Name: Student Number:

STAT 350: Summer Semester 2008

Midterm 2

14 July 2008 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. In general you need not finish doing arithmetic; I will be satisfied if your answers contain things like

$$27 \pm 1.96\sqrt{247.5/11},$$

but I have to be absolutely convinced you know what arithmetic to do! I want the answers written on the paper. The exam is out of 25. Write on the back if extra space is needed.

Appendix 1 gives data on 38 car models. (The data are quite old – 1978-79.) The response variable Y of interest is `MPG` (miles per gallon) and in this problem you will consider various ways of relating this to the other variables in the data set. In what follows X_1 is Weight, X_2 is Drive Ratio, X_3 is Horsepower and X_4 is Displacement.

1. Begin by considering the full model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_i$$

Appendix 1 gives a table of regression diagnostics: DFFITS, Cook's Distance, Leverages, Fitted Residuals and externally studentized residuals. For each diagnostic tell me if there is any case which is highlighted for attention. Then identify at most two particular cases which deserve further scrutiny. Your answer MUST include explanation in full sentences. Anything less is going to get 0. [5 marks]

The rule of thumb for DFFITS says to compare them to 1 or to $2\sqrt{p/n}$ in large data sets. This data set is not very large so I will just look for DFFITS more than 1. Only Case 1 stands out; DFFITs is very large so deletion of this case changes the fitted value for this case greatly.

The rule of thumb for Cook's distance asks us to compare it to the median of the $F_{5,33}$ distribution and to the lower 10% point of the same distribution. The median will be near 1 (it is actually 0.89). The lower 10% point is 1 over the upper 10% point of the $F_{33,5}$ distribution. From the tables the latter is between 3.14 and 3.17 (which correspond to 30 and 60 numerator df) and probably pretty close to 3.17. So the cutoff I will use is 0.315. Only Case 1 stands out as a problem.

The rule of thumb for Leverages involves looking for those over 0.5 or over 10/38=0.26 or over 0.2. None are over 0.5. Case 12 is 0.3 – over the 0.26 mark and Cases 4, 15 and 29 are over 0.2. Case 12 is the only one which seems worth looking at.

The raw residuals can't really be used easily. They need to be standardized.

The externally studentized, or case deleted standardized residuals show Case 1 is huge.

I would have a close look at Case 1 before doing anything else. That is the case I eliminated for the next bits.

2. I eliminated 1 data point from the data set (based on my investigation in the previous question) and refitted the model. In Appendix 1 I produce plots of the residuals against fitted values, residuals squared against fitted values and a QQ plot of the residuals. Criticize the model fit using these plots. [2 marks]

   The QQ plot is fine – quite straight. I see a fairly clear trend to increasing variability with fitted value in the plot of $\hat{\epsilon}^2$ against $\hat{\mu}$. I think I see some curvature in the top plot, too. I might want to try some polynomial regression model (or transform the response variable to improve the heteroscedasticity).

3. In Canada it is normal to measure gasoline economy of a car using litres per hundred kilometers. This variable LPHK is 235.2/MPG. For this part we take $Y_i = LPHK = 235.2/MPG$ and consider the full model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_i$$

   Use the tables of error sums of squares to eliminate unnecessary independent variables 1 at a time to settle on a final model. Use F tests at the level 0.05. Explain what you are doing in full sentences. [8 marks]

   You would want to eliminate a variable which had the smallest possible F statistic or equivalently the least possible change in ESS from the full model. So we try to eliminate the variable X_4. We test $H_o : \beta_4 = 0$ and compute

$$F = \frac{(14.3009 - 14.1186)/1}{14.1186/33} = 0.426$$

   which is not significant. (The relevant degrees of freedom are 1, 33.)

   Now we consider the model without Displacement as the full model. The model containing only X_1 and X_2 has the least ESS so the smallest F statistic so we test $H_o : \beta_3 = 0$. You may either use the ESS from the original full model or the ESS for the model containing predictors 1, 2 and 3 in the denominator. The F statistic is then either

$$F = \frac{(15.3141 - 14.3009)/1}{14.1186/33} = 2.37$$

   or

$$F = \frac{(15.3141 - 14.3009)/1}{14.3009/34} = 2.41.$$

The corresponding critical values are 4.14 and 4.13 at the 5% level so neither is significant. In the tables you would have to use 30 or 60 denominator degrees of freedom anyway – but the answer is the same. Not significant so out it goes.

I would eliminate Horsepower and reconsider. The next biggest F statistic would consider eliminating Drive Ratio. We would test $H_o : \beta_2 = 0$. *This F statistic has numerator*

$$(35.8309 - 15.3141)/1 = 20.5168.$$

The denominator would be either 14.1186/33 or 15.3141/35. Using the former gives

$$F = 47.95$$

and the latter 46.89. Either is far larger than the 0.05 critical points of 4.14 and 4.12 respectively. This hypothesis is rejected and we stop here ending with Weight and Drive Ratio in the model.

DISCUSSION: The choice between denominators reflects a trade-off. If we made a mistake in accepting one of the null hypotheses then our MSE for the model with the corresponding variable eliminated is biased. Its expected value is $\sigma 2$ *plus a bit which depends on the corresponding* β. *This would increase the denominator in future F tests and decrease power. But: if the accepted null hypothesis is true the estimate is not biased and our estimate of* σ^2 *is based on 1 more degree of freedom so it is a bit more precise. You see we get slightly smaller critical values and so we tend to gain power. In general the trade off is usually best made by using the denominator from the very first (the one with all 4 independent variables) model unless there are very few degrees of freedom for error in that model.*

4. Suppose now that a new categorical variable called Region is to be created based on country of origin. It will have 3 levels: US, Japan, Europe. (Other than the US and Japan all the countries are in Europe.) Consider an additive model in which the effects $\alpha_U, \alpha_J, \alpha_E$ of this categorical variable (the subscripts match United States, Japan and Europe) are subject to the constraint $\alpha_U + \alpha_J + \alpha_E = 0$. Adding your variable to the model you settled on in the previous part and using the constraint to eliminate α_J from the model equations what is the row of the design matrix corresponding to a Datsun 510? [2 marks]

We have $\alpha_J = -\alpha_U - \alpha_E$. *The model equation for the Datsun 510 is*

$$\begin{aligned} 27.2 &= \beta_0 + 2.300\beta_1 + 3.54\beta_2 + 97\beta_3 + 119\beta_4 + \alpha_J + \epsilon \\ &= \beta_0 + 2.300\beta_1 + 3.54\beta_2 + 97\beta_3 + 119\beta_4 - \alpha_U - \alpha_E + \epsilon \end{aligned}$$

If we keep the variables in the order in that formula we get

$$1 \quad 2.300 \quad 3.54 \quad 97 \quad 119 \quad -1 \quad -1$$

for the row in question. I asked for you to include only the variables in your final model so really the terms for β_3 *and* β_4 *should not be included:*

$$1 \quad 2.300 \quad 3.54 \quad -1 \quad -1$$

5. If an interaction is added between Region and the first variable in your model what is the row of the design matrix corresponding to a Datsun 510? [2 marks]

   Now we add two columns – multiply the column for Weight (or some other variable if you picked some other final model earlier) by the last two columns. The new row is

$$1 \quad 2.300 \quad 3.54 \quad -1 \quad -1 \quad -2.300 \quad -2.300$$

6. I now fit the linear model including the 4 variables $X_1, \ldots, X_4$ and add Region. I find the error sum of squares is 12.4501. Should I retain Region in this model? [4 marks]

   I am testing the hypothesis $H_o : \alpha_U = \alpha_J = \alpha_E$ and the resulting F test has 2 numerator and 31 denominator degrees of freedom. We get

$$F = \frac{(14.1186 - 12.4501)/2}{12.4501/31} = 2.08$$

   The corresponding $F_{2,31}$ critical value is 3.30 so the null hypothesis is accepted and we need not retain Region in the model.

7. Consider now three statisticians about to fit this last model. One imposes the constraint above. A second puts $\alpha_U = 0$. The third puts $\alpha_E = 0$. All three estimate $\alpha_U - \alpha_J$. Do they get different values? Explain. I only want a very brief explanation. [2 marks]

   No. They all give identical estimates of this difference in intercepts. I have not really provided clear evidence of this fact but the idea is that whichever constraint you impose changes the meaning of the other αs. Putting $alpha_U = 0$ in the model formula makes β_0 be the intercept for US cars. Then $\beta_0 + \alpha_J$ is the intercept for Japanese cars and the difference in intercepts is $\beta_0 - (\beta_0 + \alpha_J) = -\alpha_J = 0 - \alpha_J = \alpha_U - \alpha_J$. If we put $\alpha_E = 0$ the intercepts are $\beta_0 + \alpha_U$ and $\beta_0 + \alpha_J$ whose difference is again the parameter in question. The sum constraint makes the average of the three intercepts equal to β_0 and the three intercepts equal to $\beta_0 + \alpha_U$, $\beta_0 + \alpha_E$, and $\beta_0 + \alpha_J = \beta_0 - \alpha_U - \alpha_E$. The three design matrices all have the same column space so they must give the same least squares fitted values including the same predicted value at any new x.

Appendix 1: Car Data

The Data

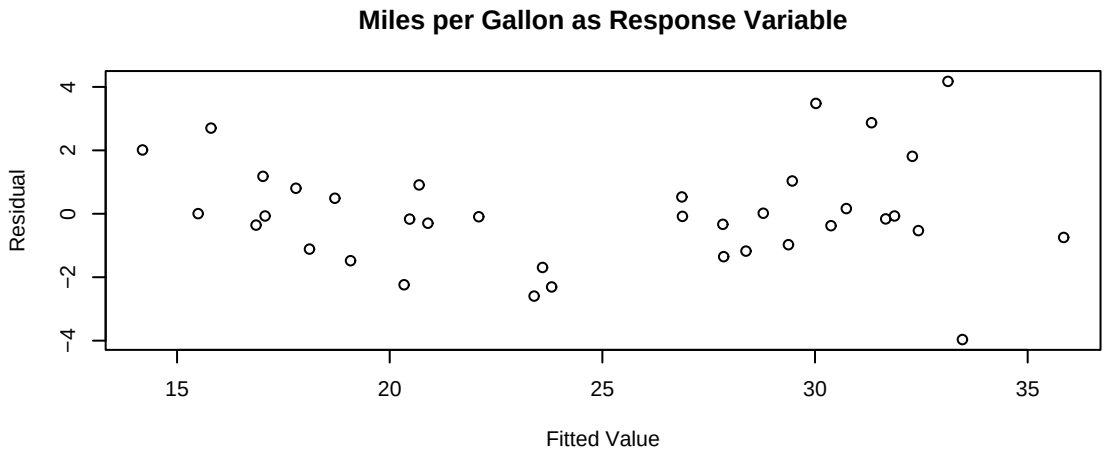
Country	Car	MPG	Weight	Drive Ratio	Horsepower	Displacement	Cylinders
U.S.	Buick Estate Wagon	16.9	4.360	2.73	155	350	8
U.S.	Ford Country Squire Wagon	15.5	4.054	2.26	142	351	8
U.S.	Chevy Malibu Wagon	19.2	3.605	2.56	125	267	8
U.S.	Chrysler LeBaron Wagon	18.5	3.940	2.45	150	360	8
U.S.	Chevette	30.0	2.155	3.70	68	98	4
Japan	Toyota Corona	27.5	2.560	3.05	95	134	4
Japan	Datsun 510	27.2	2.300	3.54	97	119	4
U.S.	Dodge Omni	30.9	2.230	3.37	75	105	4
Germany	Audi 5000	20.3	2.830	3.90	103	131	5
Sweden	Volvo 240 GL	17.0	3.140	3.50	125	163	6
Sweden	Saab 99 GLE	21.6	2.795	3.77	115	121	4
France	Peugeot 694 SL	16.2	3.410	3.58	133	163	6
U.S.	Buick Century Special	20.6	3.380	2.73	105	231	6
U.S.	Mercury Zephyr	20.8	3.070	3.08	85	200	6
U.S.	Dodge Aspen	18.6	3.620	2.71	110	225	6
U.S.	AMC Concord D/L	18.1	3.410	2.73	120	258	6
U.S.	Chevy Caprice Classic	17.0	3.840	2.41	130	305	8
U.S.	Ford LTD	17.6	3.725	2.26	129	302	8
U.S.	Mercury Grand Marquis	16.5	3.955	2.26	138	351	8
U.S.	Dodge St Regis	18.2	3.830	2.45	135	318	8
U.S.	Ford Mustang 4	26.5	2.585	3.08	88	140	4
U.S.	Ford Mustang Ghia	21.9	2.910	3.08	109	171	6
Japan	Mazda GLC	34.1	1.975	3.73	65	86	4
Japan	Dodge Colt	35.1	1.915	2.97	80	98	4
U.S.	AMC Spirit	27.4	2.670	3.08	80	121	4
Germany	VW Scirocco	31.5	1.990	3.78	71	89	4
Japan	Honda Accord LX	29.5	2.135	3.05	68	98	4
U.S.	Buick Skylark	28.4	2.670	2.53	90	151	4
U.S.	Chevy Citation	28.8	2.595	2.69	115	173	6
U.S.	Olds Omega	26.8	2.700	2.84	115	173	6
U.S.	Pontiac Phoenix	33.5	2.556	2.69	90	151	4
U.S.	Plymouth Horizon	34.2	2.200	3.37	70	105	4
Japan	Datsun 210	31.8	2.020	3.70	65	85	4
Italy	Fiat Strada	37.3	2.130	3.10	69	91	4
Germany	VW Dasher	30.5	2.190	3.70	78	97	4
Japan	Datsun 810	22.0	2.815	3.70	97	146	6
Germany	BMW 320i	21.5	2.600	3.64	110	121	4
Germany	VW Rabbit	31.9	1.925	3.78	71	89	4

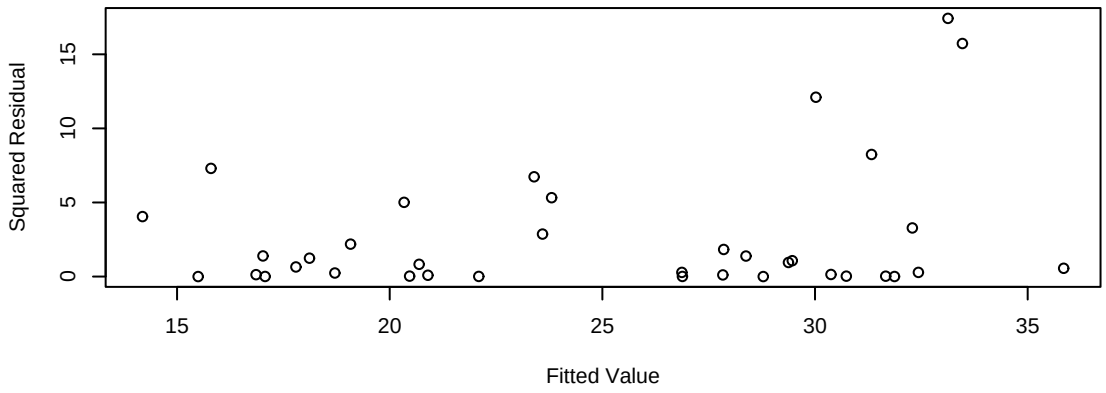
Models and Sums of Squares

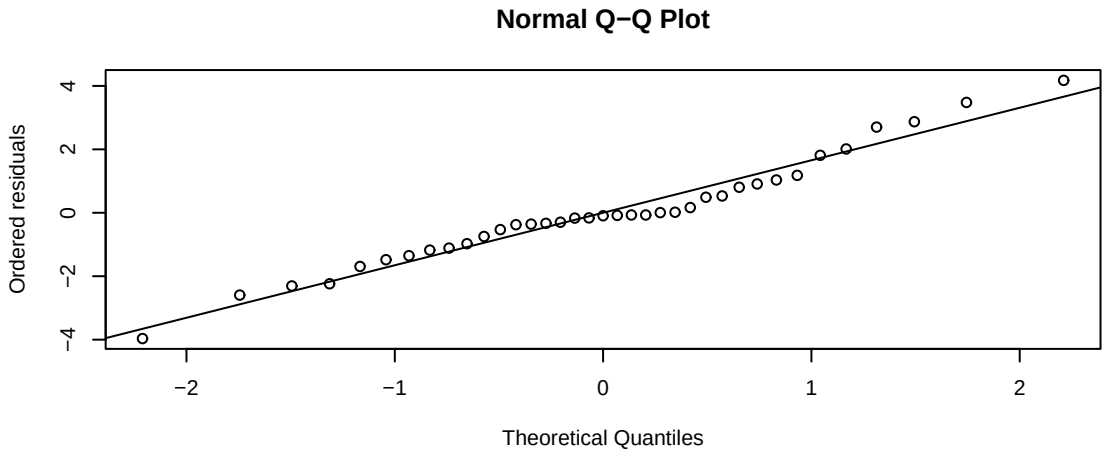
Predictors used	Error SS	Predictors used	Error SS
X_1,X_2,X_3,X_4	14.1186	X_1,X_2,X_3	14.3009
X_1,X_2,X_4	15.2223	X_1,X_3,X_4	23.1960
X_2,X_3,X_4	36.7869	X_1,X_2	15.3141
X_1,X_3	33.1432	X_1,X_4	25.8751
X_2,X_3	54.2240	X_2,X_4	53.4491
X_3,X_4	52.4630	X_1	35.8309
X_2	208.1485	X_3	56.4908
X_4	87.8645	None	259.2315

Residuals and Diagnostics

Case #	DFFITS	Cook's Distance	Leverages	$\hat{\epsilon}$	Externally Studentized Residuals
1	1.8693	0.4873	0.1857	6.2918	3.9143
2	-0.1853	0.0070	0.1387	-0.9251	-0.4617
3	-0.0062	0.0000	0.0639	-0.0497	-0.0237
4	0.5138	0.0532	0.2571	1.6115	0.8734
5	-0.0689	0.0010	0.1163	-0.3864	-0.1899
6	-0.0167	0.0001	0.0621	-0.1363	-0.0650
7	-0.2056	0.0086	0.1075	-1.2058	-0.5925
8	0.0422	0.0004	0.0546	0.3694	0.1755
9	-0.1624	0.0054	0.1476	-0.7786	-0.3902
10	-0.3681	0.0273	0.1576	-1.6729	-0.8509
11	0.1147	0.0027	0.1749	0.4899	0.2492
12	0.4768	0.0462	0.3063	1.2842	0.7175
13	-0.1125	0.0026	0.0976	-0.7028	-0.3422
14	-0.7482	0.1069	0.1801	-3.0131	-1.5962
15	0.1023	0.0022	0.2267	0.3596	0.1889
16	-0.3185	0.0198	0.0517	-2.7971	-1.3640
17	-0.1207	0.0030	0.0905	-0.7887	-0.3827
18	-0.3262	0.0213	0.0977	-2.0087	-0.9911
19	-0.2777	0.0157	0.1609	-1.2506	-0.6342
20	0.0588	0.0007	0.1007	0.3611	0.1759
21	-0.1267	0.0033	0.0448	-1.2319	-0.5850
22	-0.1917	0.0074	0.0443	-1.8623	-0.8903
23	0.3674	0.0271	0.1280	1.9128	0.9591
24	-0.0078	0.0000	0.1880	-0.0318	-0.0163
25	0.1576	0.0051	0.1588	0.7193	0.3628
26	-0.0231	0.0001	0.1285	-0.1214	-0.0601
27	-0.5998	0.0677	0.1056	-3.4167	-1.7455
28	-0.1167	0.0028	0.1831	-0.4821	-0.2465
29	0.0882	0.0016	0.2376	0.2988	0.1581
30	0.0064	0.0000	0.1529	0.0299	0.0150
31	0.6976	0.0891	0.1072	3.8817	2.0131
32	0.4275	0.0351	0.0723	3.0839	1.5313
33	0.0079	0.0000	0.1117	0.0455	0.0223
34	0.8505	0.1245	0.1023	4.7242	2.5198
35	0.1414	0.0041	0.0759	1.0235	0.4934
36	-0.1002	0.0021	0.0994	-0.6190	-0.3015
37	-0.4952	0.0481	0.1286	-2.5416	-1.2892
38	-0.0989	0.0020	0.1527	-0.4642	-0.2330


Miles per Gallon as Response Variable
Residual
Fitted Value


Squared Residual
Fitted Value


Normal Q−Q Plot
Ordered residuals
Theoretical Quantiles