STAT 350: 99-1

Final Exam, 6 April 1999 Instructor: Richard Lockhart

Instructions: This is an open book test. You should have a total of 6 pages. 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. Three shipments of glass parts are transported. One shipment is transferred once, one twice and the other three times. The number of broken parts, Y, in each shipment is recorded. If x is the number of transfers it is thought reasonable to suppose that the Y_i are independent with the mean, μ_i, of Y_i being given by $\mu_i = \beta x_i$.

   The data are

x	1	2	3
Y	1	3	8

   (a) A preliminary estimate of β is obtained by ordinary least squares. This estimate has the form $a_1 Y_1 + a_2 Y_2 + a_3 Y_3$. Derive formulas for the a_i and evaluate the estimate for the data given. [4 marks]

   (b) From now on assume that the Y_i have Poisson distributions so that the variance of Y_i is equal to its mean. Compute the mean and variance of the estimate in part a). If you couldn't do part a) you may assume (incorrectly) that $a_i = i$ for $i = 1, 2, 3$. [3 marks]

   (c) Use the estimate of β from part a) to compute estimates of the variances of Y_1, Y_2 and Y_3. If you couldn't do a) you may assume (incorrectly) that $\hat{\beta} = 2$. [3 marks]

   (d) Refit the model in (a) using weighted least squares with weights derived in (c). If you couldn't do c) you may assume (incorrectly) that the weights are 1, 2 and 4. [5 marks]

   (e) Treating the weights computed in (c) as non-random compute an estimated standard error for the estimate of β computed in (d). Your answer should include a theoretical standard error which will depend on the true value of β and an estimate of this standard error for the data given. [5 marks]

2. A group of 30 children is split at random into 2 groups of 15. Each of the children is given a reading comprehension test generating a baseline score U. Each child in one group of 15 is encouraged to play a new game which is supposed to improve reading skills. After 6 months of exposure to the new game all the children are re-tested to produce a final reading comprehension score W. Here are the data:

Game (Treatment)			No Game (Control)		
Child	U	W	Child	U	W
1	88	97	16	114	107
2	114	133	17	90	88
3	77	79	18	119	114
4	99	112	19	105	104
5	98	109	20	96	94
6	134	146	21	118	112
7	77	93	22	138	145
8	89	102	23	87	92
9	120	136	24	101	102
10	70	86	25	63	69
11	91	107	26	124	132
12	114	122	27	120	116
13	99	113	28	101	98
14	109	125	29	97	103
15	72	80	30	85	76

The goal of the experiment was to decide whether or not playing the game improved reading comprehension and if so, how much. In order to answer the question three models relating the post-test score W to the game playing and pre-test score U were considered. In model I only U affects W. Model II has additive effects for U and game playing status. Model III includes, in addition to the effects in Model II, an interaction term.

(a) Write out model equations for the responses of the first child in the treatment group and the first child in the control group for EACH of the three models. [5 marks]

(b) Appendix A contains SAS output for models I, II and III. Use the output to decide which model fit is best. [5 marks]

(c) Why did the experimenter make the pre-test measurements? [2 marks]

(d) In view of the output provided was the decision to make pre-test measurements a good one? [2 marks]

3. Measurements are made on the beaks of squid. A total of 5 different lengths are measured on the beak of each squid. These 5 measurements, X_1, X_2, X_3, X_4, X_5 are to be used to predict the weight Y of the squid. Here is a table of error sums of squares for the regression of Y on each possible subset of the predictors.

ESS	Predictors	ESS	Predictors
8.907300	X1 X2 X3 X4 X5	8.985575	X1 X2 X4 X5
9.206036	X2 X3 X4 X5	9.271223	X2 X4 X5
9.776069	X1 X3 X4 X5	9.804839	X1 X4 X5
9.889998	X1 X2 X3 X5	9.944055	X3 X4 X5
9.969040	X4 X5	9.972450	X1 X2 X5
10.172300	X1 X3 X5	10.172384	X1 X5
12.037964	X2 X3 X5	12.082856	X2 X5
12.287947	X3 X5	12.756941	X5
13.259535	X1 X2 X3 X4	13.523033	X1 X2 X3
14.242817	X1 X3 X4	14.244054	X1 X3
15.801798	X2 X3 X4	16.370593	X3 X4
17.629214	X1 X2 X4	17.642615	X1 X2
17.696791	X1 X4	17.769277	X1
18.980140	X2 X3	19.563138	X3
23.450476	X2 X4	25.663950	X4
26.635009	X2	215.92475	None

(a) Carry out forward variable selection using the significance level 0.05 for variables to enter. [10 marks]

(b) Here is a table of regression diagnostics.

OBS	X1	X2	X3	X4	X5	Y	$\hat{Y}$	$\hat{\epsilon}$	COOK	h_{ii}	PRESS	EXTST	DFFITS
1	1.31	1.07	0.44	0.75	0.35	1.95	2.194	-0.2444	0.0035	0.1320	-0.2816	-0.3627	-0.1414
2	1.55	1.49	0.53	0.90	0.47	2.90	3.860	-0.9598	0.9244	0.5647	-2.2047	-2.3390	-2.6639
3	0.99	0.84	0.34	0.57	0.32	0.72	0.787	-0.0669	0.0010	0.3098	-0.0970	-0.1109	-0.0743
4	0.99	0.83	0.34	0.54	0.27	0.81	-0.051	0.8611	0.0491	0.1441	1.0061	1.3576	0.5571
5	1.05	0.90	0.36	0.64	0.30	1.09	0.810	0.2801	0.0074	0.1855	0.3439	0.4298	0.2051
6	1.09	0.93	0.42	0.61	0.31	1.22	0.920	0.2996	0.0091	0.1949	0.3721	0.4629	0.2278
7	1.08	0.90	0.40	0.51	0.31	1.02	0.444	0.5757	0.1161	0.3888	0.9418	1.0501	0.8374
8	1.27	1.08	0.44	0.77	0.34	1.93	2.037	-0.1068	0.0007	0.1387	-0.1240	-0.1586	-0.0636
9	0.99	0.85	0.36	0.56	0.29	0.64	0.317	0.3229	0.0066	0.1395	0.3753	0.4829	0.1944
10	1.34	1.13	0.45	0.77	0.37	2.08	2.450	-0.3703	0.0056	0.0984	-0.4107	-0.5420	-0.1791
11	1.30	1.10	0.45	0.76	0.38	1.98	2.573	-0.5930	0.0090	0.0662	-0.6350	-0.8655	-0.2303
12	1.33	1.10	0.48	0.77	0.38	1.90	2.760	-0.8603	0.0516	0.1497	-1.0117	-1.3611	-0.5711
13	1.86	1.47	0.60	1.01	0.65	8.56	7.889	0.6706	0.4320	0.5578	1.5164	1.4868	1.6698
14	1.58	1.34	0.52	0.95	0.50	4.49	5.136	-0.6457	0.0361	0.1751	-0.7827	-1.0114	-0.4659
15	1.97	1.59	0.67	1.20	0.59	8.49	7.961	0.5290	0.1484	0.4596	0.9789	1.0246	0.9449
16	1.80	1.56	0.66	1.02	0.59	6.17	6.778	-0.6077	0.0335	0.1809	-0.7419	-0.9516	-0.4472
17	1.75	1.58	0.63	1.09	0.59	7.54	6.890	0.6505	0.0341	0.1662	0.7801	1.0135	0.4525
18	1.72	1.43	0.64	1.02	0.63	6.36	7.621	-1.2610	0.3421	0.3068	-1.8193	-2.4738	-1.6459
19	1.68	1.57	0.72	0.96	0.68	7.63	7.639	-0.0091	0.0001	0.6111	-0.0233	-0.0200	-0.0251
20	1.75	1.59	0.68	1.08	0.62	7.78	7.359	0.4205	0.0198	0.2084	0.5312	0.6600	0.3386
21	2.19	1.86	0.75	1.24	0.72	10.15	9.689	0.4610	0.0686	0.3748	0.7373	0.8202	0.6351
22	1.73	1.67	0.64	1.14	0.55	6.88	6.226	0.6541	0.2107	0.4471	1.1830	1.2747	1.1462

Analyze these diagnostics, identifying influential observations, possible outliers and explaining for each point identified which diagnostic makes it important and what the diagnostic measures. Your answer should look at each diagnostic, identify the most important cases and then discuss whether or not the diagnostic is big enough to demand further study. [NOTE: the column labeled COOK contains values of Cook's distance. The column labeled EXTST contains what I called externally studentized residuals or what the text calls a studntized deleted residual.] [8 marks]

4. Suppose U_1, U_2, U_3, U_4 are independent random variables and that $U_i \sim N(\beta i, \sigma^2)$. (That is, the mean of U_i is proportional to i.)

   (a) If $\mathbf{U}$ is the vector of length 4 whose entries are the U_i then we can write $\mathbf{U} = A\mathbf{Z}+b$ where $\mathbf{Z}$ is a standard multivariate normal, A is a constant matrix and b a vector of constants. What are A and b? [4 marks]

   (b) Define $Y_i = U_{i+1} - U_i$ for $i = 1, 2, 3$. What is distribution of the vector $\mathbf{Y}$ whose entries are Y_1, Y_2, Y_3? [4 marks]

Appendix A: SAS Input for Reading Comprehension Problem

```
data reading;
 infile 'reading.dat';
 input U W GAME $ ;
proc glm data=reading;
 class GAME;
 model W = GAME  ;
run;
proc glm data=reading;
 class GAME;
 model W = U  ;
run;
proc glm data=reading;
 class GAME;
 model W = U GAME ;
run;
proc glm data=reading;
 class GAME;
 model W = U | GAME  ;
run;
```

SAS output for the 4 models

```
                    Class Level Information
                   Class    Levels    Values
                   GAME          2    No Yes
            Number of observations in data set = 30
Dependent Variable: W
Source                  DF    Sum of Squares  F Value    Pr > F
Model                    1      258.13333333     0.65    0.4280
Error                   28    11173.06666667
Corrected Total         29    11431.20000000
                  R-Square              C.V.             W Mean
                  0.022581          18.77438         106.400000
Source                  DF         Type I SS  F Value    Pr > F
GAME                     1      258.13333333     0.65    0.4280
Source                  DF       Type III SS  F Value    Pr > F
GAME                     1      258.13333333     0.65    0.4280
 **************************************************************
                       MODEL I
Dependent Variable: W
Source                  DF    Sum of Squares  F Value    Pr > F
Model                    1     9477.13657305   135.80    0.0001
Error                   28     1954.06342695
```

Corrected Total	29	11431.20000000		
	R-Square	C.V.		W Mean
	0.829059	7.851429		106.400000
Source	DF	Type I SS	F Value	Pr > F
U	1	9477.13657305	135.80	0.0001
Source	DF	Type III SS	F Value	Pr > F
U	1	9477.13657305	135.80	0.0001

****

MODEL II

Dependent Variable: W

Source	DF	Sum of Squares	F Value	Pr > F
Model	2	10732.6143410	207.41	0.0001
Error	27	698.5856590		
Corrected Total	29	11431.2000000		
	R-Square	C.V.		W Mean
	0.938888	4.780643		106.400000
Source	DF	Type I SS	F Value	Pr > F
U	1	9477.13657305	366.29	0.0001
GAME	1	1255.47776796	48.52	0.0001
Source	DF	Type III SS	F Value	Pr > F
U	1	10474.4810077	404.83	0.0001
GAME	1	1255.4777680	48.52	0.0001

****

MODEL III

Dependent Variable: W

Source	DF	Sum of Squares	F Value	Pr > F
Model	3	10745.2702287	135.77	0.0001
Error	26	685.9297713		
Corrected Total	29	11431.2000000		
	R-Square	C.V.		W Mean
	0.939995	4.827380		106.400000
Source	DF	Type I SS	F Value	Pr > F
U	1	9477.13657305	359.23	0.0001
GAME	1	1255.47776796	47.59	0.0001
U*GAME	1	12.65588765	0.48	0.4947
Source	DF	Type III SS	F Value	Pr > F
U	1	10476.5001595	397.11	0.0001
GAME	1	8.7044814	0.33	0.5706
U*GAME	1	12.6558877	0.48	0.4947