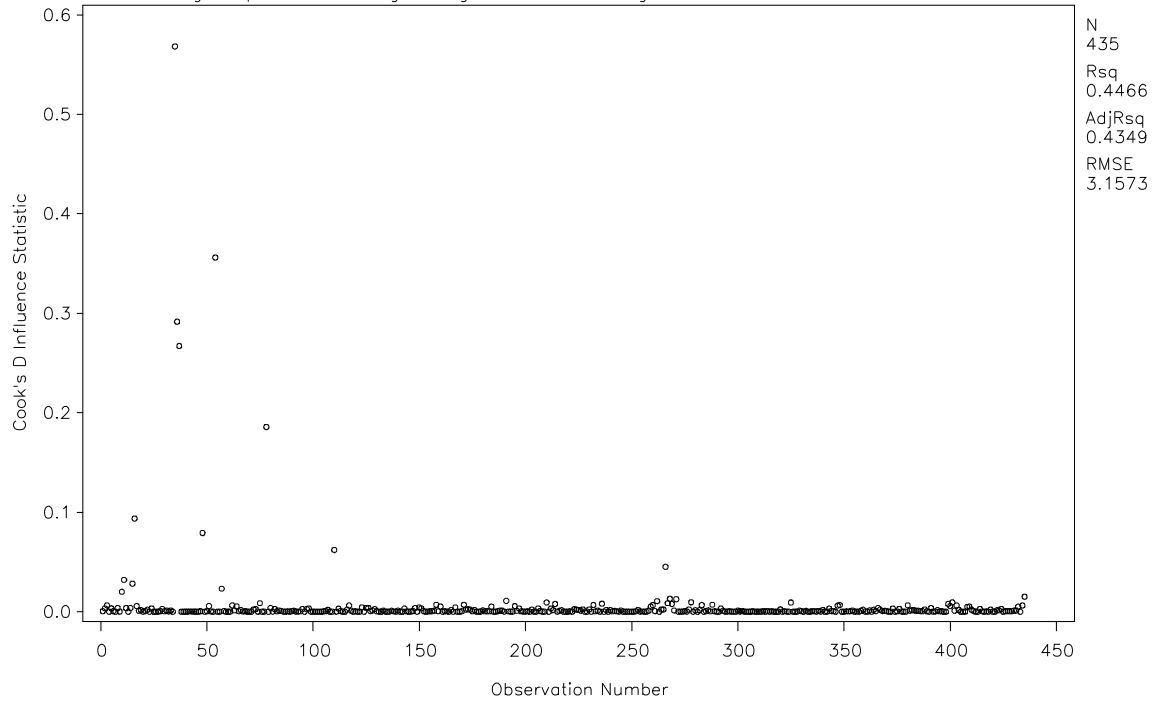
  output out =c predicted=yhat student=resid cookd=cook_dist;
  same output statement as GLM
```

3



4

math = 62.92 -1.0504 log_LEP -0.0961 pctspe -1.568 log_free_lunch -0.417 log_mobility -0.0006 opexp -789E-8 totexp
 -1.9271 log_dropout -0.2796 log_8th_grade_n +0.7089 log_K12_n



5

Observations with the largest Cook's distance:

```
proc print data=c; where (cook_dist > 0.1);
```

Obs	log	free_lunch	mobility	opexp	totexp	dropout	8th_grade_n	K12_n	cook_dist	RMSE		
35	110807	0.56841	45	0	0	0.00000	0.00000	5924	6558	0.00000	3.21888	4.60517
36	110810	0.29164	39	0	0	3.85015	0.00000	5924	6558	0.00000	3.71357	4.82028
37	110811	0.26720	39	0	0	3.52636	0.00000	5924	6558	0.00000	3.55535	4.79579
54	770070	0.35622	41	0	40	3.98898	0.00000	5890	6519	0.00000	3.68888	4.86753
78	1380050	0.18580	36	0	25	3.52636	4.15888	5779	6880	2.70805	3.58352	5.25227

6

Full model, all observations:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	3419.48739	379.94304	38.11	<.0001
Error	425	4236.55859	9.96837		
Corrected Total	434	7656.04598			

Root MSE	3.15727	R-Square	0.4466
Dependent Mean	53.45977	Adj R-Sq	0.4349
Coeff Var	5.90589		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	62.92023	2.68782	23.41	<.0001
log_LEP	1	-1.05040	0.19928	-5.27	<.0001
pctspe	1	-0.09610	0.04269	-2.25	0.0249
log_free_lunch	1	-1.56797	0.31988	-4.90	<.0001
log_mobility	1	-0.41698	0.31963	-1.30	0.1927
opexp	1	-0.00060378	0.00028682	-2.11	0.0359
totexp	1	-0.00000789	0.00020212	-0.04	0.9689
log_dropout	1	-1.92715	0.37330	-5.16	<.0001
log_8th_grade_n	1	-0.27957	0.44705	-0.63	0.5321
log_K12_n	1	0.70893	0.56729	1.25	0.2121

7

Drop the observations with errors and the unusual cases with high influence:

data b;

```
set a; modified data with log variables added
if (tot8enr > 0 and k12enr > 0); data errors
if (pctspe > 0 and pctfre > 0); unusual schools
```

Root MSE	2.70826	R-Square	0.5720
Dependent Mean	53.55269	Adj R-Sq	0.5627
Coeff Var	5.05719		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	72.24178	2.06023	35.06	<.0001
log_LEP	1	-0.77315	0.17390	-4.45	<.0001
pctspe	1	-0.20086	0.03797	-5.29	<.0001
log_free_lunch	1	-1.97327	0.30353	-6.50	<.0001
log_mobility	1	-1.10732	0.28573	-3.88	0.0001
opexp	1	-0.00040949	0.00025054	-1.63	0.1029
totexp	1	-0.00011045	0.00017472	-0.63	0.5276
log_dropout	1	-1.79090	0.32943	-5.44	<.0001
log_8th_grade_n	1	0.03660	0.33448	0.11	0.9129
log_K12_n	1	-0.45787	0.41529	-1.10	0.2709

8

Variance Inflation Factors (VIF)

From SAS Help:

The VIF option in the MODEL statement provides the Variance Inflation Factors (VIF). These factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the regressor (independent) variables. There are no formal criteria for deciding if a VIF is large enough to affect the predicted values.

Large values of VIF are a sign of collinearity, and collinearity must involve at least two predictors.

A reasonable first step is a matrix scatterplot of the predictors with largest VIF.

Consider combining collinear predictors or dropping redundant predictors.

9

```
Proc Reg; model . . . / VIF;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	72.24178	2.06023	35.06	<.0001	0
log_LEP	1	-0.77315	0.17390	-4.45	<.0001	1.63538
pctspe	1	-0.20086	0.03797	-5.29	<.0001	1.24379
log_free_lunch	1	-1.97327	0.30353	-6.50	<.0001	2.08965
log_mobility	1	-1.10732	0.28573	-3.88	0.0001	1.61097
opexp	1	-0.00040949	0.00025054	-1.63	0.1029	4.17954
totexp	1	-0.00011045	0.00017472	-0.63	0.5276	3.15637
log_dropout	1	-1.79090	0.32943	-5.44	<.0001	1.57237
log_8th_grade_n	1	0.03660	0.33448	0.11	0.9129	5.76298
log_K12_n	1	-0.45787	0.41529	-1.10	0.2709	5.14305

Examine the four predictors with VIF > 3.

```
Proc Corr data=a;  
  var totexp opexp k12enr tot8enr;
```

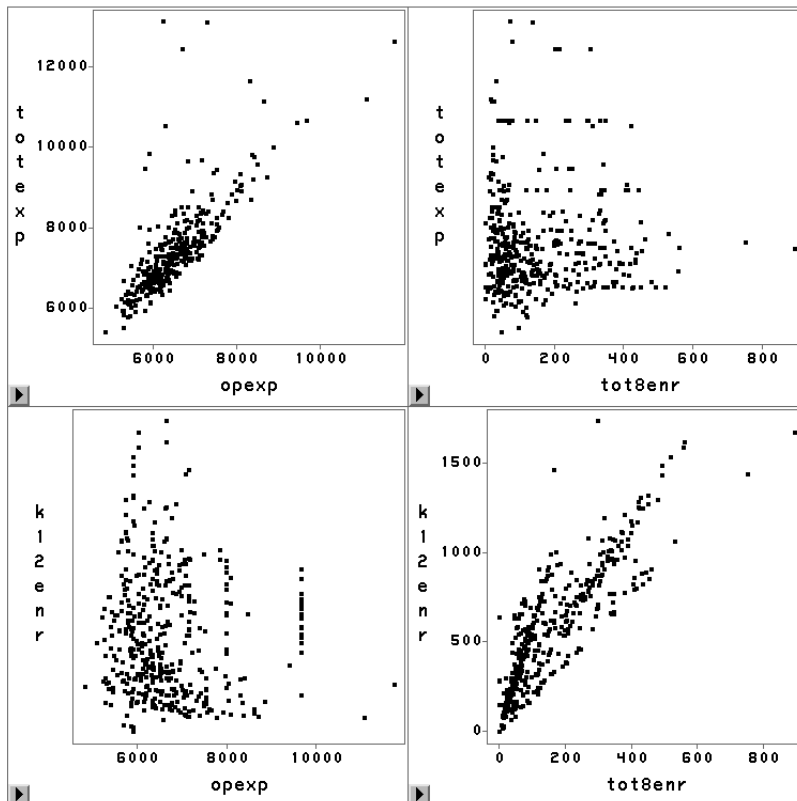
```
Proc Insight data=a;  
  scatter totexp k12enr * opexp tot8enr;
```

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
totexp	427	7724	1334	3298149	5417	13187
opexp	427	6739	1071	2877711	4836	11800
k12enr	427	544.77518	338.94196	232619	30.00000	1754
tot8enr	427	153.58548	136.68560	65581	9.00000	900.00000

Pearson Correlation Coefficients, N = 427
 Prob > |r| under H0: Rho=0

	totexp	opexp	k12enr	tot8enr
totexp	1.00000	0.81383 <.0001	-0.01064 0.8265	-0.02620 0.5892
opexp	0.81383 <.0001	1.00000	-0.07300 0.1321	-0.08579 0.0766
k12enr	-0.01064 0.8265	-0.07300 0.1321	1.00000	0.86465 <.0001
tot8enr	-0.02620 0.5892	-0.08579 0.0766	0.86465 <.0001	1.00000



What should we do about the four predictors with VIF > 3?

13

Repeat the regression with the corrected data and reduced number of predictors:

```
Proc Reg data=b;
  model math = log_LEP pctspe log_free_lunch
              log_mobility opexp log_dropout log_K12_n
              / VIF ;
  plot student. * predicted. ;
```

14

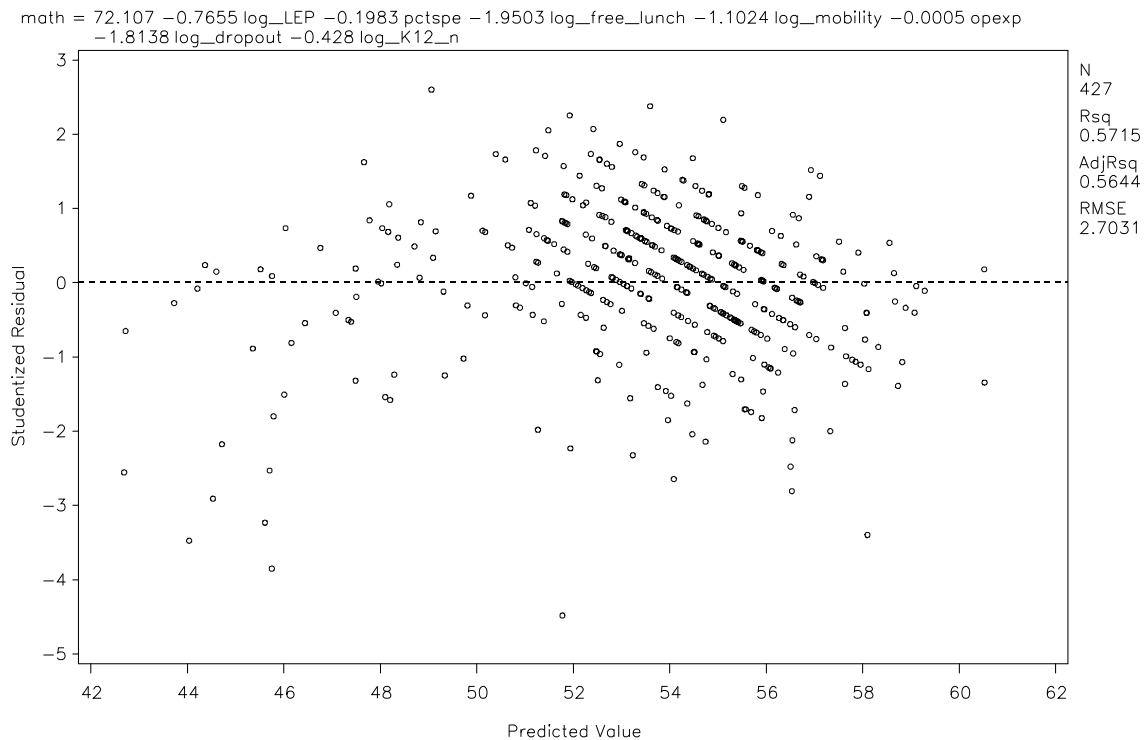
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	4084.03138	583.43305	79.85	<.0001
Error	419	3061.53302	7.30676		
Corrected Total	426	7145.56440			

Root MSE	2.70310	R-Square	0.5715
Dependent Mean	53.55269	Adj R-Sq	0.5644
Coeff Var	5.04756		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	72.10717	1.99935	36.07	<.0001	0
log_LEP	1	-0.76553	0.17217	-4.45	<.0001	1.60912
pctspe	1	-0.19833	0.03675	-5.40	<.0001	1.16938
log_free_lunch	1	-1.95030	0.29539	-6.60	<.0001	1.98662
log_mobility	1	-1.10244	0.28191	-3.91	0.0001	1.57413
opexp	1	-0.00053448	0.00015531	-3.44	0.0006	1.61217
log_dropout	1	-1.81381	0.30601	-5.93	<.0001	1.36190
log_K12_n	1	-0.42803	0.22685	-1.89	0.0599	1.54052

Other improvements to the model?

15



A pattern that disappears when you cover a few points is not really there.

16

Automatic subset selection

Proc Reg is one of several regression procedures that offers automatic selection of a smaller model from a full model.

- **Backwards** Starting from full model, sequentially drop predictors with $p <$ specified cutoff.
Done by hand not SAS, the most common simple procedure among analysts.
- **Forwards** Start with single predictor with highest p -value, sequentially add predictors.
- **Stepwise** Start with Forwards but consider Backwards at each step.
- Maximize a criterion (R^2 , adjusted R^2 , Mallows's C_p):
find models of 1, 2, 3, ..., predictors with largest values of the criterion

17

Subset selection using Mallows's C_p

Mallow's C_p measures how well a subset model predicts the observed data (see chapter 8 in Weisberg, *Applied Linear Regression*), and estimates how well the model will predict new observations.

If there are n observations and K predictors in the full model and a subset model has $p \leq K$ predictors, then

$$C_p = \frac{RSS_p}{\hat{\sigma}_p} + 2p - n = (K - p)(F_p - 1) + p.$$

For the full model, $p = K$ so $C_K = K$.

If the omitted variables really have zero coefficients, then the test statistic for this $F_p \approx 1$, so that $C_p \approx p$.

Good models have $C_p \leq p$.

18

With K predictors, there are 2^K subset models. SAS will check all of them.

```
proc reg data=b;
  model math = log_LEP pctspe log_free_lunch log_mobility
             opexp log_dropout log_K12_n / selection=cp ;
```

The REG Procedure
Dependent Variable: math

C(p) Selection Method

Number of Observations Read 427
Number of Observations Used 427

Number in Model	C(p)	R-Square	Variables in Model
7	8.0000	0.5715	log_LEP pctspe log_free_lunch log_mobility opexp log_dropout log_K12_n
6	9.5600	0.5679	log_LEP pctspe log_free_lunch log_mobility opexp log_dropout
6	17.8435	0.5594	log_LEP pctspe log_free_lunch log_mobility log_dropout log_K12_n
5	18.5088	0.5567	log_LEP pctspe log_free_lunch log_mobility log_dropout
6	21.2930	0.5559	log_LEP pctspe log_free_lunch opexp log_dropout log_K12_n
6	25.7708	0.5513	pctspe log_free_lunch log_mobility opexp log_dropout log_K12_n
5	27.6237	0.5474	log_LEP pctspe log_free_lunch opexp log_dropout
6	35.1321	0.5418	log_LEP log_free_lunch log_mobility opexp log_dropout log_K12_n
5	35.6518	0.5392	log_LEP log_free_lunch log_mobility opexp log_dropout
5	37.1561	0.5376	log_LEP pctspe log_free_lunch log_dropout log_K12_n
5	39.7377	0.5350	pctspe log_free_lunch log_mobility opexp log_dropout
5	40.4631	0.5343	log_LEP pctspe log_free_lunch log_mobility opexp
6	41.1324	0.5356	log_LEP pctspe log_free_lunch log_mobility opexp log_K12_n
4	42.6907	0.5299	log_LEP pctspe log_free_lunch log_dropout
5	42.9047	0.5318	pctspe log_free_lunch opexp log_dropout log_K12_n
4	47.1075	0.5254	log_LEP log_free_lunch log_mobility log_dropout
5	47.4360	0.5271	log_LEP log_free_lunch log_mobility log_dropout log_K12_n
5	48.6207	0.5259	log_LEP pctspe log_mobility opexp log_dropout
5	49.4585	0.5251	pctspe log_free_lunch log_mobility log_dropout log_K12_n
6	49.5918	0.5270	log_LEP pctspe log_mobility opexp log_dropout log_K12_n
5	50.6910	0.5238	log_free_lunch log_mobility opexp log_dropout log_K12_n

19

Any reasonable subset models?

Structuring SAS programs

The primary aim in writing SAS code is to get a job done. But the job is not just to produce some output. You should structure your SAS programs:

- make finding errors as easy as possible
- make the code easy to understand later by yourself or someone else
- document the time and stage of the analysis
- document data edits and corrections

Most projects have an initial data cleaning phase, followed by an analysis to write a paper. After the paper is submitted, there is a delay until the referee reports arrive, followed by another analysis with revisions.

Interruptions of weeks or months in a project are common—write your code to make it easy to pick up later.

21

Advice:

1. List program name, date, author, and revision date(s), and *what the program does* at the top. Use comments to break code into sections.
2. Edit the data and create new variables in a single data step at the beginning of the program, as far as possible. This makes it easy to find and correct problems.
3. Use as few data steps as possible. Data steps are the most confusing part of a program.
4. Use comments to explain data edits and identify data problems. Hard code all data corrections in these programs.

If there is email or other documentation, include this as comments in the code, so you don't need to search your mail files later to figure out what happened.

22

This advice also applies to a collection of programs written for a project:

1. Number the programs within their names as you write them:
PlanB_01.sas, PlanB_02.sas, etc.
2. Do all the data editing and creation of permanent datasets in the first program—no analysis.
Hard code all data corrections in these programs.
Include email correspondence as comments in the code.
3. Perform analysis in later programs that simply call the permanent datasets.
Analysis programs should only create *temporary* datasets.
Multiple versions of the same data invite trouble.

23

When there are changes in the data, all the changes happen in the first (data edit) program.

Then the analysis programs can be run again without modification.

4. Write a “table of contents” file that gives program name and a summary of which datasets it creates or what it does. Keep this up to date.

24