STATISTICS 174: APPLIED STATISTICS

FINAL EXAM

DECEMBER 10, 2002

Time allowed: 3 HOURS.

This is an open book exam: all course notes and the text are allowed, and you are expected to use your own calculator. Answers should preferably be written in a blue book.

The exam is expected to be your own work and no consultation during the exam is allowed. You are allowed to ask the instructor for clarification if you feel the question is ambiguous.

Show all working. In questions requiring a numerical solution, it is more important to demonstrate the method correctly than to obtain correct numerical answers. Even if your calculator has the power to perform high-level operations such as matrix inversion, you are expected to demonstrate the method from first principles. Solutions containing unresolved numerical expressions will be accepted provided the method of numerical calculation is clearly demonstrated.

Questions 1 and 4 are worth 40 points each; questions 2 and 3 are worth 20 points each. A score of 100 may be considered a perfect score. It is not necessary to attempt all the questions but if time allows, it is recommended that you attempt as much as possible.

1. The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

is fitted under the usual assumptions for linear models; in particular, $\{\epsilon_i\}$ are assumed to be independent $N[0, \sigma^2]$ for common unknown σ^2. Assume in addition that $\sum_i x_{ij} = 0$, $\sum_i x_{ij}^2 = n$ for $j = 1, 2, 3$, $\sum_i x_{i1} x_{i2} = 0$, $\sum_i x_{i1} x_{i3} = 0$, $\sum_i x_{i2} x_{i3} = \theta n$, where $-1 < \theta < 1$. Also write $S_0 = \sum_i y_i$, $S_j = \sum_i y_i x_{ij}$ for $j = 1, 2, 3$.

   (a) Write the least squares estimators $\widehat{\beta}_j$, $j = 0, 1, 2, 3$ as explicit algebraic expressions of $S_0, ..., S_3$, n and θ

   (b) Show that the residual sum of squares is given by

$$RSS = \sum_i y_i^2 - \frac{1}{n}\left(S_0^2 + S_1^2 + \frac{S_2^2 - 2\theta S_2 S_3 + S_3^2}{1 - \theta^2}\right). \tag{2}$$

   (c) Write down an explicit test (i.e. expressed as far as possible in terms of the quantities defined in the first two parts of this question) of the hypothesis $H_0:\ \beta_1 = 0$ against the alternative $H_1:\ \beta_1 \neq 0$.

   (d) Write down an explicit test (i.e. expressed as far as possible in terms of the quantities defined in the first two parts of this question) of the hypothesis $H_0:\ \beta_2 = \beta_3 = 0$ against the alternative that β_2 and β_3 are not both 0. (*Hint:* The residual sum of squares under H_1, written RSS_1, is given by (2). Write down the corresponding quantity under H_0, written RSS_0, and hence give a compact expression for the difference $RSS_0 - RSS_1$.)

   (e) Now suppose we are interested in the power of the test in part (d), i.e. the probability that this test rejects the null hypothesis H_0 under some explicit alternative (β_2, β_3) where β_2 and β_3 are not both 0. Show that this power may be calculated from a certain non-central F distribution $F'_{\nu_1,\nu_2;\delta}$, where you should state ν_1 and ν_2 and prove that

$$\delta^2 = \frac{n(\beta_2^2 + 2\theta\beta_2\beta_3 + \beta_3^2)}{\sigma^2}. \tag{3}$$

(f) Use the Pearson-Hartley charts to evaluate this power in the case $n = 16$, $\beta_2 = 1$, $\beta_3 = 2$, $\theta = 0.8$, $\sigma^2 = 5$. Consider both the possibilities $\alpha = 0.05$ and $\alpha = 0.01$ for the size of the test.

2. A statistician is considering the choice between just two regression models of the form

$$\begin{aligned} Y &= X_1\beta_1 + \epsilon, \\ Y &= X_2\beta_2 + \epsilon, \end{aligned}$$

where Y is an $n \times 1$ vector of observations, X_k for $k = 1, 2$ is a $n \times p_k$ design matrix, β_k is a $p_k \times 1$ vector of parameters, and ϵ is a vector of error subject to the usual assumptions of linear models.

(a) If $p_1 = p_2$, then most model selection procedures will simply select the model with the smaller residual sum of squares. Show that this is equivalent to the following: select model 1 if and only if

$$Y^T CY < 0, \tag{4}$$

where you should write down an explicit expression for the matrix C.

(b) Suppose that $p_1 < p_2$, σ^2 is known, and that we choose between models 1 and 2 using one of the criteria (i) AIC, (ii) BIC, (iii) in the case that X_1 is a submatrix of X_2, a hypothesis test in which the null hypothesis is that X_1 is the correct matrix of covariates. Show that under any of these criteria, the selection procedure is to choose model 1 if

$$Y^T CY < B, \tag{5}$$

and find B.

(c) In case (b), what is the expected value of $Y^T CY$ when model 1 is true?

3. Consider a linear model including only one covariate and no intercept:

$$y_i = \beta x_i + \epsilon_i, \;\; i = 1, ..., n, \tag{6}$$

but in which the covariance matrix of $\epsilon = \begin{pmatrix} \epsilon_1 & \dots & \epsilon_n \end{pmatrix}^T$ is of the form $\sigma^2 V$, where σ^2 is unknown and

$$V = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix} \tag{7}$$

where $-1 < \rho < 1$. (Another way to write this is to say $V = (v_{ij})$ where $v_{ij} = \rho^{|i-j|}$. In time series analysis this is known as the autoregressive model of order 1.)

(a) If $n > 2$, show that V^{-1} is of the form κW, where W is a matrix with entries w_{ij} defined by

$$w_{ij} = \begin{cases} 1 & \text{if } i = j = 1 \text{ or } i = j = n, \\ 1 + \rho^2 & \text{if } 1 < i = j < n, \\ -\rho & \text{if } |i - j| = 1, \\ 0 & \text{in all other cases,} \end{cases}$$

and κ is some constant that you have to determine.

(b) Derive the generalized least squares estimator of β and state its variance. (*Note:* If you assume the result of part (a), you can do this part even if you did not successfully complete (a).)

4. Table 2 (end of exam) is based on measurements of fine particles ($PM_{2.5}$) collected at 74 monitoring stations in the states of North Carolina, South Carolina and Georgia, during 1999. The data shown give the annual mean $PM_{2.5}$ (not corrected for missing values) at each monitor, together with a variety of covariates for that monitor, listed in Table 1. In the case of the meteorological covariates, the data are taken from the nearest meteorological station in the "Historical Climatological Network", which is an extensive data base maintained by the National Climatic Data Center. Apart from the latitude-longitude coordinates and meteorological variables, also included are indicator variables for state (NC/SC/GA), and for land use type (agricultural, commercial, forest, industrial and residential).

Name	Explanation
PM	Annual mean $PM_{2.5}$ level at monitor (μg/m^3)
LAT	Latitude of monitor
LON	Longitude of monitor
MAX	Annual mean maximum daily temperature (°F)
MIN	Annual mean minimum daily temperature (°F)
PCP	Total annual precipitation (inches)
N1	=1 if monitor is in North Carolina, 0 otherwise
S1	=1 if monitor is in South Carolina, 0 otherwise
G1	=1 if monitor is in Georgia, 0 otherwise
A1	=1 if monitor location is agricultural, 0 otherwise
C1	=1 if monitor location is commercial, 0 otherwise
F1	=1 if monitor location is in forest, 0 otherwise
I1	=1 if monitor location is industrial, 0 otherwise
R1	=1 if monitor location is residential, 0 otherwise

Table 1. Explanation of variables in Question 4.

(a) An initial regression is performed using one of $y_1 = PM$, $y_2 = \sqrt{PM}$, $y_3 = \log PM$ as the response variable of interest, and covariates $lat, lon, max, min, pcp, n1, s1, a1, c1, f1, i1$. Explain why $g1$ and $r1$ are omitted from this regression, and how one would infer a "Georgia" or "residential" effect in the absence of these covariates.

(b) An initial SAS regression using all of the above covariates resulted in error sum of squares 125.9 using y_1 as the response, 1.818 using y_2 as the response, 0.4291 using y_3 as the response. After taking the scaling of the transformation into account, which of these three models is best? (*Note:* The geometric mean of the PM observations is 16.74.)

(c) Now suppose we select y_2 as the model of interest (not necessarily the answer that you should have obtained for part (b)). A SAS run of the full model and model selection using the RSQUARE criterion produces the following (heavily edited) output:

The SAS System
Dependent Variable: y2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	6.10117	0.55465	18.92	<.0001
Error	62	1.81768	0.02932		
Corrected Total	73	7.91885			

Root MSE	0.17122	R-Square	0.7705
Dependent Mean	4.10447	Adj R-Sq	0.7297
Coeff Var	4.17164		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-8.69251	3.66715	-2.37	0.0209
lat	1	0.17669	0.04060	4.35	<.0001
lon	1	-0.05811	0.02527	-2.30	0.0249
max	1	0.02742	0.01263	2.17	0.0337
min	1	0.01105	0.00949	1.17	0.2484
pcp	1	-0.00513	0.00247	-2.08	0.0419
n1	1	-0.48338	0.12928	-3.74	0.0004
s1	1	-0.58102	0.08120	-7.16	<.0001
a1	1	-0.10472	0.07855	-1.33	0.1874
c1	1	0.01124	0.05080	0.22	0.8257
f1	1	-0.05939	0.10855	-0.55	0.5862
i1	1	0.00394	0.06749	0.06	0.9537

R-Square Selection Method

Number in Model	R-Square	Variables in Model
1	0.4819	lon
2	0.6140	lon s1
3	0.6686	pcp n1 s1
4	0.7317	lat pcp n1 s1
5	0.7411	lat max pcp n1 s1
6	0.7565	lat lon max pcp n1 s1
7	0.7639	lat lon max pcp n1 s1 a1
8	0.7690	lat lon max min pcp n1 s1 a1
9	0.7703	lat lon max min pcp n1 s1 a1 f1
10	0.7704	lat lon max min pcp n1 s1 a1 c1 f1
11	0.7705	lat lon max min pcp n1 s1 a1 c1 f1 i1

For each value of p (the number of regressors in the model), only the best model of order p has been shown. All models include an intercept.

Based on the above tables, which would you conclude is the best model for this set of data? Use forward and backward selection, AIC and BIC to make your choice.

(d) Now consider the model with $lat, pcp, n1, s1$ as covariates (not necessarily the best model you should have found in (c)). This was fitted in S-PLUS, and a range of diagnostics produced. Some of the (edited) output follows:

```
> nreg<-lm(y2~lat+pcp+n1+s1)
> summary(nreg)
Call: lm(formula = y2 ~ lat + pcp + n1 + s1)
Residuals:
     Min      1Q   Median     3Q    Max
 -0.3454 -0.1434 0.005331 0.1225 0.5022
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.8031   1.0178     0.7891   0.4328
        lat   0.1214   0.0302     4.0272   0.0001
        pcp  -0.0091   0.0019    -4.8437   0.0000
         n1  -0.6817   0.0896    -7.6071   0.0000
         s1  -0.6830   0.0604   -11.3060   0.0000
Residual standard error: 0.1755 on 69 degrees of freedom
Multiple R-Squared: 0.7317
F-statistic: 47.05 on 4 and 69 degrees of freedom, the p-value is 0

Correlation of Coefficients:
    (Intercept)     lat     pcp      n1
lat -0.9965
pcp -0.2506      0.1772
 n1  0.8285     -0.8331 -0.3172
 s1  0.3015     -0.3251 -0.0487  0.5166

> nreg1<-lm.influence(nreg)
> nreg1$hat
 0.05193476 0.05141539 0.07954896 0.07837418 0.06202485 0.05038562 0.07572711
 0.05122291 0.05849157 0.13153111 0.09117937 0.05505745 0.05569854 0.05061208
 0.15836989 0.08193201 0.05815188 0.06030476 0.07156212 0.04733088 0.04661720
 0.04795466 0.04957416 0.05647126 0.02942601 0.05437300 0.04495347 0.03872170
 0.04686207 0.03728857 0.05322722 0.08356680 0.05288008 0.03323776 0.05739777
 0.04923961 0.07710166 0.03640419 0.05464795 0.03915526 0.05615600 0.03282301
 0.07772980 0.04079687 0.08084883 0.04103693 0.03522417 0.05938889 0.23135754
 0.06673200 0.05015306 0.04773585 0.02864928 0.06583044 0.04019621 0.05177712
 0.03516634 0.07948343 0.13022523 0.09147231 0.09217416 0.10051156 0.08755229
 0.06811757 0.17500840 0.09561029 0.06920297 0.06561165 0.06605742 0.09833782
 0.06648969 0.06558622 0.09326597 0.10373491
> studres(nreg)
 0.248228  -0.414534   0.5986053 -0.8554859 -0.1517634  0.5499154  1.183984
 0.6588629  0.9578507 -0.1787045  0.8796571  0.2084996  1.744446  -1.042982
```

```
-0.3594699 -1.446977  -0.8958382  0.3578466 -1.305811  -0.4887948  0.0248971
-1.174425   0.8768566  1.092169  -0.8408888 -0.4392564 -1.33553    1.038401
-0.4303417  0.7258998  0.0934068 -1.040353  -0.6640207 -1.487854  -0.3507473
-0.9494389 -0.4214205  0.7373895  0.9929218 -1.942413  -1.162163   0.1524531
 0.9146337  0.5640672  0.1615048  3.099669  -0.1497043  2.224197   0.7584176
-1.05061    0.7116178 -1.894143  -0.177575   1.02948   -1.195616   0.585015
 0.1614823  0.7258822 -0.3393788 -2.116215  -0.1446437 -0.8431734  1.209684
-0.3734964  1.275785   2.098011   0.1511453 -0.07136735 0.7283316 -1.502677
 0.3921604  0.5221203  0.0378197 -1.011913
> dffits(nreg)
 0.0580979 -0.09650919 0.1759776 -0.2494721 -0.03902601 0.1266705  0.3388999
 0.1530894  0.2387439 -0.06954608 0.2786266  0.0503281  0.4236665 -0.2408142
-0.1559331 -0.432266  -0.2225979  0.0906524 -0.3625313 -0.1089501  0.005505392
-0.2635797  0.2002614  0.2671937 -0.1464165 -0.1053295 -0.2897498  0.2084096
-0.0954215  0.1428619  0.0221474 -0.3141575 -0.1569008 -0.2758775 -0.08655209
-0.2160672 -0.1218065  0.1433261  0.238729  -0.3921119 -0.2834752  0.02808487
 0.265529   0.1163293  0.04789927 0.6412123 -0.02860496 0.5588828  0.4160908
-0.2809347  0.1635192 -0.4240884 -0.0304965  0.2732867 -0.2446769  0.1367038
 0.0308292  0.2132989 -0.1313194 -0.6714838 -0.0460896 -0.2818559  0.3747154
-0.10098    0.5876006  0.6821538  0.0412125 -0.01891152 0.1937001 -0.4962539
 0.1046602  0.1383271  0.0121294 -0.3442608
> dfbetas(nreg)
numeric matrix: 74 rows, 5 columns.
     (Intercept)          lat          pcp           n1           s1
 1  0.0224972274 -0.019699413 -0.0159997596 -0.003102471 -0.025907835
 2 -0.0361016337  0.031425553  0.0264525976  0.006398387  0.043729128
 3  0.1082858333 -0.107699049  0.0295598103  0.028088666 -0.043340970
 4 -0.1516038218  0.150755113 -0.0427607254 -0.037565100  0.062896230
 5  0.0201411877 -0.021217846 -0.0060080496  0.031410953  0.026667906
 6 -0.0301431773  0.036696192 -0.0222417587 -0.073582316 -0.083464298
 7 -0.1929384105  0.195005491  0.1372316342 -0.285281941 -0.217991037
 8 -0.0491093438  0.055909935 -0.0102024030 -0.100644549 -0.103789014
 9 -0.1071726338  0.117221271 -0.0085758418 -0.176560714 -0.163051354
10 -0.0574554537  0.055159338  0.0235295598 -0.032035884  0.006412387
11 -0.1840663989  0.185753027  0.1098467476 -0.246776320 -0.176218895
12 -0.0199803206  0.022152137 -0.0024512545 -0.035599013 -0.034349028
13 -0.1543024407  0.175352206 -0.0603659746 -0.282908270 -0.284566389
14  0.0590102687 -0.071446747  0.0418632607  0.141107134  0.158917242
15 -0.1337064139  0.132193515  0.0107992658 -0.074561072  0.006394875
16  0.2846265133 -0.295318413 -0.0739985082  0.378391402  0.286536246
17 -0.1186320796  0.111668914  0.0254506765 -0.015436140  0.080329238
18  0.0506034103 -0.047829789 -0.0107401580  0.008815752 -0.031093372
19  0.1937457425 -0.195955605 -0.1477104534  0.298407212  0.233824187
20  0.0249243820 -0.027716828 -0.0187594932  0.068150200  0.072131179
21 -0.0010943901  0.001235963  0.0009241074 -0.003323948 -0.003615691
22 -0.0670511898  0.053705100  0.0685753328  0.047390732  0.135096798
23  0.0639681743 -0.054043524 -0.0537090877 -0.023823407 -0.096430016
24 -0.1370867953  0.127138116  0.1585729458 -0.066832185 -0.040084090
```

25	0.0008691843	-0.002742988	0.0239219644	-0.049740270	0.001115745
26	-0.0219140916	0.016663678	0.0724532122	-0.051321609	-0.004779536
27	0.1737123587	-0.174798571	-0.0247135861	0.068072640	0.056884481
28	-0.0036756831	0.011688994	-0.1023113783	0.067789662	-0.004758044
29	0.0288307543	-0.024843608	-0.0577357872	0.005373922	0.007587200
30	0.0573208923	-0.060090473	0.0228763413	0.085582272	0.019843079
31	-0.0002749601	0.001407983	-0.014520593	0.0067835175	-0.0005932851
32	-0.0512960152	0.069960854	-0.228787906	-0.0768582205	-0.0249577055
33	0.0715031194	-0.065425191	-0.094148742	0.0307868195	0.0205120967
34	0.1032859069	-0.102923963	-0.027664243	0.0027378590	0.0333740789
35	0.0195960468	-0.023924778	0.051345258	-0.0100113403	0.0082880645
36	0.0725894641	-0.081431738	0.097630035	-0.0062536125	0.0275018686
37	-0.0458283735	0.039329718	0.093844606	-0.0725172018	-0.0119888113
38	-0.0658965364	0.066390349	0.008322610	-0.0126811319	-0.0216150918
39	-0.0291591340	0.041303850	-0.149805714	0.0502315407	-0.0148698658
40	-0.0578159535	0.042985623	0.203742147	-0.1810103648	-0.0121747279
41	-0.0178883739	0.032732789	-0.187060893	-0.0604290609	-0.0124113769
42	-0.0008893209	0.001644722	-0.009523812	0.0090208563	-0.0006247769
43	0.1027600413	-0.088662207	-0.204325941	0.1601691360	0.0270931103
44	0.0433325563	-0.039729697	-0.056020218	0.0738954300	0.0124666210
45	0.0207725094	-0.018287973	-0.036602511	0.0305011111	0.0056392487
46	0.2435723325	-0.223790293	-0.308845576	0.4106538811	0.0702837966
47	0.0068778823	-0.007641535	0.008296110	-0.0038502327	0.0025728263
48	0.2094479577	-0.183134079	-0.385308443	0.3426486814	0.0562986204
49	0.1022123701	-0.130262375	0.338232378	0.0936607918	0.0456642016
50	-0.1019374046	0.115731734	-0.154825233	-0.1265621167	-0.0392334761
51	-0.0652640877	0.058720576	0.098751036	-0.0235375241	-0.0182792862
52	0.2513005570	-0.243626429	-0.154739980	0.1139047491	0.0781790575
53	0.0011895655	-0.001107706	-0.001318487	-0.0089607593	0.0003498165
54	0.1348208696	-0.146195884	0.116364491	0.1564777837	0.0488339479
55	-0.0565073237	0.047446778	0.129194393	-0.1289509342	-0.0143164522
56	-0.0510375291	0.045213465	0.086323015	-0.0177667163	-0.0139802138
57	-0.0091038429	0.008349804	0.011732171	0.0004462663	-0.0026204306
58	-0.0187085753	0.005904441	0.168954463	0.0080115990	-0.0003785890
59	-0.0932540470	0.091835289	0.039030768	-0.0804922168	-0.0957126806
60	-0.3678712988	0.360636350	0.175052971	-0.3199052656	-0.5192985374
61	-0.017569414	0.018987167	-0.014332831	-0.012719824	-0.033896097
62	-0.170145764	0.167328358	0.074158138	-0.147193342	-0.215394249
63	0.146289775	-0.156269206	0.095848942	0.108558985	0.281849975
64	0.026103082	-0.025098457	-0.018743621	0.023412420	-0.062214473
65	-0.008654952	-0.026796203	0.458178836	-0.058723299	0.267964858
66	-0.373565716	0.383755852	-0.047936858	-0.299408808	0.274698467
67	-0.007686526	0.008441770	-0.008007769	-0.005368978	0.025610548
68	-0.001645828	0.001358291	0.004067160	-0.001801513	-0.013810381
69	-0.018048594	0.021105335	-0.035319627	-0.010744482	0.129605028
70	0.288359653	-0.281669057	-0.150336674	0.252240273	-0.196653583
71	-0.012500761	0.014156706	-0.018527328	-0.008111074	0.068893165
72	-0.007385063	0.009556434	-0.026300257	-0.003060456	0.094697420

```
73 -0.006453283  0.006635789 -0.000911399 -0.005162843  0.005034201
74  0.172879171 -0.182361269  0.083525783  0.131644287 -0.133704490
```

Also shown in Fig. 1 is a plot of the diagnostics produced by the "plot(nreg)" command. Based on this output, write a detailed report on the fit of the model to the data, taking into account outliers, influential values, the fit of the normal distribution, etc.

(e) The new EPA standard for $PM_{2.5}$ includes the requirement that the annual mean at each site should be less than 15 μg/m^3. Based on this analysis, what do you conclude about the agreement with that standard?

(f) Ultimately, the EPA would like to save costs by reducing the number of monitors in its network. One criterion that it might well use is to drop a monitor if the $PM_{2.5}$ at that location can be well predicted from the rest of the data available. Suggest ways in which this kind of analysis might be used to help inform that kind of decision. (This might require more regression analyses than the ones given in the above SAS and S-PLUS output, but if so, you should indicate the kinds of analyses you would do and how you would use them.)

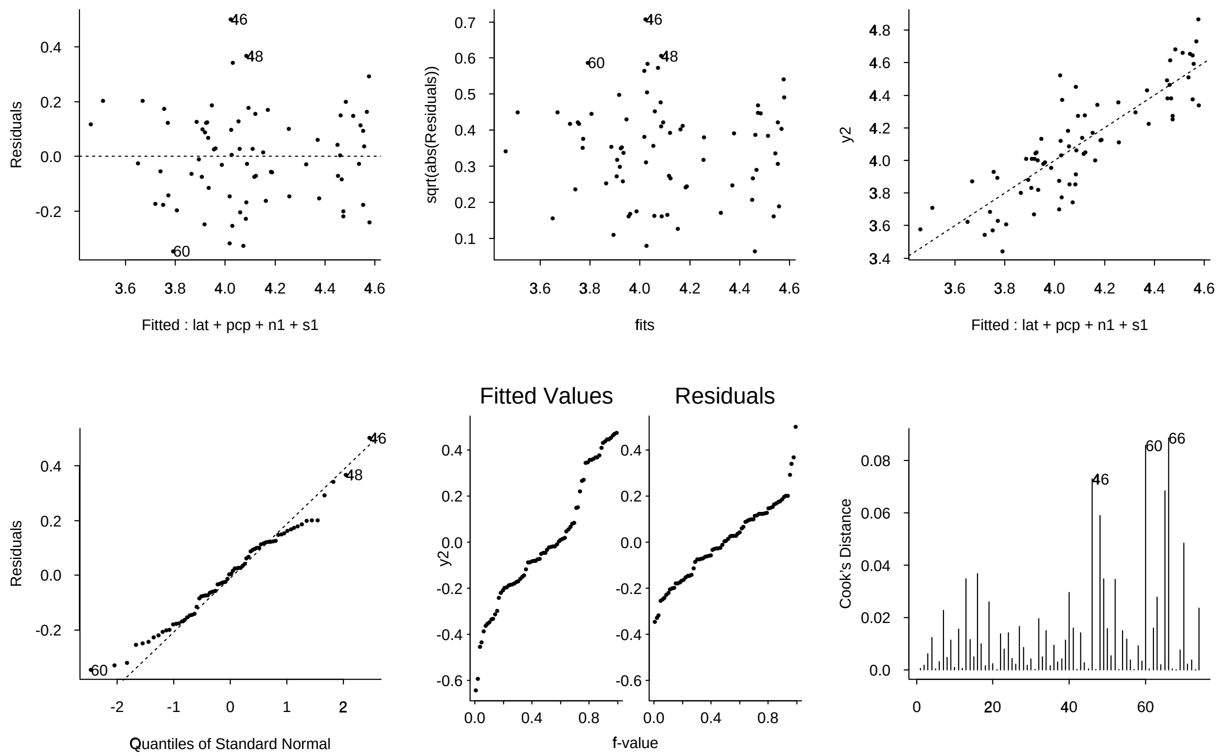

Figure 1. Diagnostic plots produced by S-PLUS "plot(nreg)" command.

Num	PM	LAT	LON	MAX	MIN	PCP	N1	S1	G1	A1	C1	F1	I1	R1
1	20.19	32.78	–83.65	78.50	50.40	36.56	0	0	1	0	0	0	1	0
2	19.20	32.80	–83.54	78.50	50.40	36.56	0	0	1	0	0	0	1	0
3	18.98	32.09	–81.14	78.10	55.10	48.78	0	0	1	0	0	0	0	1
4	16.92	32.11	–81.16	78.10	55.10	48.78	0	0	1	0	0	0	0	1
5	20.35	33.95	–83.37	72.60	50.60	42.66	0	0	1	0	1	0	0	0
6	21.60	33.61	–84.39	73.01	48.07	36.38	0	0	1	0	0	0	1	0
7	21.93	34.01	–84.61	72.00	48.50	49.30	0	0	1	0	1	0	0	0
8	21.69	33.69	–84.29	75.12	50.76	38.42	0	0	1	0	0	0	0	1
9	22.40	33.90	–84.28	75.12	50.76	38.42	0	0	1	0	1	0	0	0
10	18.45	31.58	–84.10	79.60	54.60	34.38	0	0	1	0	0	0	0	1
11	21.73	34.26	–85.27	72.00	48.50	49.30	0	0	1	0	1	0	0	0
12	21.11	33.81	–84.38	75.12	50.76	38.42	0	0	1	0	0	0	0	1
13	23.71	33.80	–84.44	73.01	48.07	36.38	0	0	1	0	1	0	0	0
14	19.15	33.62	–84.44	73.01	48.07	36.38	0	0	1	0	0	0	1	0
15	17.03	31.18	–81.50	78.90	59.90	44.38	0	0	1	0	0	0	1	0
16	18.82	34.30	–83.81	72.60	50.60	42.66	0	0	1	0	0	0	0	1
17	17.84	32.48	–84.98	74.78	49.64	40.69	0	0	1	0	1	0	0	0
18	19.64	32.43	–84.93	74.78	49.64	40.69	0	0	1	0	0	0	1	0
19	18.10	33.93	–85.05	72.00	48.50	49.30	0	0	1	1	0	0	0	0
20	19.21	33.47	–81.99	78.50	50.90	43.94	0	0	1	0	1	0	0	0
21	19.95	33.43	–82.02	78.50	50.90	43.94	0	0	1	0	0	0	0	1
22	18.26	32.97	–82.81	78.50	50.40	36.56	0	0	1	0	0	0	0	1
23	21.28	32.88	–83.33	78.50	50.40	36.56	0	0	1	0	1	0	0	0
24	17.07	36.09	–79.41	71.50	48.20	61.25	1	0	0	0	0	0	0	1
25	15.00	35.61	–82.35	68.10	45.70	46.86	1	0	0	0	1	0	0	0
26	16.33	35.51	–80.62	72.50	48.90	34.81	1	0	0	0	0	0	1	0
27	14.86	36.31	–79.47	70.10	48.20	49.17	1	0	0	1	0	0	0	0
28	18.26	35.73	–81.37	71.10	46.30	40.13	1	0	0	0	0	0	1	0
29	14.68	35.76	–79.16	71.50	48.20	61.25	1	0	0	1	0	0	0	0
30	16.08	35.04	–78.95	73.20	50.50	53.97	1	0	0	0	0	0	0	1
31	17.38	35.81	–80.26	72.50	48.90	34.81	1	0	0	0	1	0	0	0
32	12.56	34.95	–77.96	74.00	53.50	70.96	1	0	0	0	0	0	0	1
33	14.59	35.99	–78.90	71.50	48.20	61.25	1	0	0	0	1	0	0	0
34	14.25	35.95	–77.79	72.65	49.21	50.21	1	0	0	0	0	0	0	1
35	17.05	36.11	–80.23	72.50	48.90	34.81	1	0	0	0	0	0	0	1
36	16.00	36.17	–80.28	69.40	42.30	38.54	1	0	0	0	0	0	0	1
37	16.42	35.25	–81.15	74.20	50.80	30.56	1	0	0	0	0	0	0	1
38	17.49	36.08	–79.79	70.10	48.20	49.17	1	0	0	0	1	0	0	0
39	18.84	35.96	–80.00	72.50	48.90	34.81	1	0	0	0	1	0	0	0
40	14.02	35.54	–82.91	68.50	39.00	40.09	1	0	0	0	1	0	0	0

Table 2, Part 1. Fine particles data set.

Num	PM	LAT	LON	MAX	MIN	PCP	N1	S1	G1	A1	C1	F1	I1	R1
41	13.02	35.23	–77.57	72.50	50.70	65.12	1	0	0	0	1	0	0	0
42	16.69	35.69	–81.99	71.50	39.70	43.52	1	0	0	0	0	0	0	1
43	18.28	35.23	–80.88	74.20	50.80	30.56	1	0	0	0	0	0	0	1
44	16.99	35.25	–80.77	73.60	49.70	41.43	1	0	0	0	0	0	0	1
45	17.12	35.14	–80.85	74.20	50.80	30.56	1	0	0	0	0	0	0	1
46	20.48	35.24	–80.78	73.60	49.70	41.43	1	0	0	0	0	0	0	1
47	16.49	35.92	–82.07	71.50	39.70	43.52	1	0	0	0	0	0	1	0
48	19.84	35.26	–79.84	71.20	47.56	34.81	1	0	0	0	0	1	0	0
49	12.80	34.24	–77.91	73.60	51.60	89.79	1	0	0	0	1	0	0	0
50	12.76	34.77	–77.43	72.50	50.70	65.12	1	0	0	0	0	0	0	1
51	16.36	35.90	–79.06	71.50	48.20	61.25	1	0	0	0	0	0	0	1
52	13.69	36.23	–76.29	70.92	50.70	55.14	1	0	0	0	0	0	0	1
53	15.64	35.59	–77.39	72.65	49.21	50.21	1	0	0	0	0	0	0	1
54	15.45	34.62	–78.99	73.50	51.30	62.46	1	0	0	0	0	0	0	1
55	14.86	35.44	–83.44	68.50	39.00	40.09	1	0	0	0	0	0	0	1
56	16.08	35.86	–78.57	72.40	46.30	62.15	1	0	0	0	0	0	0	1
57	15.91	35.79	–78.62	72.60	48.90	55.61	1	0	0	0	0	0	0	1
58	15.15	35.37	–77.99	74.00	53.50	70.96	1	0	0	0	0	0	0	1
59	13.57	32.43	–80.68	77.42	56.60	34.97	0	1	0	0	1	0	0	0
60	11.87	32.94	–79.66	74.20	61.20	36.17	0	1	0	0	0	1	0	0
61	13.14	32.98	–80.07	77.82	52.45	52.20	0	1	0	0	0	0	0	1
62	13.19	32.79	–79.96	74.20	61.20	36.17	0	1	0	0	0	0	0	1
63	14.99	33.01	–80.97	77.74	49.93	50.37	0	1	0	1	0	0	0	0
64	14.44	34.17	–79.85	76.00	53.30	44.54	0	1	0	0	0	0	0	1
65	13.78	33.37	–79.29	74.92	53.22	72.69	0	1	0	0	0	0	0	1
66	19.12	34.90	–82.31	72.60	51.20	35.93	0	1	0	0	0	0	0	1
67	15.84	34.21	–82.17	73.80	48.90	35.15	0	1	0	0	0	0	1	0
68	15.07	33.78	–81.12	79.30	55.00	36.00	0	1	0	1	0	0	0	0
69	16.41	34.05	–81.15	79.30	55.00	36.00	0	1	0	0	1	0	0	0
70	13.46	34.80	–83.24	75.00	47.50	47.09	0	1	0	0	0	1	0	0
71	15.99	34.09	–80.96	79.30	55.00	36.00	0	1	0	0	0	0	0	1
72	16.07	33.99	–81.02	79.30	55.00	36.00	0	1	0	0	1	0	0	0
73	16.26	34.86	–82.23	72.60	51.20	35.93	0	1	0	0	0	0	0	1
74	15.34	34.94	–81.23	74.20	50.80	30.56	0	1	0	1	0	0	0	0

Table 2, Part 2.

STATISTICS 174: APPLIED STATISTICS

SOLUTIONS TO 2002 FINAL EXAM

1. In the standard notation we have

$$X^TX = \begin{pmatrix} n & 0 & 0 & 0 \\ 0 & n & 0 & 0 \\ 0 & 0 & n & \theta n \\ 0 & 0 & \theta n & n \end{pmatrix}, \quad X^TY = \begin{pmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{pmatrix}. \tag{8}$$

(a) {7 points} Exploiting the block-diagonal form to invert X^TX,

$$(X^TX)^{-1} = \frac{1}{n} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{1-\theta^2} & -\frac{\theta}{1-\theta^2} \\ 0 & 0 & -\frac{\theta}{1-\theta^2} & \frac{1}{1-\theta^2} \end{pmatrix}. \tag{9}$$

Hence from $\widehat{\beta} = (X^TX)^{-1}X^TY$,

$$\widehat{\beta}_0 = \frac{S_0}{n}, \ \widehat{\beta}_1 = \frac{S_1}{n}, \ \widehat{\beta}_2 = \frac{S_2 - \theta S_3}{n(1-\theta^2)}, \ \widehat{\beta}_3 = \frac{S_3 - \theta S_2}{n(1-\theta^2)}. \tag{10}$$

(b) {7} This follows from the sequence of identities (with H as the hat matrix)

$$\begin{aligned} RSS &= Y^T(I-H)Y \\ &= Y^TY - \widehat{Y}^T\widehat{Y} \\ &= Y^TY - \widehat{\beta}^TX^TX\widehat{\beta} \\ &= Y^TY - Y^TX(X^TX)^{-1}X^TY \\ &= Y^TY - \frac{1}{n}\begin{pmatrix} S_0 & S_1 & S_2 & S_3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{1-\theta^2} & -\frac{\theta}{1-\theta^2} \\ 0 & 0 & -\frac{\theta}{1-\theta^2} & \frac{1}{1-\theta^2} \end{pmatrix} \begin{pmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{pmatrix} \end{aligned}$$

which quickly reduces to the form given.

(c) {6} The test is: reject H_0 at size α if

$$\left| \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)} \right| > t_{n-4;1-\alpha/2} \tag{11}$$

where $SE(\widehat{\beta}_1)$ refers to the standard error of $\widehat{\beta}_1$. However, the variance of $\widehat{\beta}_1$ is σ^2/n which is estimated by $RSS/\{n(n-4)\}$, and the square root of this is the standard error. Therefore, (11) reduces to

$$|S_1|\sqrt{\frac{n-4}{n \times RSS}} > t_{n-4;1-\alpha/2}. \tag{12}$$

(d) {7} The corresponding calculation to (2) under H_0 leads to

$$RSS = \sum_i y_i^2 - \frac{1}{n}\left(S_0^2 + S_1^2\right), \tag{13}$$

in other words, RSS_1 is given by (2) and RSS_0 by (13). Therefore

$$RSS_0 - RSS_1 = \frac{S_2^2 - 2\theta S_2 S_3 + S_3^2}{n(1-\theta^2)}. \tag{14}$$

The relevant F statistic is

$$F = \frac{RSS_0 - RSS_1}{2} \cdot \frac{n-4}{RSS_1} \tag{15}$$

which may be calculated from (2) and (14). The F test at size α rejects H_0 if $F > F_{2,n-4;1-\alpha}$.

(e) {7} The alternative hypothesis is of the form $C\beta = h'$ where $C = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and $h' = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}$. Under the null hypothesis, h' is replaced by $h = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. According to page 134 of the text, the noncentrality parameter δ is given by

$$\sigma^2\delta^2 = (h-h')^T\{C(X^TX)^{-1}C^T\}^{-1}(h-h'). \tag{16}$$

However in this case,

$$C(X^TX)^{-1}C^T = \frac{1}{n}\begin{pmatrix} \frac{1}{1-\theta^2} & -\frac{\theta}{1-\theta^2} \\ -\frac{\theta}{1-\theta^2} & \frac{1}{1-\theta^2} \end{pmatrix}$$

and hence

$$\{C(X^TX)^{-1}C^T\}^{-1} = n\begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}$$

Hence

$$\sigma^2\delta^2 = n\begin{pmatrix} \beta_2 & \beta_3 \end{pmatrix}\begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \tag{17}$$

which quickly reduces to (3). The degrees of freedom ν_1 and ν_2 are 2 and $n-4$, as in (d).

(f) {6} With the given numerical values we have $\beta_2^2 + 2\theta\beta_2\beta_3 + \beta_3^2 = 8.2$ and hence

$$\delta^2 = \frac{16 \times 8.2}{5} = 26.24$$

and hence $\phi = \frac{\delta}{\sqrt{1+\nu_1}} = \sqrt{\frac{26.24}{3}} = 2.957$. From the Pearson-Hartley charts with $\nu_1 = 2$, $\nu_2 = 12$, the power is approximately .89 in the case $\alpha = 0.01$ and .984 in the case $\alpha = 0.05$. (More precise values from the S-PLUS "pearsonhartley" function are .8966 and .9852.)

2. (a) {5} If RSS_k denotes the residual sum of squares under model $k = 1, 2$, then $RSS_k = Y^T(I - H_k)Y$ where $H_k = X_k(X_k^T X_k)^{-1} X_k^T$. Then

$$RSS_1 - RSS_2 = Y^T(H_2 - H_1)Y.$$

For this to be negative, condition (4) is satisfied with $C = H_2 - H_1$.

(b) {8} (i) With σ^2, AIC selects model 1 if

$$\frac{SSE_1}{\sigma^2} + 2p_1 < \frac{SSE_2}{\sigma^2} + 2p_2.$$

This is equivalent to

$$Y^T CY = SSE_1 - SSE_2 < 2\sigma^2(p_2 - p_1),$$

so (5) is satisfied with $B = 2\sigma^2(p_2 - p_1)$.
(ii) BIC replaces $2p_k$ with $p_k \log n$ for $k = 1, 2$, so $B = \sigma^2(p_2 - p_1)\log n$.
(iii) With σ^2 known, the most direct test is a χ^2 test: reject H_0 that model 1 is correct with significance level α if

$$\frac{SSE_1 - SSE_2}{\sigma^2} > \chi^2_{p_2 - p_1; 1-\alpha} \tag{18}$$

so (5) is satisfied if $B = \sigma^2 \chi^2_{p_2 - p_1; 1-\alpha}$.
If we used an F test instead of a χ^2 test, the result would be to reject H_0 if

$$\frac{SSE_1 - SSE_2}{p_2 - p_1} \cdot \frac{n - p_2}{SSE_2} > F_{p_2 - p_1, n - p_2; 1-\alpha} \tag{19}$$

which is of form (5) with

$$B = \frac{SSE_2}{n - p_2} \cdot (p_2 - p_1) F_{p_2 - p_1, n - p_2; 1-\alpha}.$$

(c) {7} $\mathrm{E}\{Y^T CY\} = \mathrm{E}\{\mathrm{tr}(Y^T CY)\} = \mathrm{E}\{\mathrm{tr}(CYY^T)\} = \mathrm{tr}(C\mathrm{E}\{YY^T\})$ and

$$\mathrm{E}\{YY^T\} = X_1 \beta_1 \beta_1^T X_1^T + \sigma^2 I_n$$

(I_n is the $n \times n$ identity matrix). Therefore,

$$\mathrm{E}\{Y^T CY\} = \mathrm{tr}(CX_1 \beta_1 \beta_1^T X_1^T) + \sigma^2 \mathrm{tr}(C). \tag{20}$$

However $\mathrm{tr}(H_k) = p_k$ from theory developed in Chapter 3, so in (20), $\mathrm{tr}(C)$ may be replaced by $p_2 - p_1$.
In the case of nested models (X_1 a submatrix of X_2) it follows directly from Theorem 3.1 that $\mathrm{E}(RSS_k) = (n - p_k)\sigma^2$ and therefore that $\mathrm{E}(RSS_1 - RSS_2) = \sigma^2(p_2 - p_1)$. Therefore, in this case, the first term of (20) may be omitted entirely.

3. (a) {10} We have to show

$$\sum_j w_{ij} v_{jk} = \begin{cases} \kappa^{-1} & \text{if } k = i, \\ 0 & \text{if } k \neq i. \end{cases} \tag{21}$$

For $i = 1$,

$$\sum_j w_{ij} v_{jk} = v_{1k} - \rho v_{2k}$$
$$= \begin{cases} 1 - \rho^2 & \text{if } k = 1, \\ \rho^{k-1} - \rho^{k-1} = 0 & \text{if } k > 1, \end{cases}$$

while if $2 < i < n$,

$$\sum_j w_{ij} v_{jk} = -\rho v_{i-1,k} + (1 + \rho^2) v_{i,k} - \rho v_{i+1,k}$$
$$= \begin{cases} -\rho^2 + (1 + \rho^2) - \rho^2 = 1 - \rho^2 & \text{if } k = i, \\ -\rho^{k-i+2} + (1 + \rho^2)\rho^{k-i} - \rho^{k-i} = 0 & \text{if } k > i, \\ -\rho^{i-k} + (1 + \rho^2)\rho^{i-k} - \rho^{i-k+2} = 0 & \text{if } k < i. \end{cases}$$

The case $i = n$ is similar to the case $i = 1$.

Thus for all cases, we have proved (21) with $\kappa = (1 - \rho^2)^{-1}$.

(b) {10} The GLS estimator is $\widehat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$. We may replace V^{-1} by W, since the constant κ cancels from the numerator and denominator. Thus

$$X^T W Y = x_1 y_1 + x_n y_n + (1 + \rho^2) \sum_{i=2}^{n-1} x_i y_i - \rho \sum_{i=1}^{n-1} (x_i y_{i+1} + x_{i+1} y_i)$$

and similarly

$$X^T W X = x_1^2 + x_n^2 + (1 + \rho^2) \sum_{i=2}^{n-1} x_i^2 - 2\rho \sum_{i=1}^{n-1} x_i x_{i+1}.$$

Therefore

$$\widehat{\beta} = \frac{x_1 y_1 + x_n y_n + (1 + \rho^2) \sum_{i=2}^{n-1} x_i y_i - \rho \sum_{i=1}^{n-1} (x_i y_{i+1} + x_{i+1} y_i)}{x_1^2 + x_n^2 + (1 + \rho^2) \sum_{i=2}^{n-1} x_i^2 - 2\rho \sum_{i=1}^{n-1} x_i x_{i+1}}.$$

The variance of $\widehat{\beta}$ is

$$\sigma^2 (X^T V^{-1} X)^{-1} = \sigma^2 \kappa^{-1} (X^T W X)^{-1}$$
$$= \frac{\sigma^2 (1 - \rho^2)}{x_1^2 + x_n^2 + (1 + \rho^2) \sum_{i=2}^{n-1} x_i^2 - 2\rho \sum_{i=1}^{n-1} x_i x_{i+1}}.$$

4. (a) {3} For all rows, $N1 + S1 + G1 = 1$ and $A1 + C1 + F1 + I1 + R1 = 1$ so the $G1$ and $R1$ variables are exactly collinear with some of the others. Therefore, we have to omit some variables to make X of full rank. However, we could still infer an effect for $G1$ from the coefficients for $N1$ and $S1$ and similarly for $R1$ from the coefficients for $A1, C1, F1, I1$

(b) {5} The scaled variable has to be multiplied by C where $C = 1$ for y_1, $C = 2\sqrt{\dot{PM}}$ for y_2 and $C = \log \dot{PM}$ for y_3. Then RSS is multiplied by C^2, i.e. $4\dot{PM} = 66.96$ for y_2 and $(\dot{PM})^2 = 280.2$ for y_3. This makes the rescaled RSS values 125.9, 121.7, 120.2 respectively for y_1, y_2, y_3, i.e. y_3 appears to be the best.

(c) {13} The RSS values are of the form $(1 - R^2) SSTO$ where $SSTO = 7.91885$; therefore, the RSS for the 11 models at the bottom of page 4 are

```
4.079 3.057 2.624 2.125 2.050 1.928 1.870 1.829 1.819 1.818 1.817
```

Ignoring some constants, $AIC = n \log RSS + 2_p$, $BIC = n \log RSS + 2_p$, where $n = 74$ and $p = 2, 3, ..., 12$ for the 11 models, so the AIC and BIC values are

```
108.033 88.682 79.396 65.766 65.127 62.589 62.305 62.689 64.272 66.240 68.207

112.641 95.594 88.613 77.286 78.951 78.717 80.738 83.426 87.312 91.584 95.856
```

The best model is the one with 7 covariates (lat,lon,max,pcp,n1,s1,a1) by AIC, 4 covariates (lat,pcp,n1,s1) by BIC.

Successive F statistics for the model in row i against the model in row $i+1$ are of the form

$$\frac{RSS_i - RSS_{i+1}}{1} \cdot \frac{n-i-2)}{RSS_{i+1}}, \ i = 1, ..., 10$$

which leads to values

```
23.75 11.53 16.23 2.47 4.24 2.07 1.44 0.36 0.03 0.03
```

Note that the model is nested in every case except the test of row 2 against row 3.

Without detailed looking up of tables, we may interpret the values of 4.24 and higher to be significant, but not the smaller values. This means that forward selection would stop after the first 3 tests (i.e. the model with 4 covariates) while backward selection would select the model with 6 covariates.

(d) {13} For this model $p = 5$ (counting the intercept) while $n = 74$. The critical value for h_i is $2p/n = .135$, exceeded for $i = 15, 49, 65$. We have $|studres| > 2$ for $i = 46, 48, 60, 66$; only the value $i = 46$ for which $studres = 3.10$ seems truly an outlier. The critical value for *dffits* is $2\sqrt{p/n} = 0.520$, exceeded in magnitude for $i = 46, 48, 60, 65, 66$ (see also Cook's distance on Fig. 1 which is similar but not identical). Critical value for *dfbetas* is $2/\sqrt{n} = 0.232$ which is exceeded in numerous places, see esp. row 60, value for $s1$. From this we conclude that there are a number of potentially influential values but observations 46, 60 and 66 are most critical. The normality plot shown as part of Fig. 1 seems fine, but note that this is for ordinary residuals and not studentized residuals; however even for the latter, with only one significant outlier, the fit to the normal distribution does not seem bad.

(e) {3} Based on raw data and fitted values, many sites are not in agreement with the standard. Sites are more likely to be out of compliance in Georgia (in the data, all the Georgia sites have mean $PM_{2.5}$ greater than 15), and it also appears that low-rainfall sites are more likely to be out of compliance.

(f) {3} For a proposed reduction of the network, repeat the regression on reduced data set and use to predict $PM_{2.5}$ at the deleted sites. A good network will be one in which the prediction MSE at the deleted sites is small. However, this simple suggestion ignores the effect of direct spatial correlation between the sites. One possible extension of the analysis would be to include values at observed neighboring sites among the covariates of the regression.