STATISTICS 174: APPLIED STATISTICS

MIDTERM EXAM

OCTOBER 16, 2001

Time allowed: 75 minutes.

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.

Statistical tables are not provided, but are not needed for the questions being asked. When the calculations call for a percentage point of a t or F distribution, be sure to state precisely the degrees of freedom and the required tail probability, but a numerical value is not required.

1. **Question One**.

   A metallic alloy contains K elements in proportions $\beta_1, ..., \beta_K$, such that each $\beta_k > 0$ $(1 \leq k \leq K)$ and $\sum_{k=1}^{K} \beta_k = 1$. A chemical separation processes allows each β_k to be measured with error. Specifically, we have observations of the form

   $$y_i = \beta_{k(i)} + \epsilon_i, \quad 1 \leq i \leq n,$$

   where there are n observations in total, the ith observation is of element $k(i)$, and the ϵ_i are independent normal random variables with mean 0 and common unknown variance σ^2.

   An experiment is taken in which there are a measurements of each of the first $K-1$ elements and b measurements of the Kth, where $(K-1)a+b = n$. Thus

   $$\begin{aligned}
   k(i) &= 1, \quad i = 1, 2, ..., a, \\
   k(i) &= 2, \quad i = a+1, a+2, ..., 2a, \\
   &\vdots \\
   k(i) &= K-1, \quad i = (K-2)a+1, (K-2)a+2, ..., (K-1)a, \\
   k(i) &= K, \quad i = (K-1)a+1, (K-1)a+2, ..., n.
   \end{aligned}$$

(a) Show how this experiment can be set up as a linear model to estimate the parameters $\beta_1, ..., \beta_K$. Specifically, write the model in the form $y = X\beta + \epsilon$ and derive the least squares equations for β.

(b) Calculate the variances of the least squares estimators $\widehat{\beta}_1, ..., \widehat{\beta}_K$.

(c) How should a and b be determined if the objective is to minimize the sum of the variances of $\widehat{\beta}_1, ..., \widehat{\beta}_K$?

(d) Show how to construct an F test of the hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_K = \frac{1}{K}$ against the alternative H_1 that $\beta_1, ..., \beta_K$ are not all equal.

(e) Suppose in fact the test in the previous part is carried out when $\beta_1 = \beta_2 = ... = \beta_{K-1}$ but $\beta_K = 2\beta_1$. Explain how to calculate the power of the test in this case.

2. **Question Two**.

Table 1 gives the results of a survey of 30 new car buyers. The data tabulated are the buyer's income, age, sex (0=male, 1=female) and whether they live in the north-east United States (variable NE; 1 for yes, 0 for no). The last column give the price they paid for their car. We are interested in determining the effect that the first four variables has on the price someone is willing to pay for their car.

Some sample S-PLUS output follows:

```
> nreg<-lm(price~income+age+sex+ne)
> summary(nreg)

Call: lm(formula = price ~ income + age + sex + ne)
Residuals:
   Min    1Q Median    3Q   Max
 -4780 -978    493 1396 2787

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -6685.268  1795.950     -3.722     0.001
     income     0.310     0.014     22.644     0.000
        age   112.185    35.066      3.199     0.004
        sex  1032.362   831.964      1.241     0.226
         ne  2196.588   836.634      2.626     0.015

Residual standard error: 1990 on 25 degrees of freedom
Multiple R-Squared: 0.966
F-statistic: 177 on 4 and 25 degrees of freedom, the p-value is 0
```

Income	Age	Sex	NE	Price
67500	50	0	0	19690
30500	45	0	1	5220
52000	41	0	0	14720
75500	46	1	0	22390
117400	48	0	1	37850
96800	77	1	1	33100
24000	40	1	1	8830
152700	44	0	1	50110
101100	34	0	0	27470
97700	46	0	1	29520
20100	23	1	0	1650
42200	39	0	1	13850
24600	50	0	0	8050
63000	48	1	0	21320
49200	47	1	0	16750
66300	70	0	0	22550
54500	60	0	1	20620
91100	59	0	0	29190
75600	30	0	0	19700
69700	47	0	0	21700
32100	46	1	0	8560
34400	45	1	0	12830
94100	52	0	0	25230
54500	48	0	1	20190
46800	39	1	0	13920
86900	28	0	0	24260
49000	38	1	0	12790
51500	49	1	0	15180
84000	50	0	0	21170
57200	46	1	0	15860

Table 1: Prices of cars bought by owner characteristics

```
Correlation of Coefficients:
       (Intercept) income    age    sex
income -0.399
   age -0.775      -0.179
   sex -0.355       0.395 -0.059
    ne  0.032      -0.075 -0.186  0.181
> nregi<-lm.influence(nreg)
> nregi$h
 [1] 0.0920452 0.2299902 0.1028753 0.1204862 0.1885222 0.4664165
 [7] 0.2620485 0.3965901 0.1674193 0.1405760 0.2436487 0.2034206
[13] 0.2107578 0.0990998 0.0898913 0.2872482 0.1926143 0.1477387
[19] 0.1509200 0.0833289 0.1029539 0.0981935 0.1113893 0.1477021
[25] 0.0971376 0.1877002 0.1012046 0.0940675 0.0925591 0.0914544
> studres(nreg)
         1        2       3         4        5        6        7
 -0.065646 -3.20175 0.36996 -0.255819 0.336784 -1.43467 0.212244
       8         9       10        11       12       13       14       15
 1.59719 -0.516322 -0.74768 -0.86091 0.499449 0.849626 1.10958 1.00084
       16       17       18        19       20       21       22       23
 0.504168 0.837838 0.567792 -0.203391 0.801056 -0.46094 1.5109 -1.68005
      24       25       26        27        28       29        30
 1.33949 0.369588 0.497155 -0.515024 -0.313288 -2.11142 -0.705356
> dffits(nreg)
          1        2        3          4        5        6        7
 -0.0209015 -1.74982 0.125281 -0.0946847 0.162328 -1.34134 0.126477
       8         9       10        11       12       13       14
 1.29486 -0.231532 -0.30239 -0.488628 0.252391 0.439051 0.368006
       15       16       17        18         19       20        21
 0.314541 0.320063 0.409227 0.236401 -0.0857495 0.24152 -0.156156
       22        23       24       25       26        27        28
 0.498562 -0.594823 0.557617 0.121227 0.238982 -0.172821 -0.100952
        29        30
 -0.674335 -0.223788
> dfbetas(nreg)
   (Intercept)      income          age         sex         ne
 1 -0.00234853  0.00374479 -0.006767847  0.01296468  0.0117083
 2 -0.79245008  1.18169752  0.082916321  0.75669778 -1.0416852
 3  0.07790895 -0.05171408 -0.016093846 -0.08630357 -0.0517410
 4  0.02218545 -0.04865560  0.001787028 -0.06967405  0.0232583
 5 -0.02317834  0.09837792 -0.028634383  0.00949713  0.0864374
 6  1.07939873 -0.38422426 -0.853947281 -0.70165365 -0.4363010
 7  0.03833759 -0.04632935 -0.031456719  0.04889383  0.0941037
 8 -0.21415052  1.06512543 -0.391404590  0.25776056  0.4652660
 9 -0.09078387 -0.12825498  0.127170053  0.04551251  0.0761590
```

```
10 -0.01557750 -0.10825270  0.070666632  0.02097368 -0.1998887
11 -0.38068859  0.12813282  0.340047752 -0.12752312 -0.0184090
12  0.14943149 -0.12572209 -0.079874665 -0.09480882  0.1674394
13  0.14507360 -0.33358736  0.144320872 -0.29134491 -0.1415496
14 -0.08694090  0.10085630  0.051449093  0.25650800 -0.0998792
15 -0.02401122 -0.00555435  0.045474015  0.19198259 -0.0787853
16 -0.14074970 -0.07499606  0.269845282 -0.12667710 -0.1403442
17 -0.04979950 -0.19243781  0.197607740 -0.14709264  0.2088308
18 -0.09440073  0.04876077  0.140064353 -0.08206431 -0.1301252
19 -0.06289742 -0.01417740  0.058241987  0.03211930  0.0239300
20  0.05774630 -0.02493931  0.035736478 -0.14985722 -0.1348899
21 -0.01831981  0.05816404 -0.022743084 -0.06697212  0.0306937
22  0.07158260 -0.16002212  0.041844582  0.23052399 -0.0969907
23  0.11787211 -0.21728788 -0.179640687  0.21038779  0.3388384
24  0.16023002 -0.24460599  0.000918646 -0.21087809  0.3821256
25  0.03614211  0.00133376 -0.036784016  0.07186249 -0.0185144
26  0.15515931  0.08168745 -0.172627595 -0.06220665 -0.0593673
27 -0.05461910 -0.01177685  0.062407371 -0.10417344  0.0246856
28  0.01859685 -0.00138014 -0.024929044 -0.06160790  0.0272674
29  0.02469754 -0.13053886 -0.172840761  0.31783473  0.3953922
30  0.02306201 -0.03904611 -0.011777988 -0.15225402  0.0561703
```

Explanation of this output. It is not necessary to be an S-PLUS expert to understand this! The "lm" command fits a linear model in which price is the y variable and income, age, sex and ne are x variables. "summary(nreg)" gives the coefficients and standard errors, estimate of the residual standard error s, and the correlations of the coefficients $\widehat{\beta}_j$. The remaining output gives the diagonal elements of the hat matrix (nregi$h), the (externally) studentized residuals, and the values of dffits and dfbetas.

(a) Based on the above output, what would you consider the best model for the data? In other words, which (if any) of the four covariates would you drop from the model?

(b) A second run of the "lm" command but including only income and age as covariates results in a residual standard error of 2180 on 27 degrees of freedom. Based on this, perform an F test of the hypothesis that both the "sex" and "ne" coefficients are 0.

(c) Discuss the questions of leverage, outlyingness and influence for this data set. In particular, say which observations appear to have high leverage, which ones you suspect of being outliers, and which ones have high influence.

(d) Table 2 gives the profiles of five new customers. For the first of these (income $80,000, age 47, male, not in NE) show how to calculate a

95% prediction interval for the price he will pay for his car. (*Note:* the actual numerical calculations may be a little involved for this, but I am not looking for a numerical answer: what is needed is a clear statement of the method for calculating it.)

(e) Now consider all five potential new customers in Table 2 and suppose we want a simultaneous 95% *confidence* interval (not prediction intervals) for the mean prices they would be expected to pay for their cars. Describe the method for calculating this. (Once again, actual numerical calculations are not required to long as the method is clearly explained. Be sure to indicate the degrees of freedom and tail probability of any t or F percentage points that are involved.)

Income	Age	Sex	NE
80000	47	0	0
50000	35	1	0
72500	54	1	0
32000	27	0	0
102100	61	1	0

Table 2: New owners for confidence and prediction interval calculations

SOLUTIONS

1. **Question One**

Writing $\beta_K = 1 - \beta_1 - ... - \beta_{K-1}$, the X matrix is of the form

$$\begin{array}{lll} (1 & 0 & 0 \quad \ldots \quad 0) \qquad (a \text{ times}), \\ (0 & 1 & 0 \quad \ldots \quad 0) \qquad (a \text{ times}), \\ (0 & 0 & 1 \quad \ldots \quad 0) \qquad (a \text{ times}), \\ \vdots & & \\ (0 & 0 & 0 \quad \ldots \quad 1) \qquad (a \text{ times}), \\ (-1 & -1 & -1 \quad \ldots \quad -1) \qquad (b \text{ times}). \end{array}$$

Hence $X^T X$ is the matrix with $a+b$ on the main diagonal and b everywhere else, i.e. in the notation of Section 3.2.4, it is $aI_{K-1} + bJ_{K-1}$. Hence by the result of page 118, $(X^T X)^{-1}$ is $cI_{K-1} + dJ_{K-1}$, where

$$c = \frac{1}{a}, \qquad d = -\frac{b}{a\{a + (K-1)b\}}.$$

Also

$$X^T Y = \begin{pmatrix} S_1 - S_K \\ S_2 - S_K \\ \vdots \\ S_{K-1} - S_K \end{pmatrix}$$

where S_k denotes the sum of all observations on alloy k.

(a) The least squares estimates are of the form

$$\widehat{\beta}_k = (c+d)(S_k - S_K) + d \sum_{j \neq k} (S_j - S_K),$$

for $1 \leq k \leq K-1$, where c and d are as above. The estimate for $\widehat{\beta}_K$ is obtained by subtraction.

(b) The variances of $\widehat{\beta}_1, .., \widehat{\beta}_{K-1}$ are each

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

For $\widehat{\beta}_K$, using the subtraction formula the answer is

$$\begin{aligned} & (K-1)(c+d)\sigma^2 + (K-1)(K-2)d\sigma^2 \\ = \; & \frac{(K-1)\sigma^2}{a} \left[\frac{a + (K-2)b}{a + (K-1)b} - \frac{(K-2)b}{a + (K-1)b} \right] \\ = \; & \frac{(K-1)\sigma^2}{a + (K-1)b}. \end{aligned}$$

(c) The sum of the variances is

$$\begin{aligned} & \frac{(K-1)\sigma^2}{a} \cdot \frac{a+(K-2)b}{a+(K-1)b} + \frac{(K-1)\sigma^2}{a+(K-1)b} \\ = \quad & \frac{\sigma^2}{a} \cdot \frac{(K-1)(a+(K-2)b)+(K-1)a}{a+(K-1)b} \\ = \quad & \frac{(K-1)\sigma^2\{2a+(K-2)b\}}{a\{a+(K-1)b\}}. \end{aligned}$$

Hence the optimization problem chooses a and b to minimize

$$\frac{2a+(K-2)b}{a\{a+(K-1)b\}}$$

subject to $(K-1)a+b=n$. After substituting for b, the problem becomes to minimize

$$\phi(a) = \frac{(K-2)n-K(K-3)a}{a\{(K-1)n-K(K-2)a\}}.$$

Strictly speaking we should minimize this subject to the constraint that a be integer, but if we ignore this and treat a as a continuous variable, it is an easy guess (because of the symmetry of the resulting configuation) that the optimal a is n/K. We shall now demonstrate this to the extent of showing that $\phi'(n/K)=0$. Writing

$$\log\phi(a) = \log\{(K-2)n-K(K-3)a\} - \log\{(K-1)n-K(K-2)a\} - \log a$$

and differentiating,

$$\frac{\phi'(a)}{\phi(a)} = -\frac{K(K-3)}{(K-2)n-K(K-3)a} + \frac{K(K-2)}{(K-1)n-K(K-2)a} - \frac{1}{a},$$

which is 0 when $a=n/K$.

(d)

$$F = \frac{SSE_0 - SSE_1}{K-1} \cdot \frac{n-K+1}{SSE_1} \sim F_{K-1,n-K+1} \qquad \text{under } H_0.$$

Here $SSE_1 = \sum(y_i - \widehat{\beta}_{k(i)})^2$ and SSE_0 is the same thing with $\widehat{\beta}_{k(i)}$ replaced by $\frac{1}{K}$.

(e) The simplest way to do this is using equation (3.43), page 133. Here $C = I_{K-1}$, $h = \frac{1}{K}\mathbf{1}_{K-1}$ ($\mathbf{1}_{K-1}$ is the $(K-1)$-dimensional vector of ones) and $h' = \frac{1}{K+1}\mathbf{1}_{K-1}$. Note that $h - h' = \frac{1}{K(K+1)}\mathbf{1}_{K-1}$. Hence

(3.43) becomes

$$\begin{aligned}\sigma^2\delta^2 &= \left\{\frac{1}{K(K+1)}\right\}^2 \mathbf{1}_{K-1}(X^TX)\mathbf{1}_{K-1} \\ &= \left\{\frac{1}{K(K+1)}\right\}^2 \{(K-1)(a+b)+(K-1)(K-2)b\} \\ &= \left\{\frac{1}{K(K+1)}\right\}^2 (K-1)\{a+(K-1)b\}.\end{aligned}$$

We solve this equation to get δ; the power is then derived from the $F'_{K-1,n-K+1;\delta}$ distribution.

2. **Question Two**

(a) Drop sex because this has a t value of only 1.2 (p-value .226 > .05); all the rest appear clearly significant.

(b) The S-PLUS output quotes a residual standard error of 1990 on 25 degrees of freedom, which translates to a SSE_1 of $25 \times 1990 \times 1990$. Similarly, $SSE_0 = 27 \times 2180 \times 2180$. Hence

$$F = \frac{SSE_0 - SSE_1}{2} \cdot \frac{25}{SSE_1} = 3.7$$

which has an $F_{2,25}$ distribution. Although you could not have determined the "significance" of this without having F tables to hand, it is in fact significant (the actual p-value is .039).

(c) $p = 5$ and $n = 30$ so $\frac{2p}{n} = .333$, $2\sqrt{\frac{p}{n}} = .816$, $\frac{2}{\sqrt{n}} = .365$.

The high leverage values ($h_{ii} > \frac{2p}{n}$) are observations 6 and 8.

The outliers are observation 2 (studentized residual –3.2) and possibly 29 (–2.11 — the .975-quantile of the t_{25} distribution is actually 2.06).

By the DFFITS criterion, the influential observations are 2, 6 and 8.

By the DFBETAS criterion, there is at least one parameter estimate strongly influenced by observations 2, 6, 8, 11, 24, 29, though in several of these cases the evidence does not appear strong.

Overall, it appears observation 2 is the worst outlier but is not high leverage, while observations 6 and 8 combine high leverage with moderately high magnitudes of the studentized residuals (–1.43 and +1.60) and are therefore influential.

(d) The point prediction is

$$-6685.268 + .310 \times 80000 + 112.185 \times 47 = 23387.$$

The prediction variance is of the form

$$\begin{aligned} & 1795.950^2 + (80000 \times .014)^2 + (47 \times 35.066)^2 \\ + \quad & 2 \times (-.399) \times 80000 \times 1795.95 \times .014 \\ + \quad & 2 \times (-.775) \times 47 \times 1795.95 \times 35.066 \\ + \quad & 2 \times (-.179) \times 80000 \times 47 \times .014 \times 35.066 \\ + \quad & 1990^2, \end{aligned}$$

where the last term corresponds to the $+\sigma^2$ term that arise in going from a confidence interval to a prediction interval (note that $s = 1990$ here). Although you were not expected to evaluate the above expression numerically, the actual value is $4302347 = 2074^2$. Thus, the prediction standard error is 2074 and the 95% prediction interval is $23387 \pm t_{.975;25} \times 2074$.

(e) We repeat the above calculation corresponding to each of the five potential buyers but omitting the last term ($+1990^2$) in the prediction variance calculation, because we are now asked for confidence intervals rather than prediction intervals. The simultaneous prediction intervals are then of the form

$$\text{point prediction } \pm G\sqrt{\text{prediction variance}}$$

where $G^2 = qF_{q;n-p;.95}$ as in equation (3.41), page 130, of the text. In this question, $n = 30$, $p = 5$, $q = 4$. The reason q is 4 rather than 5 in this instance is because each of the five c_k values for which a prediction is required has NE=0; therefore (counting first coordinate 1 for each c_k since the regression has an intercept), the actual dimension of the space in which all the c_k lie is 4 rather than 5.