STATISTICS 174: APPLIED STATISTICS

FINAL EXAM

DECEMBER 8, 2001

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 1–3 are theoretical questions and each is worth 20 points. Question 4 is worth 60 points. A score of 100 may be considered a perfect score. A table of 95% points for the F distribution is provided.

1. In the world of Scotch whisky, a single malt is a whisky made entirely from one kind of barley at one distillery, while a blended whisky consists of many different types of whisky mixed together (usually mixed with grain whisky as well). In a tasting experiment of blended whiskies, k different single malt whiskies are taken, and blended whiskies are formed by mixing some number $m < k$ single malts in each blend. Assume that in each blend, the different single malts that make up the blend are mixed in equal proportions. Assume that during the course of the experiment, every possible combination of m out of the k single malts is tried.

   [Thus, the total number of blends tried is

   $$n = \binom{k}{m} = \frac{k!}{m!(k-m)!}.$$

   If k and m are not very small, this could be rather a large number of blends. Let's just say the experiment need not be completed in a single sitting.]

   After trying out all n whisky blends, a satisfaction score y_i is assessed for the taste of each blend. A statistical analysis is then performed to determine the desirability of each single malt when used in a blend. A plausible model for such an analysis is

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \epsilon_i,$$

where x_{ij} is 1 if single malt j is a constituent of blended whisky i, and 0 otherwise.

Show how to formulate this problem as a linear model, give algebraic expressions for the least squares estimators $\widehat{\beta}_j$ as functions of the observations y_i, and calculate the variances of the estimates $\widehat{\beta}_j$. Assume the ϵ_i are independent errors with common mean 0 and variance σ^2.

2. A furnace is controlled by opening an air vent to a prescribed aperture x. Allowing for possible feedback effects, the temperature in the furnace is believed to be a quadratic function of x. After measuring the temperature $y_1, ..., y_n$ corresponding to a series of apertures $x_1, ..., x_n$, an attempt is made to determine the aperture which would correspond to a desirable temperature T. The assumed model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

where as usual the ϵ_i are assumed uncorrelated with mean 0 and variance σ^2, and we also assume the x_i values are centered and scaled so that $\sum x_i = 0$, $\sum x_i^2 = A$, $\sum x_i^3 = 0$, $\sum x_i^4 = B$, for known constants A and B.

(a) Show how to formulate this as a linear model and calculate the covariance matrix of the least squares estimates $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$.

(b) Describe how to construct a 95% confidence interval (or, if it doesn't turn out to be an interval, some other kind of confidence set) for the value or values of x that satisfy $\beta_0 + \beta_1 x + \beta_2 x^2 = T$ for a given value of T. You should find that the boundary points of this interval (or set) satisfy an equation of the form $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 = 0$ where $\alpha_0, ..., \alpha_4$ are functions of $n,\ A,\ B$, the least squares estimates $\widehat{\beta}_0, ..., \widehat{\beta}_2$ and the estimated residual standard deviation s; give explicit expressions for $\alpha_0, ..., \alpha_4$ in terms of these quantities.

3. Consider a simple weighing design in which there are four objects, each weighed two at a time. Thus, a suitable model is

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 + \epsilon_1, \\ y_2 &= \beta_1 + \beta_3 + \epsilon_2, \\ y_3 &= \beta_1 + \beta_4 + \epsilon_3, \end{aligned}$$

$$
\begin{aligned}
y_4 &= \beta_2 + \beta_3 + \epsilon_4, \\
y_5 &= \beta_2 + \beta_4 + \epsilon_5, \\
y_6 &= \beta_3 + \beta_4 + \epsilon_6.
\end{aligned}
$$

Once again we make the usual assumptions for $\{\epsilon_i\}$, i.e. uncorrelated, mean 0, common variance σ^2.

(a) Give an explicit formula for the least squares estimate $\widehat{\beta}_1$ as a linear combination of the observations $y_1, ..., y_6$. What is its variance? Note that by the symmetry of the experiment, the variances of $\widehat{\beta}_j,\ j = 2, 3, 4$, will be the same.

Suppose we use a ridge regression estimate, which, for computational simplicity in what follows, we define as $(X^TX + cI)^{-1}X^TY$ where X is derived directly from the above equations without rescaling to $\sum_i x_{ij} = 0,\ \sum_i x_{ij}^2 = n$ as in the usual treatment of ridge regression.

(b) For the ridge regression estimate $(\tilde{\beta}_1^{(c)}, \tilde{\beta}_2^{(c)}, \tilde{\beta}_3^{(c)}, \tilde{\beta}_4^{(c)})$, calculate directly (i) the bias of $\tilde{\beta}_1^{(c)}$, (ii) the variance of $\tilde{\beta}_1^{(c)}$, (iii) the value of c which minimizes the mean squared error. (You don't need to give an explicit expression for c, but state clearly the minimization problem that has to be solved. Of course, here again, if we can solve the problem for β_1 then the same solution will hold by symmetry for $\beta_j,\ j = 2, 3, 4$.)

(c) Suppose the objective were not to estimate the values of β_j with maximum precision, but instead to predict the y_i values in a future experiment. To be precise, assume a future experiment is to be conducted for the same model but with y_i replaced by y_i^* and ϵ_i replaced by an independent ϵ_i^*. Suppose the predictor $\tilde{y}_i^{(c)}$ is formed by summing the relevant $\tilde{\beta}_j^{(c)}$, $\tilde{y}_1^{(c)} = \tilde{\beta}_1^{(c)} + \tilde{\beta}_2^{(c)}$. The symmetry of the experiment implies that the mean squared prediction error of $\tilde{y}_i^{(c)}$ will be the same for each i, so we can take $i = 1$ for definiteness.

Outline how the calculations in (b) would have to be changed if the objective were to choose c minimize the mean squared error of $\tilde{y}_1^{(c)}$ rather than $\tilde{\beta}_1^{(c)}$.

4. Tables 2–4 (Appendix B at the end of this exam) are based on a large study (known as the NMMAPS study) of the health effects of particulate matter based on the 88 largest cities in the continental U.S. In this study, an analysis of the effects of particulate matter on health (similar to the analyses discussed at various points of this course) was conducted separately for each city. Ignoring all the other covariates used in the analysis, the regression coefficient for the effect of particulate matter on mortality

for city i is denoted y_i, and its standard error is denoted s_i. Units are percent increase in deaths corresponding to a 10 μg/m^3 rise in PM_{10}. Thus, for example, for the first city in Table 2 (Los Angeles), we have $y_1 = .38$ and $s_1 = .19$. This means that using the data in Los Angeles, we estimate that a 10 μg/m^3 rise in PM_{10} gives rise to a 0.38% rise in deaths, and the standard error of that estimate is 0.19%.

The purpose of the NMMAPS study was to find out what could be learned by combining these results, possibly using regression methods as part of that process. This differs from examples seen at various points in the course, because here, y_i is used as the input data to a regression model rather than as an end-result in its own right. (It's partly for that reason that the notation is y_i rather than something like $\widehat{\theta}_i$.) Our objective is to treat y_i as given observations and then regress them on the other covariates defined for each city. The hope is that by doing this, we will understand what factors explain why the y_i estimates differ from city to city, and also that the analysis will lead to improved estimates of the overall effect by combining all the y_i. Another issue is geographic variation, e.g. it has been suggested that the effects of particulate matter on health are different in the eastern and western halves of the U.S., and that this may be due to different compositions of atmospheric particulates in different parts of the country.

Tables 2–4 show the name of the city (five-letter abbreviation — for example, the first four are Los Angeles, New York, Chicago and Dallas); region (classified as 1–7 by geography); latitude (°N); longitude (°W); Population in millions; Mean levels of particulate matter (PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and carbon monoxide (CO); the estimate y_i and its standard error s_i.

For the purpose of the analysis, the data were recoded as follows. The "region" variable was converted into seven indicator variables r1–r7; for example, Los Angeles is in Region 3 so r3=1 and r1=r2=r4=r5=r6=r7=0. The latitude and longitude variables were converted to decimal degrees (instead of degrees and minutes, as in Tables 2–4). The other variables were taken directly from the tables. A typical SAS analysis was coded as

```
options ls=77 ps=58;
data nmm1;
infile 'nmm2.txt';
input lon lat y se pop r1-r7 pm o3 no2 so2 co;
wt1=1/se*se;
run;
;
proc reg;
model y=r1-r7 pop pm o3 no2 so2 co /selection=rsquare ;
```

```
weight wt1;
output p=predval r=resid1;
run;
;
```

in which data were read from a file 'nmm2.txt' and variable selection performed on all the variables except latitude and longitude using the 'rsquare' option (which calculates the best model of order p for each p and ranks them using R^2).

Note the use of the 'weight' statement, which weights each observation according to the reciprocal of the variance (so the calculated estimates are actually WLS rather than OLS estimates). However, except for that one statement, the analyses are exactly the same as in a standard linear regression using the OLS estimates, so for the rest of this question you can ignore the distinction between OLS and WLS.

(a) Based on the above variable selection, Table 1 gives the value of the error sum of squares SSE, and the selected variables, for various model orders from 0 to 12. (For model 0, the SSE is the same as $SSTO$, the total sum of squares.) Note that $R^2 = 1 - SSE/SSTO$.

p	R^2	Variables	SSE
0	0		83.0046
1	.0408	r6	79.6180
2	.0599	r3 so2	78.0326
3	.0939	r3 pm so2	75.2096
4	.1057	r3 r7 pm so2	74.2310
5	.1117	r3 r6 r7 pm so2	73.7330
6	.1157	r3 r6 r7 pm so2 co	73.4010
7	.1183	r2 r3 r6 r7 pm so2 co	73.1852
8	.1191	r2 r3 r6 r7 pm o3 so2 co	73.1188
9	.1196	r2 r3 r4 r6 r7 pm o3 so2 co	73.0772
10	.1200	r2 r3 r4 r6 r7 pop pm o3 so2 co	73.0440
11	.1200	r2 r3 r4 r6 r7 pop pm o3 no2 so2 co	73.0440
12	.1200	r1 r2 r4 r5 r6 r7 pop pm o3 no2 so2 co	73.0440

Table 1: Best model of order p for each p

Which of the above models might be considered "best" using (i) F tests (where applicable) to compare the different models in Table 4a, (ii) AIC, (iii) BIC?

(b) For a study of this nature, in which the regressions performed at the level of the individual cities are supposed to take all relevant covariates into account, there is no obvious reason why there should be *any* relationship between the values of y_i and the city-wide covariates. Indeed, all the R^2 values in Table 4a are quite low. How would you decide this point, i.e. whether any of the regressions are "significant"?

(c) Some of the initial press commentary on the results of this study highlighted the fact that the North-East U.S.A. (region 6 in the above analysis) had the highest overall death rates. Comment on this conclusion in the light of the above regression analyses.

We shall now go into more detail about one of the models in Table 4a, for which $p = 3$ and the variables are r3, pm, so2. (This is an obvious candidate to be the best overall model, though it may not be the model you identified as best in part (a) of this question.) Some more SAS code reads

```
proc reg;
model y=r3 pm so2 /collin influence r cli clm vif covb ;
weight wt1;
output p=predval r=resid1;
run;
;
```

which creates all the diagnostics for this model (with the "weight" command again, but for the purpose of the question, you can assume that the interpretation of the diagnostics in a WLS regression is exactly the same as in a OLS regression).

Appendix A at the end of this question gives edited SAS output generated by the above commands.

Now answer the following questions about this SAS output.

(d) Do there appear to be any outliers? If so, give details.

(e) Are there points of high leverage? If so, give details.

(f) Are there influential data points? If so, give details.

(g) Is multicollinearity a problem with this data set? If so, give details.

The final part of this question addresses the overall objectives of the regression exercise.

(h) If the objective is to calculate the overall average effect of particulate matter on health, there would seem to be (at least) two ways to do it: (i) simply average over all the y_i's (presumably a weighted average), (ii) average over the fitted values $\widehat{y}_i$ resulting from the regression (again with suitable weights). What would be the advantages and disadvantages of method (b) as opposed to (a)?

Note: As in earlier parts of the question, you can ignore the fact that this is really a WLS regression: answer the question as if it was being asked for OLS regression.

Appendix A: SAS Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7.79501	2.59834		
Error	84	75.20959	0.89535		
Corrected Total	87	83.00460			

Root MSE	0.94623	R-Square	0.0939
Dependent Mean	0.47239	Adj R-Sq	0.0616
Coeff Var	200.30861		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.75100	0.53214	1.41	0.1619	0
r3	1	1.12099	0.49519	2.26	0.0262	1.76464
pm	1	-0.03076	0.01733	-1.77	0.0796	1.46967
so2	1	0.09382	0.03759	2.50	0.0145	1.24572

Covariance of Estimates

Variable	Intercept	r3	pm	so2
Intercept	0.2831760814	0.0760953991	-0.0080856	-0.006256668
r3	0.0760953991	0.2452169262	-0.004760968	0.007966508
pm	-0.0080856	-0.004760968	0.0003004355	-0.000090439
so2	-0.006256668	0.007966508	-0.000090439	0.001412667

Collinearity Diagnostics

Number	Eigenvalue	Condition Index
1	2.91291	1.00000

2	0.96433	1.73801
3	0.10381	5.29715
4	0.01895	12.39767

	----------------	----------------Proportion	of Variation----------------	
Number	Intercept	r3	pm	so2
1	0.00409	0.00891	0.00410	0.01611
2	0.00031029	0.47585	0.00004872	0.01888
3	0.05720	0.31836	0.05794	0.95647
4	0.93840	0.19688	0.93791	0.00853

Output Statistics

Obs	Weight Variable	Dep Var est	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	1.0000	0.3800	0.6355	0.3608	-0.0821	1.3531
2	1.0000	1.1100	1.0661	0.2669	0.5353	1.5969
3	1.0000	0.3100	0.0877	0.1759	-0.2622	0.4375
4	1.0000	-0.4100	0.1222	0.2274	-0.3299	0.5743
5	1.0000	0.1800	0.0910	0.1711	-0.2492	0.4313
6	1.0000	1.1000	0.9981	0.3950	0.2126	1.7836
7	1.0000	0.6900	0.8437	0.3714	0.1051	1.5823
8	1.0000	0.6500	-0.1601	0.2596	-0.6764	0.3562
9	1.0000	0.4800	0.0935	0.2340	-0.3719	0.5590
10	1.0000	0.7000	0.5141	0.1180	0.2794	0.7488
11	1.0000	0.7700	0.5910	0.1959	0.2014	0.9807
12	1.0000	0.4800	0.1676	0.1730	-0.1764	0.5117
13	1.0000	0.2800	0.5264	0.1212	0.2853	0.7675
14	1.0000	0.3100	0.3696	0.1101	0.1506	0.5885
15	1.0000	-0.0500	0.3302	0.3192	-0.3046	0.9650
16	1.0000	0.2500	0.7997	0.3737	0.0566	1.5428
17	1.0000	0.3900	1.1113	0.3133	0.4883	1.7344
18	1.0000	2.0600	0.4957	0.1138	0.2693	0.7220
19	1.0000	0.0500	0.2036	0.1659	-0.1263	0.5335
20	1.0000	0.6900	0.5725	0.1357	0.3026	0.8425
21	1.0000	0.8500	0.3102	0.3934	-0.4720	1.0925
22	1.0000	0.2000	0.3566	0.1105	0.1370	0.5763
23	1.0000	-0.4500	0.2804	0.1328	0.0163	0.5444
24	1.0000	0.8500	0.8854	0.2147	0.4585	1.3123
25	1.0000	-0.0500	0.8904	0.1919	0.5087	1.2721
26	1.0000	0.9500	0.4126	0.1062	0.2015	0.6238
27	1.0000	0.2000	0.8156	0.2430	0.3324	1.2988

28	1.0000	1.5000	0.5818	0.1391	0.3052	0.8583
29	1.0000	0.4000	0.1796	0.1809	-0.1802	0.5394
30	1.0000	1.9000	0.6124	0.1207	0.3724	0.8524
31	1.0000	0.4000	0.4571	0.1094	0.2394	0.6747
32	1.0000	1.3500	0.4892	0.1272	0.2363	0.7422
33	1.0000	0.6000	0.6398	0.1711	0.2995	0.9800
34	1.0000	0.9500	0.5272	0.1446	0.2396	0.8148
35	1.0000	0.2500	0.1519	0.1498	-0.1460	0.4499
36	1.0000	1.8000	1.0531	0.2336	0.5886	1.5177
37	1.0000	3.2500	0.6968	0.1578	0.3829	1.0107
38	1.0000	0	0.1936	0.2217	-0.2473	0.6345
39	1.0000	0.1000	0.0378	0.1888	-0.3377	0.4133
40	1.0000	0.9000	0.7155	0.3579	0.003798	1.4271
41	1.0000	1.1500	0.5918	0.1337	0.3259	0.8577
42	1.0000	1.9000	0.8895	0.1869	0.5179	1.2612
43	1.0000	1.4000	0.4106	0.1131	0.1857	0.6355
44	1.0000	0.7500	0.9344	0.2145	0.5080	1.3609
45	1.0000	0.0500	0.5356	0.1238	0.2894	0.7819
46	1.0000	1.2500	0.6919	0.1608	0.3721	1.0117
47	1.0000	0.2000	0.3376	0.2494	-0.1583	0.8335
48	1.0000	0.6000	0.4997	0.1065	0.2878	0.7115
49	1.0000	1.1000	0.6556	0.1690	0.3196	0.9916
50	1.0000	-0.6000	0.4618	0.1084	0.2463	0.6773
51	1.0000	0.8500	0.8168	0.1956	0.4279	1.2057
52	1.0000	0.6500	0.5173	0.3976	-0.2735	1.3080
53	1.0000	0.4000	1.1879	0.2763	0.6385	1.7373
54	1.0000	1.8000	0.3603	0.1116	0.1383	0.5823
55	1.0000	-0.4000	0.8428	0.2271	0.3912	1.2944
56	1.0000	0.9000	0.5802	0.1154	0.3508	0.8097
57	1.0000	-0.1500	0.3312	0.1829	-0.0325	0.6949
58	1.0000	-0.3000	0.4126	0.1062	0.2015	0.6238
59	1.0000	0.0500	-0.2890	0.2833	-0.8524	0.2745
60	1.0000	0.8000	0.7847	0.2302	0.3269	1.2426
61	1.0000	1.3000	0.3352	0.1569	0.0232	0.6473
62	1.0000	2.9500	0.5172	0.1188	0.2809	0.7534
63	1.0000	-0.3000	0.5172	0.1188	0.2809	0.7534
64	1.0000	-1.0500	0.4140	0.1278	0.1598	0.6682
65	1.0000	1.9500	0.4957	0.1138	0.2693	0.7220
66	1.0000	0	0.3992	0.1143	0.1719	0.6265
67	1.0000	-1.6500	-0.3532	0.3135	-0.9766	0.2701
68	1.0000	0.5000	0.4486	0.2063	0.0383	0.8588
69	1.0000	0.3000	0.1973	0.1652	-0.1313	0.5259
70	1.0000	-2.7000	0.2014	0.1756	-0.1477	0.5506
71	1.0000	-1.2000	0.2993	0.1309	0.0389	0.5596

72	1.0000	-0.4000	0.1881	0.1693	-0.1485	0.5247
73	1.0000	2.1500	0.2071	0.1883	-0.1673	0.5816
74	1.0000	-0.8500	0.2127	0.1587	-0.1029	0.5283
75	1.0000	0.6500	0.4127	0.1596	0.0953	0.7301
76	1.0000	-0.1500	0.0852	0.2269	-0.3660	0.5363
77	1.0000	-1.3500	0.5799	0.1178	0.3456	0.8143
78	1.0000	-1.7500	0.4926	0.1132	0.2674	0.7177
79	1.0000	0.7500	0.5049	0.1158	0.2745	0.7352
80	1.0000	-0.1000	0.5792	0.1287	0.3232	0.8351
81	1.0000	-0.9000	0.5326	0.1229	0.2881	0.7771
82	1.0000	-0.1000	0.5139	0.1206	0.2740	0.7539
83	1.0000	-1.0000	0.6279	0.1571	0.3154	0.9404
84	1.0000	1.5000	0.6771	0.1786	0.3220	1.0322
85	1.0000	1.2500	0.3080	0.1240	0.0614	0.5547
86	1.0000	0.3000	-0.1428	0.2876	-0.7148	0.4291
87	1.0000	0.9000	0.6064	0.1484	0.3113	0.9015
88	1.0000	1.8000	0.4126	0.1062	0.2015	0.6238

Output Statistics

```
                                      Std Error  Student
Obs     95% CL Predict     Residual   Residual Residual   -2-1 0 1 2

  1   -1.3784    2.6494   -0.2555     0.875    -0.292 |      |      |
  2   -0.8890    3.0212    0.0439     0.908    0.0484 |      |      |
  3   -1.8263    2.0016    0.2223     0.930     0.239 |      |      |
  4   -1.8130    2.0575   -0.5322     0.919    -0.579 |     *|      |
  5   -1.8212    2.0032    0.0890     0.931    0.0956 |      |      |
  6   -1.0410    3.0372    0.1019     0.860     0.119 |      |      |
  7   -1.1777    2.8651   -0.1537     0.870    -0.177 |      |      |
  8   -2.1113    1.7912    0.8101     0.910     0.890 |      |*     |
  9   -1.8449    2.0319    0.3865     0.917     0.422 |      |      |
 10   -1.3822    2.4104    0.1859     0.939     0.198 |      |      |
 11   -1.3306    2.5126    0.1790     0.926     0.193 |      |      |
 12   -1.7453    2.0805    0.3124     0.930     0.336 |      |      |
 13   -1.3707    2.4235   -0.2464     0.938    -0.263 |      |      |
 14   -1.5248    2.2639   -0.0596     0.940   -0.0634 |      |      |
 15   -1.6556    2.3161   -0.3802     0.891    -0.427 |      |      |
 16   -1.2234    2.8228   -0.5497     0.869    -0.632 |     *|      |
 17   -0.8708    3.0935   -0.7213     0.893    -0.808 |     *|      |
 18   -1.3996    2.3909    1.5643     0.939     1.665 |      |***   |
 19   -1.7068    2.1140   -0.1536     0.932    -0.165 |      |      |
 20   -1.3284    2.4735    0.1175     0.936     0.125 |      |      |
 21   -1.7276    2.3481    0.5398     0.861     0.627 |      |*     |
```

22	-1.5378	2.2511	-0.1566	0.940	-0.167		
23	-1.6198	2.1805	-0.7304	0.937	-0.780	*	
24	-1.0441	2.8149	-0.0354	0.922	-0.0384		
25	-1.0296	2.8104	-0.9404	0.927	-1.015	**	
26	-1.4809	2.3061	0.5374	0.940	0.572		*
27	-1.1272	2.7583	-0.6156	0.914	-0.673	*	
28	-1.3201	2.4837	0.9182	0.936	0.981		*
29	-1.7362	2.0954	0.2204	0.929	0.237		
30	-1.2845	2.5093	1.2876	0.939	1.372		**
31	-1.4372	2.3513	-0.0571	0.940	-0.0607		
32	-1.4094	2.3878	0.8608	0.938	0.918		*
33	-1.2724	2.5520	-0.0398	0.931	-0.0427		
34	-1.3763	2.4307	0.4228	0.935	0.452		
35	-1.7532	2.0571	0.0981	0.934	0.105		
36	-0.8850	2.9913	0.7469	0.917	0.815		*
37	-1.2109	2.6045	2.5532	0.933	2.737		*****
38	-1.7391	2.1262	-0.1936	0.920	-0.210		
39	-1.8810	1.9566	0.0622	0.927	0.0671		
40	-1.2963	2.7272	0.1845	0.876	0.211		
41	-1.3086	2.4922	0.5582	0.937	0.596		*
42	-1.0285	2.8076	1.0105	0.928	1.089		**
43	-1.4845	2.3057	0.9894	0.939	1.053		**
44	-0.9950	2.8638	-0.1844	0.922	-0.200		
45	-1.3621	2.4334	-0.4856	0.938	-0.518	*	
46	-1.2168	2.6006	0.5581	0.932	0.599		*
47	-1.6083	2.2835	-0.1376	0.913	-0.151		
48	-1.3939	2.3932	0.1003	0.940	0.107		
49	-1.2559	2.5670	0.4444	0.931	0.477		
50	-1.4322	2.3558	-1.0618	0.940	-1.130	**	
51	-1.1047	2.7382	0.0332	0.926	0.0359		
52	-1.5238	2.5583	0.1327	0.859	0.155		
53	-0.7724	3.1481	-0.7879	0.905	-0.871	*	
54	-1.5344	2.2551	1.4397	0.940	1.532		***
55	-1.0923	2.7779	-1.2428	0.919	-1.353	**	
56	-1.3154	2.4759	0.3198	0.939	0.340		
57	-1.5853	2.2477	-0.4812	0.928	-0.518	*	
58	-1.4809	2.3061	-0.7126	0.940	-0.758	*	
59	-2.2532	1.6753	0.3390	0.903	0.375		
60	-1.1518	2.7213	0.0153	0.918	0.0166		
61	-1.5721	2.2426	0.9648	0.933	1.034		**
62	-1.3793	2.4136	2.4328	0.939	2.592		*****
63	-1.3793	2.4136	-0.8172	0.939	-0.871	*	
64	-1.4848	2.3128	-1.4640	0.938	-1.561	***	
65	-1.3996	2.3909	1.4543	0.939	1.548		***

66	-1.4961	2.2946	-0.3992	0.939	-0.425		
67	-2.3355	1.6290	-1.2968	0.893	-1.452	**	
68	-1.4773	2.3744	0.0514	0.923	0.0557		
69	-1.7128	2.1075	0.1027	0.932	0.110		
70	-1.7124	2.1152	-2.9014	0.930	-3.120	******	
71	-1.6004	2.1989	-1.4993	0.937	-1.600	***	
72	-1.7235	2.0997	-0.5881	0.931	-0.632	*	
73	-1.7115	2.1257	1.9429	0.927	2.095		****
74	-1.6953	2.1207	-1.0627	0.933	-1.139	**	
75	-1.4955	2.3210	0.2373	0.933	0.254		
76	-1.8498	2.0202	-0.2352	0.919	-0.256		
77	-1.3163	2.4761	-1.9299	0.939	-2.056	****	
78	-1.4025	2.3877	-2.2426	0.939	-2.387	****	
79	-1.3909	2.4006	0.2451	0.939	0.261		
80	-1.3199	2.4782	-0.6792	0.937	-0.724	*	
81	-1.3649	2.4301	-1.4326	0.938	-1.527	***	
82	-1.3830	2.4109	-0.6139	0.939	-0.654	*	
83	-1.2796	2.5353	-1.6279	0.933	-1.745	***	
84	-1.2378	2.5920	0.8229	0.929	0.886		*
85	-1.5897	2.2058	0.9420	0.938	1.004		**
86	-2.1095	1.8239	0.4428	0.901	0.491		
87	-1.2983	2.5111	0.2936	0.935	0.314		
88	-1.4809	2.3061	1.3874	0.940	1.476		**

Output Statistics

Obs	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	0.004	-0.2905	0.1454	1.2226	-0.1198
2	0.000	0.0481	0.0796	1.1396	0.0141
3	0.001	0.2378	0.0346	1.0837	0.0450
4	0.005	-0.5771	0.0577	1.0957	-0.1429
5	0.000	0.0950	0.0327	1.0841	0.0175
6	0.001	0.1178	0.1743	1.2696	0.0541
7	0.001	-0.1756	0.1541	1.2383	-0.0749
8	0.016	0.8892	0.0753	1.0923	0.2537
9	0.003	0.4195	0.0612	1.1080	0.1071
10	0.000	0.1969	0.0156	1.0637	0.0247
11	0.000	0.1922	0.0429	1.0941	0.0407
12	0.001	0.3340	0.0334	1.0795	0.0621
13	0.000	-0.2611	0.0164	1.0631	-0.0337
14	0.000	-0.0630	0.0135	1.0633	-0.0074
15	0.006	-0.4248	0.1138	1.1736	-0.1522

16	0.018	−0.6301	0.1560	1.2194	−0.2708
17	0.020	−0.8062	0.1096	1.1421	−0.2829
18	0.010	1.6834	0.0145	0.9307	0.2040
19	0.000	−0.1640	0.0307	1.0809	−0.0292
20	0.000	0.1247	0.0206	1.0703	0.0181
21	0.021	0.6249	0.1728	1.2447	0.2857
22	0.000	−0.1657	0.0136	1.0622	−0.0195
23	0.003	−0.7777	0.0197	1.0395	−0.1102
24	0.000	−0.0382	0.0515	1.1059	−0.0089
25	0.011	−1.0152	0.0411	1.0414	−0.2103
26	0.001	0.5692	0.0126	1.0460	0.0643
27	0.008	−0.6709	0.0659	1.0991	−0.1783
28	0.005	0.9808	0.0216	1.0239	0.1457
29	0.001	0.2360	0.0366	1.0860	0.0460
30	0.008	1.3793	0.0163	0.9740	0.1774
31	0.000	−0.0604	0.0134	1.0631	−0.0070
32	0.004	0.9172	0.0181	1.0261	0.1244
33	0.000	−0.0425	0.0327	1.0844	−0.0078
34	0.001	0.4500	0.0234	1.0637	0.0696
35	0.000	0.1043	0.0251	1.0755	0.0167
36	0.011	0.8129	0.0610	1.0823	0.2071
37	0.054	2.8503	0.0278	0.7427	0.4822
38	0.001	−0.2092	0.0549	1.1077	−0.0504
39	0.000	0.0667	0.0398	1.0924	0.0136
40	0.002	0.2095	0.1430	1.2216	0.0856
41	0.002	0.5936	0.0200	1.0525	0.0847
42	0.012	1.0906	0.0390	1.0313	0.2197
43	0.004	1.0538	0.0143	1.0092	0.1269
44	0.001	−0.1990	0.0514	1.1038	−0.0463
45	0.001	−0.5154	0.0171	1.0538	−0.0680
46	0.003	0.5962	0.0289	1.0620	0.1028
47	0.000	−0.1499	0.0695	1.1262	−0.0409
48	0.000	0.1061	0.0127	1.0620	0.0120
49	0.002	0.4751	0.0319	1.0719	0.0862
50	0.004	−1.1315	0.0131	0.9999	−0.1305
51	0.000	0.0357	0.0427	1.0958	0.0075
52	0.001	0.1537	0.1766	1.2726	0.0712
53	0.018	−0.8693	0.0853	1.1060	−0.2654
54	0.008	1.5448	0.0139	0.9498	0.1835
55	0.028	−1.3598	0.0576	1.0193	−0.3362
56	0.000	0.3387	0.0149	1.0590	0.0416
57	0.003	−0.5161	0.0374	1.0759	−0.1017
58	0.002	−0.7560	0.0126	1.0337	−0.0854
59	0.003	0.3735	0.0897	1.1447	0.1172

60	0.000	0.0165	0.0592	1.1151	0.0041
61	0.008	1.0343	0.0275	1.0249	0.1739
62	0.027	2.6857	0.0158	0.7637	0.3399
63	0.003	-0.8692	0.0158	1.0279	-0.1100
64	0.011	-1.5752	0.0183	0.9498	-0.2148
65	0.009	1.5614	0.0145	0.9481	0.1892
66	0.001	-0.4229	0.0146	1.0555	-0.0515
67	0.065	-1.4623	0.1097	1.0644	-0.5134
68	0.000	0.0554	0.0475	1.1013	0.0124
69	0.000	0.1096	0.0305	1.0814	0.0194
70	0.087	-3.2990	0.0344	0.6637	-0.6229
71	0.012	-1.6151	0.0191	0.9450	-0.2257
72	0.003	-0.6294	0.0320	1.0633	-0.1144
73	0.045	2.1393	0.0396	0.8813	0.4344
74	0.009	-1.1413	0.0281	1.0143	-0.1942
75	0.000	0.2530	0.0285	1.0765	0.0433
76	0.001	-0.2546	0.0575	1.1096	-0.0629
77	0.017	-2.0967	0.0155	0.8668	-0.2632
78	0.021	-2.4577	0.0143	0.8036	-0.2962
79	0.000	0.2596	0.0150	1.0616	0.0320
80	0.002	-0.7224	0.0185	1.0424	-0.0992
81	0.010	-1.5393	0.0169	0.9535	-0.2017
82	0.002	-0.6519	0.0163	1.0448	-0.0838
83	0.022	-1.7665	0.0276	0.9308	-0.2975
84	0.007	0.8844	0.0356	1.0478	0.1700
85	0.004	1.0042	0.0172	1.0171	0.1328
86	0.006	0.4890	0.0924	1.1426	0.1560
87	0.001	0.3125	0.0246	1.0705	0.0496
88	0.007	1.4861	0.0126	0.9565	0.1678

Output Statistics

	DFBETAS			
Obs	Intercept	r3	pm	so2
1	0.0153	-0.0788	-0.0154	-0.0021
2	-0.0030	0.0044	-0.0018	0.0130
3	-0.0161	-0.0293	0.0324	-0.0203
4	-0.0990	0.0361	0.0375	0.1146
5	0.0044	-0.0092	0.0040	-0.0136
6	0.0198	0.0484	-0.0230	0.0039
7	-0.0185	-0.0626	0.0201	-0.0009
8	-0.1186	-0.1734	0.2089	-0.1289
9	-0.0753	-0.0625	0.0956	-0.0111

10	0.0152	0.0003	-0.0109	-0.0014
11	-0.0254	-0.0057	0.0198	0.0250
12	0.0333	-0.0227	-0.0050	-0.0479
13	-0.0218	-0.0015	0.0163	0.0016
14	-0.0001	0.0029	-0.0020	0.0012
15	0.1308	0.0545	-0.1247	-0.0535
16	-0.0759	-0.2206	0.0750	0.0125
17	0.1038	-0.0707	-0.0065	-0.2629
18	0.1132	-0.0077	-0.0742	-0.0143
19	0.0153	0.0171	-0.0225	0.0049
20	0.0135	0.0026	-0.0113	-0.0003
21	-0.0869	0.1128	0.1148	-0.0469
22	-0.0031	0.0075	-0.0032	0.0055
23	0.0359	0.0594	-0.0663	0.0206
24	0.0024	-0.0019	0.0002	-0.0077
25	-0.1110	-0.0914	0.1482	-0.1146
26	0.0134	-0.0183	0.0033	-0.0088
27	0.0935	-0.0120	-0.0471	-0.1456
28	0.1108	0.0237	-0.0943	-0.0013
29	0.0280	-0.0144	-0.0076	-0.0352
30	0.0153	0.0104	-0.0245	0.0861
31	0.0005	0.0018	-0.0015	-0.0010
32	-0.0436	-0.0295	0.0469	0.0449
33	0.0037	0.0004	-0.0025	-0.0053
34	-0.0320	-0.0132	0.0288	0.0337
35	-0.0022	-0.0103	0.0090	-0.0090
36	0.0684	0.0979	-0.1251	0.1522
37	0.3545	0.1467	-0.3564	0.0976
38	-0.0386	0.0092	0.0184	0.0375
39	0.0037	-0.0071	0.0029	-0.0112
40	-0.0016	0.0627	0.0003	0.0030
41	-0.0219	-0.0046	0.0149	0.0476
42	0.0229	0.0759	-0.0800	0.1726
43	-0.0219	-0.0450	0.0447	0.0073
44	0.0061	-0.0135	0.0078	-0.0403
45	-0.0456	-0.0045	0.0351	0.0027
46	-0.0314	0.0060	0.0125	0.0743
47	0.0330	0.0158	-0.0329	-0.0125
48	0.0034	-0.0012	-0.0017	0.0012
49	0.0726	0.0233	-0.0671	0.0023
50	-0.0559	0.0179	0.0269	0.0128
51	-0.0022	0.0013	0.0002	0.0063
52	-0.0302	0.0332	0.0296	0.0054
53	-0.0487	-0.1262	0.1353	-0.2216

54	−0.0061	−0.0766	0.0573	−0.0305
55	0.1429	−0.0426	−0.0512	−0.2836
56	0.0169	0.0032	−0.0149	0.0104
57	−0.0836	0.0096	0.0483	0.0604
58	−0.0178	0.0243	−0.0043	0.0117
59	−0.0333	−0.0802	0.0830	−0.0813
60	0.0037	0.0015	−0.0037	0.0003
61	0.1303	−0.0266	−0.0672	−0.0999
62	0.2111	0.0070	−0.1532	−0.0181
63	−0.0683	−0.0023	0.0496	0.0058
64	−0.1457	0.0256	0.0803	0.0801
65	0.1050	−0.0071	−0.0688	−0.0133
66	−0.0253	0.0113	0.0092	0.0165
67	0.2028	0.3549	−0.4035	0.3200
68	0.0114	0.0009	−0.0083	−0.0054
69	−0.0099	−0.0115	0.0149	−0.0037
70	−0.3887	0.1865	0.1166	0.4642
71	−0.1119	0.0744	0.0202	0.1289
72	0.0600	0.0679	−0.0892	0.0217
73	0.2994	−0.1080	−0.1162	−0.3217
74	0.0938	0.1134	−0.1444	0.0369
75	0.0360	−0.0014	−0.0231	−0.0195
76	−0.0402	0.0188	0.0121	0.0522
77	−0.1302	−0.0256	0.1134	−0.0509
78	−0.1613	0.0137	0.1038	0.0215
79	0.0187	−0.0004	−0.0129	−0.0020
80	−0.0670	−0.0138	0.0572	−0.0059
81	−0.1335	−0.0119	0.1021	0.0084
82	−0.0542	−0.0019	0.0392	0.0067
83	−0.2433	−0.0702	0.2195	−0.0045
84	0.1454	0.0499	−0.1367	0.0059
85	−0.0316	−0.0674	0.0689	−0.0243
86	−0.0990	−0.1020	0.1417	−0.0511
87	0.0394	0.0101	−0.0347	0.0002
88	0.0350	−0.0478	0.0085	−0.0229

Appendix B: Data Tables

City	Reg.	Lat.	Lon.	Pop.	PM_{10}	O_3	NO_2	SO_2	CO	y_i	s_i
Los A	3	34: 3	118:14	8.86	46.0	22.8	39.4	1.9	1.51	.38	.19
New Y	6	40:47	73:58	8.20	28.8	19.6	38.9	12.8	2.04	1.11	.29
Chica	5	41:59	87:54	5.11	35.6	18.6	24.3	4.6	.79	.31	.10
Dalla	7	32:54	97: 2	3.31	23.8	25.3	13.8	1.1	.74	−.41	.63
Houst	7	29:58	95:21	2.82	30.0	20.5	18.8	2.8	.89	.18	.33
San D	3	32:44	117:10	2.50	33.6	31.6	22.9	1.7	1.10	1.10	.47
Santa	3	33:50	117:55	2.41	37.4	23.0	35.1	1.3	1.23	.69	.52
Phoen	2	33:26	112: 1	2.12	40.3	22.5	16.6	3.5	1.27	.65	.54
Detro	5	42:14	83:20	2.11	40.9	22.6	21.3	6.4	.66	.48	.19
Miami	7	25:49	80:17	1.94	25.7	25.9	11.0	5.9	1.06	.70	.73
Phila	6	39:53	75:15	1.59	35.4	20.5	32.2	9.9	1.18	.77	.48
Minne	4	44:53	93:13	1.52	26.9	24.9	17.6	2.6	1.18	.48	.28
Seatt	1	47:27	122:18	1.51	25.3	19.4	22.1	5.9	1.78	.28	.30
San J	1	37:20	121:53	1.50	30.4	17.9	25.1	5.9	.94	.31	.33
Cleve	5	41:25	81:52	1.41	45.1	27.4	25.2	10.3	.85	−.05	.22
San B	3	34: 7	117:19	1.42	37.0	35.9	27.9	.7	1.03	.25	.68
Pitts	5	40:30	80:13	1.34	31.6	20.7	27.6	14.2	1.22	.39	.15
Oakla	1	37:49	122:16	1.28	26.3	17.2	21.2	5.9	.91	2.06	.56
Atlan	7	33:45	84:23	1.19	36.1	25.1	26.0	6.0	.89	.05	.83
San A	2	29:32	98:28	1.19	23.8	22.2	22.1	5.9	1.01	.69	.89
River	3	33:59	117:22	1.17	52.0	33.4	25.0	.4	1.12	.85	.47
Denve	1	39:44	104:59	1.12	29.6	21.4	27.9	5.5	1.03	.20	.25
Sacra	1	38:35	121:29	1.04	33.3	26.7	16.3	5.9	.94	−.45	.52
St Lo	5	38:37	90:12	.99	30.1	22.8	22.5	11.3	1.05	.85	1.23
Buffa	5	42:53	78:53	.97	21.7	22.9	19.0	8.6	.73	−.05	.92
Colum	5	39:58	83: 0	.96	29.0	26.0	22.1	5.9	.76	.95	.57
Cinci	5	39: 6	84:31	.87	34.2	25.8	26.7	11.9	1.00	.20	.40
St Pe	7	27:46	82:39	.85	23.5	24.6	11.8	5.9	.71	1.50	1.00

Table 2: NMMAPS Data, Part 1

City	Reg.	Lat.	Lon.	Pop.	PM_{10}	O_3	NO_2	SO_2	CO	y_i	s_i
Kansa	4	39: 6	94:35	.84	25.9	27.6	9.2	2.4	.62	.40	1.00
Tampa	7	27:57	82:27	.83	28.3	23.5	21.2	7.8	.78	1.90	1.05
Memph	7	35: 8	90: 3	.83	30.3	29.0	26.8	6.8	1.19	.40	1.10
India	5	39:46	86: 9	.80	32.0	31.9	20.2	7.7	.90	1.35	.53
Newar	6	40:44	74:10	.78	32.9	15.2	33.6	9.6	.87	.60	.70
Balti	6	39:17	76:37	.74	32.9	21.2	32.9	8.4	.92	.95	.42
Salt	1	40:45	111:53	.73	32.9	23.0	29.6	4.4	1.35	.25	.18
Roche	6	43:10	77:37	.71	21.9	22.7	22.1	10.4	.63	1.80	1.20
Worce	6	42:16	71:48	.71	22.2	30.0	25.2	6.7	.89	3.25	1.13
Orlan	7	28:33	81:23	.68	22.7	24.1	11.4	1.5	.93	.00	1.75
Jacks	7	30:20	81:39	.67	29.9	28.2	14.8	2.2	.92	.10	1.05
Fresn	3	36:44	119:47	.67	43.4	29.4	21.7	1.9	.68	.90	.50
Louis	5	38:15	85:46	.66	30.8	19.8	22.4	8.4	1.12	1.15	.97
Bosto	6	42:22	71:94	.66	26.0	17.9	29.9	10.0	1.13	1.90	.95
Birmi	7	33:31	86:48	.65	31.2	22.4	22.1	6.6	1.05	1.40	.70
Washi	6	38:54	77: 6	.61	28.2	17.5	25.6	11.2	1.23	.75	1.02
Oklah	2	35:30	97:30	.60	25.0	28.4	13.9	5.9	.71	.05	1.02
Provi	6	41:49	71:24	.60	30.9	25.4	21.9	9.5	1.00	1.25	.88
El Pa	2	31:45	106:29	.59	41.2	24.4	23.6	9.1	1.25	.20	.30
Tacom	1	47:14	122:26	.59	28.0	23.8	22.1	6.5	1.66	.60	.85
Austi	2	30:17	97:45	.58	21.1	25.5	22.1	5.9	1.02	1.10	1.45
Dayto	5	39:45	84:12	.57	27.4	26.6	22.1	5.9	.83	−.60	1.20
Jerse	6	40:44	74: 4	.55	30.5	19.7	28.7	10.7	2.01	.85	.57
Baker	3	35:23	119: 1	.54	53.2	33.3	19.4	3.0	1.05	.65	.48
Akron	5	41: 5	81:31	.51	22.4	30.5	22.1	12.0	.70	.40	.80
Charl	7	35:13	80:51	.51	30.7	29.3	16.2	5.9	1.11	1.80	1.30
Nashv	7	36:10	86:47	.51	32.4	16.2	22.1	11.6	1.12	−.40	.60
Tulsa	7	36:10	95:55	.50	26.6	31.4	16.6	6.9	.65	.90	1.15
Grand	5	42:58	85:40	.50	22.8	27.7	22.1	3.0	.57	−.15	1.08
New O	7	29:58	90: 4	.50	29.0	20.5	21.3	5.9	.94	−.30	.95

Table 3: NMMAPS Data, Part 2

City	Reg.	Lat.	Lon.	Pop.	PM_{10}	O_3	NO_2	SO_2	CO	y_i	s_i
Stock	1	37:58	121:17	.48	39.0	22.6	24.2	1.7	.82	.05	.67
Albuq	2	35: 5	106:39	.48	16.9	25.8	22.1	5.9	.79	.80	1.35
Syrac	6	43: 3	76: 9	.47	24.5	23.7	22.1	3.6	1.17	1.30	1.15
Toled	5	41:39	83:33	.46	25.6	27.1	22.1	5.9	1.03	2.95	1.27
Ralei	7	35:46	78:38	.42	25.6	35.4	12.7	5.9	1.61	–.30	2.05
Wichi	4	37:42	97:20	.40	25.6	24.2	22.1	4.8	.65	–1.05	1.73
Color	1	38:50	104:49	.40	26.3	24.3	22.1	5.9	1.09	1.95	1.77
Baton	7	30:27	91:11	.38	27.3	21.2	16.6	5.2	.43	.00	1.75
Modes	1	37:39	121: 0	.37	41.7	26.1	24.2	1.9	.91	–1.65	1.02
Madis	5	43: 4	89:24	.37	19.9	29.7	22.1	3.3	1.04	.50	2.25
Spoka	1	47:40	117:24	.36	36.0	32.6	22.1	5.9	2.19	.30	.25
Littl	7	34:45	92:17	.35	25.8	27.7	9.3	2.6	1.02	–2.70	1.40
Green	7	36: 4	79:48	.35	27.5	24.9	22.1	4.2	1.22	–1.20	1.60
Knoxv	7	35:58	83:55	.34	36.3	29.6	22.1	5.9	1.36	–.40	1.20
Shrev	7	32:31	93:45	.33	24.7	28.2	22.1	2.3	1.02	2.15	1.67
Des M	4	41:35	93:37	.33	35.5	11.8	22.1	5.9	.86	–.85	.68
Fort	5	41: 4	85: 9	.30	23.2	32.1	22.1	4.0	1.44	.65	2.08
Corpu	2	27:47	97:24	.29	24.7	23.9	22.1	1.0	1.02	–.15	1.83
Norfo	6	36:51	76:17	.26	26.0	24.9	19.6	6.7	.73	–1.35	1.83
Jacks	7	32:18	90:12	.25	26.4	23.9	22.1	5.9	.79	–1.75	1.88
Hunts	7	34:44	86:35	.24	26.0	30.4	12.9	5.9	.63	.75	1.38
Lexin	5	38: 3	84:30	.23	24.5	32.8	16.4	6.2	.88	–.10	1.65
Lubbo	2	33:35	101:51	.22	25.1	24.9	22.1	5.9	1.02	–.90	.85
Richm	6	37:33	77:27	.20	25.4	24.9	23.7	5.8	.66	–.10	2.05
Arlin	6	38:53	77: 7	.17	22.0	29.0	25.5	5.9	.66	–1.00	1.75
Kings	6	41:56	73:59	.17	20.4	24.9	22.1	5.9	1.02	1.50	1.75
Evans	5	37:58	87:35	.17	32.4	24.9	22.1	5.9	1.02	1.25	1.88
Kansa	4	36: 7	94:38	.16	43.4	18.5	17.6	4.7	.82	.30	1.25
Olymp	1	47:35	122:10	.16	22.7	24.9	22.1	5.9	1.27	.90	.95
Topek	4	39: 3	95:40	.16	29.0	24.9	22.1	5.9	1.02	1.80	1.85

Table 4: NMMAPS Data, Part 3

STATISTICS 174: APPLIED STATISTICS

SOLUTIONS TO 2001 FINAL EXAM

1. Each single malt appears in the experiment $\binom{k-1}{m-1}$ times, since after one malt is chosen, there are this number of ways of selecting $m-1$ other whiskies from the other $k-1$ choices.

   By the same argument, each pair of single malts appears in the experiment $\binom{k-2}{m-2}$ times.

   Therefore, the X^TX matrix is of the form $aI_n + bJ_n$ where

   $$a+b=\binom{k-1}{m-1}, \quad b=\binom{k-2}{m-2}. \tag{1}$$

   Note that this implies

   $$a=\binom{k-2}{m-1}. \tag{2}$$

   By the results in Section 3.2.4, the inverse matrix is of the form

   $$X^TX = cI_n + dJ_n$$

   where c and d are given by

   $$c=\frac{1}{a}, \quad d=-\frac{b}{a(a+nb)}. \tag{3}$$

   If we denote S_j as the sum of y_i for all blends in which single malt j is one of the constituents, then

   $$X^TY = \begin{pmatrix} S_1 \\ S_2 \\ \ldots \\ S_k \end{pmatrix}.$$

   Since $\widehat{\beta} = (X^TX)^{-1}X^TY$, it follows that

   $$\widehat{\beta}_j = cS_j + d\sum_{\ell} S_\ell. \tag{4}$$

   Combining equations (1)–(4) gives the desired explicit expression.

   Also, the variance of $\widehat{\beta}_j$ is

   $$(c+d)\sigma^2 = \frac{a+(n-1)b}{a(a+nb)}\sigma^2.$$

2. We have

$$X^T X = \begin{bmatrix} n & 0 & A \\ 0 & A & 0 \\ A & 0 & B \end{bmatrix}, \quad (X^T X)^{-1} = \begin{bmatrix} \frac{B}{nB-A^2} & 0 & -\frac{A}{nB-A^2} \\ 0 & \frac{1}{A} & 0 \\ -\frac{A}{nB-A^2} & 0 & \frac{n}{nB-A^2} \end{bmatrix}.$$

(a) The covariance matrix of $\widehat{\beta}$ is $(X^T X)^{-1}\sigma^2$ so this follows immediately from the above equation for $(X^T X)^{-1}$.

(b) The confidence interval consists of all x for which a null hypothesis $H_0: \ \beta_0 + \beta_1 x + \beta_2 x^2 = T$ is accepted at the .05 level. Using the answer to (a), the variance of $\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2$ is

$$\left\{ \frac{B}{nB - A^2} + \frac{x^2}{A} + \frac{x^4 n}{nB - A^2} - \frac{2x^2 A}{nB - A^2} \right\} \sigma^2.$$

This may be written in the form $(f + gx^2 + hx^4)\sigma^2$ where

$$f = \frac{B}{nB - A^2}, \quad g = \frac{1}{A} - \frac{2A}{nB - A^2}, \quad h = \frac{n}{nB - A^2}. \tag{5}$$

The obvious test statistic for H_0 is then

$$\frac{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 - T}{s\sqrt{f + gx^2 + hx^4}} \sim t_{n-3}.$$

We accept x for which

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 - T)^2 \le t^2_{n-3;.95} s^2 (f + gx^2 + hx^4). \tag{6}$$

Writing (6) in the form

$$\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 \le 0,$$

one possible specification of the constants $\alpha_0, ..., \alpha_4$ is

$$\begin{aligned}
\alpha_0 &= (\widehat{\beta}_0 - T)^2 - t^2_{n-3;.95} s^2 f, && (7) \\
\alpha_1 &= 2(\widehat{\beta}_0 - T)\widehat{\beta}_1, && (8) \\
\alpha_2 &= \widehat{\beta}_1^2 + 2(\widehat{\beta}_0 - T)\widehat{\beta}_2 - t^2_{n-3;.95} s^2 g, && (9) \\
\alpha_3 &= 2\widehat{\beta}_1 \widehat{\beta}_2, && (10) \\
\alpha_4 &= \widehat{\beta}_2^2 - t^2_{n-3;.95} s^2 h. && (11)
\end{aligned}$$

The final answer is obtained by combining (5) with (7)–(11).

3. Formulating this as a linear model in the usual way, we find

$$X^TX = \begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 \end{bmatrix}.$$

Thus in the notation of Section 3.2.4, we have

$$X^TX+cI_4 = (2+c)I_4+J_4, \quad (X^TX+cI)^{-1} = \frac{1}{2+c}I_4 - \frac{1}{(2+c)(6+c)}J_4. \tag{12}$$

(a) With $c=0$, $(X^TX)^{-1}$ is just $\frac{1}{2}I_4 - \frac{1}{12}J_4$. We also have

$$X^TY = \begin{bmatrix} y_1+y_2+y_3 \\ y_1+y_4+y_5 \\ y_2+y_4+y_6 \\ y_3+y_5+y_6 \end{bmatrix}.$$

Hence the first component of $\widehat{\beta} = (X^TX)^{-1}X^TY$ is

$$\begin{aligned} & \frac{1}{2}(y_1+y_2+y_3) - \frac{1}{12}(2y_1+2y_2+2y_3+2y_4+2y_5+2y_6) \\ = \quad & \frac{1}{3}(y_1+y_2+y_3) - \frac{1}{6}(y_4+y_5+y_6). \end{aligned}$$

The associated variance of $\widehat{\beta}_1$ is $\frac{5}{12}\sigma^2$.

(b) By the results in Section 5.2.4 of the notes, the variance of the ridge regression estimator is $(X^TX+cI)^{-1}X^TX(X^TX+cI)^{-1}\sigma^2$ and the bias is $-c(X^TX+cI)^{-1}\beta$. In the present case, we calculate

$$\begin{aligned} & \left[\frac{1}{2+c}I_4 - \frac{1}{(2+c)(6+c)}J_4\right][2I_4+J_4] \\ = \quad & \left[\frac{2}{2+c}I_4 + \frac{c}{(2+c)(6+c)}J_4\right], \\ & \left[\frac{2}{2+c}I_4 + \frac{c}{(2+c)(6+c)}J_4\right]\left[\frac{1}{2+c}I_4 - \frac{1}{(2+c)(6+c)}J_4\right] \\ = \quad & \left[\frac{2}{(2+c)^2}I_4 + \frac{c^2-12}{(2+c)^2(6+c)^2}J_4\right]. \end{aligned}$$

In particular, the variance of $\tilde{\beta}_1^{(1)}$ is

$$\left\{\frac{2}{(2+c)^2} + \frac{c^2-12}{(2+c)^2(6+c)^2}\right\}\sigma^2 = \frac{3(c^2+8c+20)}{(2+c)^2(6+c)^2}\sigma^2. \tag{13}$$

The bias is

$$-c\left[\frac{1}{2+c}I_4-\frac{1}{(2+c)(6+c)}J_4\right]\beta$$

and the first component of this is

$$\begin{aligned} & -c\left[\frac{1}{2+c}\beta_1-\frac{1}{(2+c)(6+c)}(\beta_1+...+\beta_4)\right] \\ = & -\frac{c(5+c)}{(6+c)}\beta_1+\frac{c}{(2+c)(6+c)}(\beta_2+\beta_3+\beta_4). \end{aligned} \quad (14)$$

The optimization problem therefore chooses c to minimize $S+B^2$, where S is given by (13) and B by (14).

(c) In this case, $\tilde{y}_1^{(c)}=\tilde{\beta}_1^{(c)}+\tilde{\beta}_2^{(c)}$ so the bias of $\tilde{y}_1^{(c)}$ is the sum of the biases for $\tilde{\beta}_1^{(c)}$ and $\tilde{\beta}_2^{(c)}$, i.e. the sum of (14) and the corresponding expression with β_1 and β_2 interchanged.

By the independence of past and future observations, the variance of $\tilde{y}_1^{(c)}$

$$\sigma^2+\mathrm{Var}(\tilde{\beta}_1^{(c)})+\mathrm{Var}(\tilde{\beta}_2^{(c)})+2\mathrm{Cov}(\tilde{\beta}_1^{(c)},\tilde{\beta}_2^{(c)}). \quad (15)$$

The variances of $\tilde{\beta}_1^{(c)}$ and $\tilde{\beta}_2^{(c)}$ are both given by (13), while the covariance is

$$\frac{c^2-12}{(2+c)^2(6+c)^2}\sigma^2. \quad (16)$$

The variance S of $\tilde{y}_1^{(c)}$ is derived by combining (13), (15) and (16), while the bias B is given as the sum of (14) and the corresponding expression with β_1 and β_2 interchanged. The optimal value of c is again that which minimizes $S+B^2$.

4. Problem about NMMAPS study.

(a) Successive F tests of one model against the next yield F statistics 3.66 (for $p=0$ against $p=1$); 1.73 ($p=1$ against $p=2$, though note this combination is not nested); 3.15, 1.09, 0.55 etc. The 95% point for $F_{1,\nu}$ where $\nu\approx 88$ is about 4.00; thus, none of these tests is significant. On this basis, it looks as though either forward or backward selection would result in $p=0$. On the other hand, if we test $H_0:\ p=0$ against $H_1:\ p=3$ (based on the r3, pm, so2 variables) we get an F statistic of 2.90 and the corresponding $F_{3,84,.95}$ value is about 2.7. So this is significant.

AIC, BIC calculations are as in Table 5 based on

$$AIC=n\log SSE+2p, \quad BIC=n\log SSE+p\log n,$$

and suggest that the best-AIC model is $p=3$ and the best-BIC model is $p=0$. The choice appears to be between those two.

p	SSE	AIC	BIC
0	388.86	388.86	388.86
1	385.20	387.20	389.68
2	383.43	387.43	392.38
3	380.18	386.18	393.61
4	379.03	387.03	396.94

Table 5: AIC and BIC calculations

(b) As noted in (a), both BIC and successive F testing suggest $p = 0$ as the optimal model, which would therefore support the statement that there is no effect due to any of the regressors. On the other hand, a direct test of $p = 0$ against $p = 3$ does produce a significant result. The answer to the direct question, whether any of the models is significant against the null model, is "yes" in the case of $p = 3$.

(c) The model with $p = 1$ has r6 as the only significant variable, so presumably the coefficient is positive and this confirms that region 6 has the highest mortality ratio (though not significantly, according to this analysis). On the other hand, the $p = 3$ model has both pm and so2 as covariates, and r3 as the only significant "region" covariate (with a positive coefficient, from the SAS output). Therefore, it looks as though when the model is properly adjusted to allow for variable background levels in PM_{10} and SO_2, it is region 3 (southern California), not region 6, which has the highest mortality ratios.

(d) Large studentized residuals include observation 37 (2.737), 62 (2.592), 70 (–3.120) and 78 (–2.387).

(e) $p = 4$ (counting the intercept) so $\frac{2p}{n} = .0909$. Large h_{ii} values include observations 1,6,7,15,16,17,21,40,52,67,86. In other words, there are many points of possibly high leverage here.

(f) For DFFITS, the cutoff is $2\sqrt{\frac{p}{n}} = .426$ and by this criterion observations 37, 70, 73 are influential.

For DFBETAS, the cutoff is $\frac{2}{\sqrt{n}} = .213$ and (in addition to the foregoing) this means each of observations 16, 17, 53, 55, 67, 83 is influential in at least one β_j.

(g) The largest VIF is 1.76; largest condition index is 12.4. No problem with multicollinearity.

(h) The choice is between $\frac{1}{88}\sum y_i$ and $\frac{1}{88}\sum \widehat{y}_i$ as estimate of overall average effect (ignoring the weightings). $\sum \widehat{y}_i$ could be biased if we did not identify the correct regression model. Normally, we would expect it to have lower variance, however. The two could be examined analytically because $\sum y_i$ has variance $n\sigma^2$ while the variance of $\sum \widehat{y}_i$ is of the form $\mathrm{tr}\{(X^TX)^{-1}X^TJX\}\sigma^2$ where J is the $n \times n$ matrix of ones. To see this, note that the vector $\widehat{Y} = HY$ has covariance matrix $H\sigma^2$ where H is the hat matrix, so the variance of $\sum \widehat{y}_i$ is $\mathbf{1}^T H \mathbf{1} \sigma^2$ where $\mathbf{1}$ is the column vector of ones. But

$$\begin{aligned}
\mathbf{1}^T H \mathbf{1} &= \operatorname{tr}\{\mathbf{1}^T X (X^T X)^{-1} X^T \mathbf{1}\} \\
&= \operatorname{tr}\{(X^T X)^{-1} X^T \mathbf{1}\mathbf{1}^T X\} \\
&= \operatorname{tr}\{(X^T X)^{-1} X^T J X\}
\end{aligned}$$

which reduces the variance expression to the given form. Of course, we can't tell how much $\operatorname{tr}\{(X^T X)^{-1} X^T J X\}$ is less than 1 without actually doing the calculations, but if it was a great deal less than 1, that would be an argument in favor of using the regression-based calculation. There was no definitive "right answer" to this question, but definite bonus points if you discussed the $\operatorname{tr}\{(X^T X)^{-1} X^T J X\}\sigma^2$ formula or anything equivalent to it.