STAT 350: 97-1

Final Exam, 8 April 1997 — Instructor: Richard Lockhart

Instructions: This is an open book test. You may use notes, text, other books and a calculator. Your presentations of statistical analysis will be marked for clarity of explanation. I expect you to explain what assumptions you are making and to comment if those assumptions seem unreasonable. The exam is out of 60.

1. When a weight is hung from a wire, the wire stretches (returning to its original length when the weight is removed). A 1 kilogram weight is hung from a piece of wire and the length stretched is measured. This is repeated and the two resulting lengths are $L_{1,1}$ and $L_{1,2}$. Then a 2 kilogram weight is tried 3 times resulting in lengths $L_{2,1}$, $L_{2,2}$ and $L_{2,3}$. To save analysis effort the experimenter averages the two measurements made with the 1 kilogram weight, obtaining $Y_1 = (L_{1,1} + L_{1,2})/2$ and the 3 measurements made with the two kilogram weight, obtaining Y_2. [Total of 20 marks]

   (a) Assume that the individual lengths satisfy, for i from 1 to 2 and j from 1 to 2 (for $i = 1$) or 1 to 3 (for $i = 2$),

   $$L_{i,j} = x_{i,j}\beta + \epsilon_{i,j}$$

   where the errors $\epsilon_{i,j}$ are independent normal variables and have mean 0 and variance σ^2. What is the design matrix for this linear model? [2 marks]

   (b) Give an explicit, simple, formula for the least squares estimate of β; I do not want a general formula such as $(X^T X)^{-1} X^T Y$. [4 marks]

   (c) Give the mean and variance of the estimator in (b). [2 marks]

   (d) The average measurements Y_i also satisfy a linear model

   $$Y_i = x_i\gamma + \epsilon_i$$

   i. What is γ in terms of β? HINT: What is $E(Y_i)$? [1 marks]

   ii. What is the joint distribution of (ϵ_1, ϵ_2)? In particular what are the variances and means of each ϵ_i? [3 marks]

   iii. What is the design matrix of this linear model? [1 marks]

   (e) Show that the weighted least squares estimate of γ is

   $$\hat{\gamma} = (Y_1 + 3Y_2)/7$$

   [4 marks]

   (f) What is the distribution of $\hat{\gamma}$. [2 marks]

   (g) Why would analysis of the original variables $L_{i,j}$ be better than analysis of the Y_i? [1 mark]

2. A variable Y (a measurement of oxygen taken up by a system) is regressed on 4 predictors $X_1, \ldots, X_4$. A total of 20 measurements were made and Y was regressed on various subsets of the predictor variables leading to the following table of Error Sums of Squares.

Vars	ESS	Vars	ESS	Vars	ESS	Vars	ESS
X_1	154	X_1, X_2	109	X_2, X_4	133	X_1, X_3, X_4	139
X_2	156	X_1, X_3	144	X_3, X_4	175	X_2, X_3, X_4	132
X_3	203	X_1, X_4	146	X_1, X_2, X_3	106	All	104
X_4	250	X_2, X_3	150	X_1, X_2, X_4	107	None	506

(a) Does adding the variables X_3 and X_4 to the model containing X_1 and X_2 significantly improve the fit? [6 marks]

(b) Use Backwards selection with a 10% significance level to stay to select a suitable subset of regression variables. [8 marks]

(c) If the estimated slope associated with X_1 in the model including X_1 and X_2 only as predictors is positive what is the value of the t statistics for testing the hypothesis that the true coefficient of X_1 is 0? [1 mark]

3. Five different treatments, A, B, C, D and E, are to be examined for their effect on blood pressure. Fifty patients are randomly split into 5 groups of 10. The initial blood pressure X of each patient is measured, the treatment is applied and then the final blood pressure Y is measured. Let $i = 1, \ldots, 5$ label the treatment and j running from 1 to 10 label the patient within the treatment group. Three models were fitted:

Model I

$$Y_{i,j} = \alpha + \beta X_{i,j} + \epsilon_{i,j}$$

the error sum of squares is 85355 and the estimates are

$\hat{\alpha}$	$\hat{\beta}$
37.26	0.65

Model II

$$Y_{i,j} = \mu_i + \beta X_{i,j} + \epsilon_{i,j}$$

the error sum of squares is 66115 and the estimates are

$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$	$\hat{\beta}$
14.2424	67.5325	48.3918	49.6033	68.7786	0.5509

For this model

$$(X^T X)^{-1} = \begin{bmatrix} 1.026 & 0.995 & 0.924 & 0.927 & 0.929 & -0.00784 \\ & 1.168 & 0.993 & 0.995 & 0.998 & -0.00842 \\ & & 1.023 & 0.925 & 0.927 & -0.00782 \\ & & & 1.028 & 0.930 & -0.00784 \\ & & & & 1.032 & -0.00786 \\ & & & & & 6.63 \times 10^{-5} \end{bmatrix}$$

Model III

$$Y_{i,j} = \mu_i + \beta_i X_{i,j} + \epsilon_{i,j}$$

the error sum of squares is 62433 and the estimates are

$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$
52.04954	-68.05918	62.48453	46.66416	112.529
$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
0.2309892	1.619385	0.4313726	0.5757114	0.1818949

(a) Of the three models, based on the information available to you, which model provides the best fit to the data. [10 marks]

(b) There are 10 possible comparisons between pairs of treatments. It is desired to give simultaneous 95% confidence intervals for all possible comparisons based based on the second model above. I want you to show clearly that you know how to get these ten confidence intervals. Your answer will include a clear description of the parameters for which intervals are needed, written in terms of the notation used above for the second model and the resulting confidence interval for the difference between treatment A and treatment B with all the numbers filled in. You need not work it out to the point of a numerical value for the lower and upper limit. [5 marks]

(c) Examine the residual plots attached for the three models. Is there anything wrong with our fit? If so suggest what you might try next. Be quite clear. [5 marks]

(d) I attach a table of regression diagnostics for the fit to model II above. For each diagnostic review the values and comment on whether or not they show any problems and which cases might warrant further examination. [5 marks]

Diagnostics for Model II for Question 3

Obs #	h_{ii}	Ext'ly Stud'zed Residual	DFFITS	Cooks D_i	Obs #	h_{ii}	Ext'ly Stud'zed Residual	DFFITS	Cooks D_i
1	0.120	-0.777	-0.287	0.014	26	0.100	-0.096	-0.032	0.000
2	0.108	0.407	0.142	0.003	27	0.100	2.018	0.674	0.071
3	0.129	0.047	0.018	0.000	28	0.158	0.768	0.333	0.019
4	0.103	0.868	0.295	0.015	29	0.104	-0.475	-0.162	0.004
5	0.101	0.141	0.047	0.000	30	0.106	-0.997	-0.343	0.020
6	0.124	-0.377	-0.142	0.003	31	0.102	-1.133	-0.383	0.024
7	0.102	0.681	0.229	0.009	32	0.144	-0.139	-0.057	0.001
8	0.150	-0.578	-0.243	0.010	33	0.154	-0.201	-0.086	0.001
9	0.148	-0.180	-0.075	0.001	34	0.103	1.186	0.401	0.027
10	0.127	-0.261	-0.099	0.002	35	0.137	-0.009	-0.004	0.000
11	0.121	1.073	0.398	0.026	36	0.134	0.607	0.238	0.010
12	0.100	-1.076	-0.359	0.021	37	0.114	0.184	0.066	0.001
13	0.102	-0.179	-0.060	0.001	38	0.101	0.069	0.023	0.000
14	0.130	0.329	0.127	0.003	39	0.109	0.372	0.130	0.003
15	0.106	3.436	1.186	0.188	40	0.101	-0.934	-0.312	0.016
16	0.180	-0.613	-0.288	0.014	41	0.115	-2.130	-0.766	0.091
17	0.104	-0.306	-0.104	0.002	42	0.146	-0.732	-0.303	0.015
18	0.100	0.516	0.172	0.005	43	0.126	1.295	0.491	0.040
19	0.110	-1.138	-0.401	0.027	44	0.101	3.038	1.016	0.145
20	0.110	-1.742	-0.611	0.059	45	0.107	-1.635	-0.565	0.051
21	0.117	0.211	0.076	0.001	46	0.148	1.019	0.425	0.030
22	0.130	0.385	0.149	0.004	47	0.115	0.417	0.150	0.004
23	0.152	-0.699	-0.296	0.015	48	0.103	0.105	0.036	0.000
24	0.111	-0.320	-0.113	0.002	49	0.143	-0.911	-0.372	0.023
25	0.104	-0.715	-0.243	0.010	50	0.142	-0.333	-0.135	0.003