STATISTICS 174: APPLIED STATISTICS

MIDTERM EXAM

OCTOBER 15, 2002

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.

Answer both Parts, A and B, as far as you are able to within the time allowed. Credit will be given for all partially correct answers.

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: you do not need to know precise values for any distributions to be able to answer the following questions.

Part A.

This is about a variant of the weighing problem (Section 3.2.4, page 117) in which the scale is subject to an unknown bias, i.e. if an object of weight β is placed in the pan, the mean of the measured weight is not β but $\beta + \gamma$, with γ common to all observations, but unknown. As in Section 3.2.4, all observations are independent and have an additional random error with variance σ^2.

For this scenario, the weighing designs proposed in Section 3.2.4 do not work directly because it is not possible to separate the estimation of γ from the actual weights of the objects. (You are not asked to prove this; just take it as given.) However, some alternative estimation schemes do make it possible to estimate all the weights.

1. Consider a scheme with just two objects to weigh, and three weighings set out as follows:

   (a) Weigh object 1

   (b) Weigh object 2

   (c) Weigh objects 1 and 2 together

Write the model in the form

$$y_i = \beta_0 + \beta_1(x_{i1} - \bar{x}_{\cdot 1}) + \beta_2(x_{i2} - \bar{x}_{\cdot 2}) + \epsilon_i, \tag{1}$$

where β_1 and β_2 are the weights of the two objects; x_{i1} is 1 if object 1 is in the pan on weighing i, 0 otherwise; x_{i2} is 1 if object 2 is in the pan on weighing i, 0 otherwise; $\bar{x}_{\cdot 1}$ is the mean over all $x_{i1},\ i = 1, 2, 3$; $\bar{x}_{\cdot 2}$ is the mean over all $x_{i2},\ i = 1, 2, 3$; and $\gamma = \beta_0 - \beta_1\bar{x}_{\cdot 1} - \beta_2\bar{x}_{\cdot 2}$.

Show that this formulation leads to an X matrix

$$\begin{pmatrix} 1 & \frac{1}{3} & -\frac{2}{3} \\ 1 & -\frac{2}{3} & \frac{1}{3} \\ \cdots & \cdots & \cdots \end{pmatrix}$$

and fill in the dots in the last row.

Hence write down X^TX, find $(X^TX)^{-1}$ (which confirms that the least squares estimation equations have a unique solution) and hence find the common variances of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ under the least squares estimation scheme. Also find the least squares estimator $\widehat{\beta}_1$ as a linear combination of $y_1,\ y_2$ and y_3.

2. Now consider the same problem with four objects to weigh, and suppose (as a combination of the two weighing schemes considered in Section 3.2.4) we weigh each object on its own once, and each of the six possible pairs of objects once, for a total of ten weighings. With the obvious extension of the notation in (1), write the model in the form

$$y_i = \beta_0+\beta_1(x_{i1}-\bar{x}_{\cdot 1})+\beta_2(x_{i2}-\bar{x}_{\cdot 2})+\beta_3(x_{i3}-\bar{x}_{\cdot 3})+\beta_4(x_{i4}-\bar{x}_{\cdot 4})+\epsilon_i, \tag{2}$$

where x_{ij} is 1 if object j is in the pan on weighing i, 0 otherwise, and $\bar{x}_{\cdot j}$ is the mean of the x_{ij} over $i = 1, 2, ..., 10$.

For this problem, write down X^TX, verify that $(X^TX)^{-1}$ exists, and calculate it. Hence find the (common) variance of $\widehat{\beta}_1, ..., \widehat{\beta}_4$ using the least squares estimators. (For this part of the question, you are not asked to find the least squares estimators.)

Part B.

In a clinical trial of a new drug designed to reduce blood pressure, 30 men are sampled and their age in years, weight in kg. and reduction in blood pressure (B.P.) recorded, as follows:

Age x_{i1}	Weight x_{i2}	B.P. y_i	Age x_{i1}	Weight x_{i2}	B.P. y_i	Age x_{i1}	Weight x_{i2}	B.P. y_i
51	59.3	12.9	55	128.3	15.0	46	103.3	9.0
44	100.7	14.1	39	132.5	6.1	42	122.1	7.5
49	73.3	-0.3	38	88.7	3.5	62	89.3	1.1
52	83.7	10.7	53	66.8	13.2	55	92.7	18.1
46	77.3	9.8	54	92.0	21.6	52	82.2	13.5
30	60.3	9.3	59	96.3	12.3	60	100.6	34.1
49	73.4	2.0	44	95.9	15.6	63	83.7	23.0
43	68.9	6.2	47	73.4	10.5	28	79.0	-2.2
34	106.3	1.3	37	81.1	-3.5	56	79.6	4.1
46	72.4	10.6	51	83.3	6.1	43	70.9	8.7

Assume B.P. for the ith patient is y_i, age is x_{i1} and weight is x_{i2}. The following may be assumed without checking: $\sum y_i = 293.9, \sum (y_i - \bar{y})^2 = 1846.67$, $\sum y_i(x_{i1} - \bar{x}_{\cdot 1}) = 1137.46$, $\sum y_i(x_{i2} - \bar{x}_{\cdot 2}) = 659.4343$, $\sum (x_{i1} - \bar{x}_{\cdot 1})^2 = 2309.2$, $\sum (x_{i2} - \bar{x}_{\cdot 2})^2 = 9800.354$, $\sum (x_{i1} - \bar{x}_{\cdot 1})(x_{i2} - \bar{x}_{\cdot 2}) = 204.42$. Here $\bar{x}_{\cdot j},\ j = 1, 2$ means the average of x_{ij} over $i = 1, ..., 30$.

Write the model in the form

$$y_i = \beta_1 + \beta_2(x_{i1} - \bar{x}_{\cdot 1}) + \beta_3(x_{i2} - \bar{x}_{\cdot 2}) + \epsilon_i, \tag{3}$$

with $\{\epsilon_i\}$ assumed independent $N[0, \sigma^2]$ for some unknown σ^2.

1. Using the computational formulae for linear regression on two covariates with an intercept, calculate point estimates $\widehat{\beta}_j,\ j = 1, 2, 3$.
2. Calculate s^2.
3. Hence calculate the standard errors of $\widehat{\beta}_j,\ j = 1, 2, 3$.
4. Test the hypotheses $\beta_2 = 0$ and $\beta_3 = 0$. What do you conclude?
5. The following diagnostics are computed. The 30 values of h_{ii} are

```
 [1] 0.11988922 0.05834409 0.05441139 0.04329834 0.04426311 0.23335081
 [7] 0.05412480 0.07538321 0.15539322 0.05653703 0.22390930 0.28184081
[13] 0.07378709 0.09070571 0.05286586 0.09623293 0.04718068 0.05292954
[19] 0.08476036 0.04018092 0.06126449 0.17467493 0.12319317 0.05940402
[25] 0.04473446 0.11528190 0.13849663 0.20402401 0.07108001 0.06845797
```

The values of the externally studentized residuals are

```
 0.46875921  0.79424610 -1.54423164 -0.15305068  0.19914681  1.66431746
-1.16755649 -0.04586624 -0.46479608  0.36124336 -0.12249977 -0.35573657
-0.25456885  0.29313257  1.28308925 -0.54395312  1.06428922  0.26464545
-1.20258139 -0.76206507 -0.13879668 -0.24710292 -2.77067342  0.65613355
 0.27234328  3.14658055  0.92988566 -0.31857324 -1.45304215  0.31074510
```

The values of DFFITS are

```
 0.17301000  0.19770033 -0.37043008 -0.03255987  0.04285734  0.91821099
-0.27929242 -0.01309633 -0.19936605  0.08843088 -0.06579837 -0.22285382
-0.07185220  0.09258255  0.30313688 -0.17749866  0.23682986  0.06256369
-0.36596829 -0.15592196 -0.03545781 -0.11367910 -1.03854827  0.16489162
 0.05893533  1.13583950  0.37283843 -0.16128721 -0.40194146  0.08423936
```

Comment on these values from the point of view of determining which observations are (a) point of high leverage, (b) outliers, (c) influential.

SOLUTIONS

Part A

1. The last row of X is $\begin{pmatrix} 1 & \frac{1}{3} & -\frac{2}{3} \end{pmatrix}$. Hence

$$X^TX = \begin{pmatrix} 3 & 0 & 0 \\ 0 & \frac{2}{3} & -\frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}, \quad (X^TX)^{-1} = \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

The variances of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are both $2\sigma^2$.

To find $\widehat{\beta}_1$, note that

$$\widehat{\beta} = \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} y_1+y_2+y_3 \\ \frac{y_1-2y_2+y_3}{3} \\ \frac{-2y_1+y_2+y_3}{3} \end{pmatrix}$$

so that

$$\widehat{\beta}_1 = \frac{2y_1 - 4y_2 + 2y_3 - 2y_1 + y_2 + y_3}{3} = y_3 - y_2$$

as could equally well be guessed as the only combination of (y_1, y_2, y_3) that gives an unbiased estimator.

2. If we don't subtract off the $\bar{x}_{\cdot j}$'s, the X matrix is

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

After subtracting $\bar{x}_{\cdot j},\ j = 2, 3, 4, 5$, this becomes

$$\begin{pmatrix} 1 & \alpha & \beta & \beta & \beta \\ 1 & \beta & \alpha & \beta & \beta \\ 1 & \beta & \beta & \alpha & \beta \\ 1 & \beta & \beta & \beta & \alpha \\ 1 & \alpha & \alpha & \beta & \beta \\ 1 & \alpha & \beta & \alpha & \beta \\ 1 & \alpha & \beta & \beta & \alpha \\ 1 & \beta & \alpha & \alpha & \beta \\ 1 & \beta & \alpha & \beta & \alpha \\ 1 & \beta & \beta & \alpha & \alpha \end{pmatrix}$$

with $\alpha = \frac{3}{5}$, $\beta = -\frac{2}{5}$ to make the column sums (except the first) all 0. Then

$$X^TX = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \frac{12}{5} & -\frac{3}{5} & -\frac{3}{5} & -\frac{3}{5} \\ 0 & -\frac{3}{5} & \frac{12}{5} & -\frac{3}{5} & -\frac{3}{5} \\ 0 & -\frac{3}{5} & -\frac{3}{5} & \frac{12}{5} & -\frac{3}{5} \\ 0 & -\frac{3}{5} & -\frac{3}{5} & -\frac{3}{5} & \frac{12}{5} \end{pmatrix}.$$

The key step is to note that this matrix is of block-diagonal form, and the lower-right 4×4 block is of the form $aI_4 + bJ_4$ with $a = 3,\ b = -\frac{3}{5}$. By the usual formulae, the inverse is $cI_4 + dJ_4$ with $c = \frac{1}{3},\ d = \frac{1}{3}$. Therefore,

$$(X^TX)^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \frac{2}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

In particular, the variances of $\widehat{\beta}_1, ..., \widehat{\beta}_4$ (the least squares estimators) are all $\frac{2}{3}\sigma^2$, which answers the question as asked.

Footnote. Although the question didn't ask you to find the least squares estimators, it is not hard to guess what they are: for instance

$$\widehat{\beta}_1 = \frac{y_5 + y_6 + y_7 - y_2 - y_3 - y_4}{3}$$

with variance $6 \times \frac{\sigma^2}{9} = \frac{2}{3}\sigma^2$ (the fact that this particular estimator achieves the minimum variance is proof that it is the LSE, since the LSE is unique). With a little more algebra, one can derive this formula directly from $(X^TX)^{-1}X^TY$, but I won't show that here.

Part B

1. The relevant formulae are those on page 115 of the text (replacing $x_{i1},\ x_{i2}$ everywhere by $x_{i1} - \bar{x}_{\cdot 1},\ x_{i2} - \bar{x}_{\cdot 2}$). Thus $\Delta = 2309 \times 9800.354 - (204.42)^2 = 22589190$ and

$$\begin{aligned} \widehat{\beta}_1 &= \frac{239.9}{30} = 9.797, \\ \widehat{\beta}_2 &= \frac{1137.46 \times 9800.354 - 659.4343 \times 204.42}{22589190} = .4875 \\ \widehat{\beta}_3 &= \frac{-1137.46 \times 204.42 + 659.4343 \times 2309.2}{22589190} = .05712. \end{aligned}$$

2. Using $\sum (y_i - \bar{y})^2 = 1846.67$, we have

$$s^2 = \frac{1846.67 - .4875 \times 1137.46 - .05712 \times 659.4343}{27} = 46.46 = (6.816)^2.$$

3. The diagonal entries of $(X^TX)^{-1}$ are $\frac{1}{30} = .03333$, $\frac{9800.354}{22589190} = .0004339$ and $\frac{2309.2}{22589190} = .0001022$ so the standard errors of $\widehat{\beta}_j$, $j = 1, 2, 3$ are $s\sqrt{.03333} = 1.244$, $s\sqrt{.0004339} = .1420$, $s\sqrt{.0001022} = .0689$.

4. The t ratios (estimate divided by standard error) for β_2 and β_3 are $\frac{.4875}{.1420} = 3.43$ and $\frac{.05712}{.06892} = .83$. Without detailed checking of t tables, one can see that β_2 is significant and β_3 is not, i.e. the change in blood pressure is affected by age but not by weight.

5. The critical value for h_{ii} is $\frac{2p}{n} = 0.2$ (here $p = 3$, $n = 30$) which is exceeded by observations 6, 11, 12, 28, i.e. these are the points of high leverage. The values of the studentized residuals that are greater than 2 in magnitude are observations 23 and 26, i.e. these are the possible outliers. For DFFITS, the critical value is $2\sqrt{\frac{p}{n}} = .632$, and this is exceeded by observations 6, 23 and 26. It looks as though observation 6 is influential because it combines high leverage with a moderately large studentized residual (1.66), while observations 23 and 26 are influential because they have the largest residuals in magnitude, as well as being of moderately high leverage (h_{ii} between .1 and .2).