STAT 350

Assignment 3

1. The following table gives the carbon monoxide emission rates in grams per mile for two vehicles. Readings were taken approximately every 1000 miles for each car.

-2 VEHICLE 1 -4		VEHICLE 2	
Mileage	Emission Rate	Mileage	Emission Rate
0	50	0	40
1000	56	1100	49
2000	58	2200	58
3000	60	3000	65
4200	58	4000	75
5000	63	5300	77
6000	73	6000	86
6900	71	7000	93
8000	76	8100	98
9200	73	9000	103
10000	80	10000	109

(a) Plot the data.

(b) Consider the following 4 models for the data:

i. Two straight lines, one for each vehicle, with different slopes and intercepts,

ii. Two parallel straight lines.

iii. Two lines with the same intercept but different slopes.

iv. One straight line.

Write out model equations for data points number 2 and 22 for the first 3 models. To be clear for the fourth model these equations would be

$$Y_2 = \beta_1 + 1000\beta_2 + \epsilon_2$$

and

$$Y_{22} = \beta_1 + 10000\beta_2 + \epsilon_{22}\,.$$

Other models would have different numbers of parameters of course.

(c) Fit all 4 models. Hand in: estimates of the slopes and intercepts and of σ. Do NOT just hand in output from SAS or Minitab.

(d) Using formal hypothesis tests and plots select the best of these models. Again, I do not want computer output but discussion. You may attach computer output in order to say things like: "The Sum of Squares for ... is on page ..." and to hand in plots on which you comment in the discussion but I will not be looking through the output.

(e) For the final selected model estimate the total emissions of CO in grams for each vehicle over the first 10000 miles. (This is the area under the fitted straight line from 0 to 10000 and is a linear combination of the parameter estimates.) Attach a standard error.

(f) Are the emissions of the two vehicles different over the first 10000 miles?

(g) Suppose the two cars are of the same make but that one vehicle was equipped with a special pollution control device. In 4 or 5 sentences comment on the experimental design as a method of determining whether or not the device reduces emissions and on what else you would want to find out from the experimenter to help interpret the results.

Some Help: Some of the models have design matrices which do not naturally have a column of ones. To fit these in SAS you will need to add `/ NOINT` to the end of the model statement. So, for example, to fit a straight line relating emissions RATE and MILEAGE which passed through the origin you might use the statements

```
proc glm
 model RATE = MILEAGE / NOINT ;
```

You will also need to create one or more data files for SAS to read. The table above is available in the assignment lab. For some models you will have to create columns of the design matrix yourself, using some text editor such as Microsoft Word or whatever. You will have to use the model equations from part 2 to see what goes in the columns of the design matrix and then create a data set which has these columns in it.

You may also want to use `proc gplot` to do the plots since these plots are much higher resolution. This procedure is just like `proc plot` but produces better graphs. Here is some SAS example code:

```
data insure;
  infile 'insure.dat' firstobs=2;
  input year cost;
  code = year - 1975.5 ;
proc glm  data=insure;
   model cost = code ;
   estimate 'fit1980' intercept 1 code  4.5  / E;
   output out=insfit p=fitted r=resid student=isr press=press rstudent=e
run ;
proc rank data=insfit out=qqdat normal=blom ;
   var resid;
   ranks nscores;
run;
proc gplot data=qqdat;
   plot resid*nscores;
run;
```

The output is The output statement produces internally standardized residuals is isr, press residuals in press and externally studentized residuals in esr. The purpose of `proc rank` is to compute the plotting points for a Q-Q plot and store the residuals together with the corresponding plotting points in a data set called qqdat. Then `proc gplot` plots them. You can use the on-line help for `proc gplot` to find out how to customize the axes.

In MINITAB, too, it is possible to ask for no intercept in a regression model.

2. Data below are from a nitrogen balance experiment on Kangaroo Island Wallabies, taken from Barker,S. (1968). "Nitrogen balance and Water Intake in the Kangaroo Island Wallaby" *Austral. J. Experimental Biology and Medical Science*, **46**, 17-32.

Y Nitrogen Excreted	X_1 Body Weight	X_2 Dry Intake	X_3 Water Intake	X_4 Nitrogen Intake
162	3.386	16.6	41.7	54
174	3.033	18.1	40.9	99
119	3.477	13.4	25.0	46
205	3.278	22.6	39.2	188
312	3.368	26.5	47.4	345
157	2.932	21.4	51.6	66
184	3.128	30.3	71.6	171
155	3.251	17.6	27.1	81
192	3.396	21.3	37.7	175
331	3.497	29.9	50.5	399
114	3.182	12.8	28.4	38
159	3.234	19.6	34.3	106
260	3.139	36.2	77.6	228
265	3.434	35.0	58.9	291
387	2.970	32.9	55.3	449
146	3.230	22.9	46.2	72
233	3.470	32.9	67.4	176
261	3.000	35.7	77.1	235
287	3.224	34.4	74.9	288
412	3.366	36.2	60.7	485
174	3.264	29.9	65.4	92
171	3.292	21.7	51.2	126
259	3.525	35.0	66.8	224
298	3.036	29.7	65.8	276
407	3.356	29.2	48.1	386

Fit the model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

by least squares. Get estimates and standard errors for all the parameters and an estimate of σ. Suggest a simpler model for the data, and fit it. Check the fit of the model, graphically and, if the model seems poor, modify it appropriately. Hand in a dsicussion of your findings bolstered by output used only as an appendix. I will be marking the discussion, not sorting through the output.