Mathematics & Statistics 523B Assignment #2
Due Monday 3 March 2003 in class

Notes:

- This Assignment comprises 7 questions on 5 pages.
- The data are available in R format with the extension `.dput` or in text format with the extension `.dat` with the first row containing the variable names.
- Data files in R format can be read in using the `dget` command and `attach`'ed as in Assignment 1.
- If you use the R format, be aware that the **factors may not be coded as such**. You may have to modify them appropriately.

1. The data in `malepigs.*` and `fempigs.*` are from an experimental piggery arranged for individual feeding of six pigs in each of five pens (covariate `pen`). From each of five litter, six young pigs, three males and three females, were selected and allotted to one of the pens. Three feeding treatments denoted by 1, 2, 3 (covariate `food`), containing increasing proportions ($p_1 < p_2 < p_3$) of protein, were used and each given to one male and one female in each pen. The pigs were individually weighed each week for 16 weeks. For each pig, the growth rate in pounds per week (covariate `growth`) was calculated as the slope of a line fitted by least-squares. The weight of each pig at the beginning of the experiment is also included (covariate `weight`). The variables in each of the files are summarized in the following table:

Column	Var. name	Factor/ Continous	Description	Values
Col. 1	`food`:	Factor	Type of food	1,2,3
Col. 2	`pen`:	Factor	Pen number	1,2,3,4,5
Col. 3	`growth`:	Continuous	Average growth rate	(in lbs/wk)
Col. 4	`weight`:	Continuous	Original weight	(in lbs)

The male students should use `malepigs.*` and the female students `fempigs.*`. Use the data to show that

(a) `pen` and `food` are orthogonal for the outcome `growth` with respect to the intercept.

(b) `pen` and `food` are not orthogonal for the outcome `growth` with respect to both the intercept and `weight`.

2. We wish to choose between the two models

$$\begin{aligned} M_1 : \mathbb{E}[Y_i] &= x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p \\ M_2 : \mathbb{E}[Y_i] &= x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + x_{i,p+1}\beta_{p+1} \end{aligned}$$

for $i = 1, \ldots, n$, where the $x_{i,j}$ and β_j, $j = 1, \ldots, p+1$, are respectively covariates and unknown scalars.
Show that the model chosen by the criterion $\widehat{\text{MSEP}}_3$ is approximately the same as that which is chosen by a t-test of the hypothesis

$$H_0 : \beta_{p+1} = 0$$

at the $\sqrt{2}$ critical value (i.e., we choose M_1 if $|t| \leq \sqrt{2}$ and M_2 if $|t| > \sqrt{2}$, for $t = \hat{\beta}_{p+1}/\widehat{\text{s.e.}}(\hat{\beta}_{p+1})$).

3. Show that $\widehat{\text{MSEP}}_2$ (PRESS) is an upwardly biased estimator of MSEP.

4. The file `cars.*` contains data on cars that were advertised for sale in a French newspaper in September 1985. There are 164 observations, each containing 4 variables:

Column	Var. name	Factor/ Continous	Description	Values
Col. 1	`type`:	Factor	Type of car	1=Peugeot 104
				2= Citroën 2CV
				3= Peugeot 304/305
				4= Renault 4
				5= Renault 5
				6= Peugeot 504/505
				7= Renault 18
Col. 2	`year`:	Continuous	Year of car	66, 67,. . . ,85
Col. 3	`kilo`:	Continuous	Kilometrage	(in 1000km)
Col. 4	`price`:	Continuous	Asking price	(in 1000F)

We are interested in the relation between price and the other variables.

(a) Make a plot of price against year. Does there seem to be a strong linear relationship?

(b) Calculate the variable `logpr<-log(price)` and produce a plot of `logpr` against `year`. Notice the improvement in linearity. Test that there is a significant linear relationship between `logpr` and `year`.

(c) Is it possible that there are different basic prices for each type of car? Fit a model with different intercepts according to type, and test the hypothesis

that these intercepts are really different, assuming (for the moment) that the linear relationship with year is the same for each type.

(d) Do the different types of car depreciate in value at the same rate? Test the hypothesis that the linear effect of year is the same for all types of car, or, if you like, that there is an interaction between type and year. Which type of car depreciates the most? the least?

(e) Is there an effect of `kilo` on `logpr` allowing for `type`, `year` and their interaction?

(f) Is there an interaction between `kilo` and `type`, allowing for the variables fitted in part **(e)**?

5. **(a)** Among the models that you have fitted in Question **4.** (and any others that you think are reasonable), which is the best from the point of view of prediction? Use the $\widehat{\text{MSEP}}_3$ (Mallow's C_p) criterion.

(b) A particular Renault 5, having logged 45,000 km, was purchased in France in 1979 for 16,000F. Use the model chosen in part **(a)** to predict the asking price for such a car.

6. Let the model M_0 be given by

$$M_0 : \mathbb{E}[\mathbf{Y}] = X\boldsymbol{\beta}$$

where $\mathbf{Y}$ is an $n \times 1$ random column vector, X is a $n \times (n-1)$ real matrix with $\text{rank}(X) = n - 1$ and $\boldsymbol{\beta}$ is an $(n-1) \times 1$ vector of parameters.
Define the k^{th} standardized residual to be

$$\hat{Z}_k = \frac{Y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}_0^2(1 - \mathbf{x}_k'(X'X)^{-1}\mathbf{x}_k)}}$$

where $\hat{\boldsymbol{\beta}}$ is the least-squares estimator of β under M_0, x_i' is a vector of covariates corresponding to the i^{th} row of X and $\hat{\sigma}_0^2 = \text{SSE}_0/[n - (n-1)] = \text{SSE}_0$.
Show that $\hat{Z}_k$ =-1, 0 or 1. (This is an extreme example of the non-normality and non-t-ness of $\hat{Z}_k$.

Hint: You may wish to use the following two facts:

- Let $\mathbf{R}$ and $\mathbf{H}$ be two $p \times p$ matrices such that
  i. $\mathbf{R}$ and $\mathbf{H}$ are symmetric and idempotent, i.e. $\mathbf{R} = \mathbf{R}' = \mathbf{R}^2$ and $\mathbf{H} = \mathbf{H}' = \mathbf{H}^2$;
  ii. $\mathbf{RH} = 0$;

  then $\text{rank}(\mathbf{R}) + \text{rank}(\mathbf{H}) = \text{rank}(\mathbf{R} + \mathbf{H})$.
- If $\mathbf{R}$ is a $p \times p$ symmetric, non-negative definite matrix with rank $q \leq p$, then there exists a $p \times p$ matrix $\mathbf{A}$ with rank q such that $\mathbf{R} = \mathbf{AA}'$.

7. Strikes in OECD countries

The data frame `strikes` consists of annual observations on the level of strike volume (labour days lost due to industrial disputes per 1000 wage salary earners), and covariates in 18 Organization for Economic Cooperation and Development (OECD) countries from 1951 to 1985. The data are described below.

Variable name	Variable values	Description Description
Country	factor, 18 levels	country name
Year	continuous, ¿0	year minus 1950 (between 1 and 35)
Strvol	continuous, =0	strike volume in workdays per 1000 wage earners
Unemp	continuous, =100, =0	unemployment rate (%)
Inflat	continuous, =0	Yearly inflation rate (%)
Socdem	continuous, =100, =0	Social democratic parliamentary representation (%)
Union	continuous, =100, =0	Union centralization index (fixed by country)

SOURCE: Western, B. (1996). Vague theory and model uncertainty in macrosociology. *Sociological Methodology* **26**: 165–192.

There are 625 instead of 18 x 35 = 630 observations because years 1981 to 1985 are missing for Belgium. Fit the outcome variable `logsv=log(strvol+1)` (not in the data frame), under a Normal error model with common variance.

(a) Consider a model involving only the countries as an explanatory variable. Test that countries have different expected `logsv`.

(b) The following code allows you to obtain the estimated correlation matrix of the estimated coefficients:

```
correl<-function(model) {
cormat<-summary(model)$cov.scaled
secoef<-sqrt(diag(cormat))
cormat<-sweep(sweep(cormat,1,secoef,FUN="/"),2,secoef,FUN="/")
return(cormat)}
```

Type in this code; you can now, from a `glm` object named, e.g., `foo`, extract the correlation matrix of the estimates by typing `correl(foo)`.

- **i.** In a model including only the `Country` levels, use this function to obtain the correlation matrix between the estimates.
- **ii.** Is the result that you obtained in **i.** a coincidence? Explain.
- **iii.** Fit a model as in **(a)** such that Australia is aliased into the intercept. Apply a suitable Bonferroni correction to determine which countries have significantly different `logsv` from Australia at the 5% level, allowing for all other countries. Argue the appropriateness of the Bonferroni correction and be clear about your method.

(c) Test for an effect of `year` in the presence of `Country`. What is your conclusion?

(d) Test for a country-specific effect of year, allowing for a distinct intercept for each country. What is your conclusion? In one sentence, reconcile your results from parts **(c)** and **(d)**.

(e) The continuous covariate `union` is fixed for each country, regardless of the year. Fit the models `logsv~Country+union` and `logsv~Country`. Explain the relationship between the residual sums of squares (deviances) of the two models. Bonus marks for the 3 shortest correct explanations.

(f) Consider the following R code (we assume that the data frame `strikes` is `attach`'ed):

```
model1<-glm(logsv~Country*year)
model2<-glm(logsv~Country*year+inflat)
strikes2<-data.frame(Country,year,logsv,fitted(model1),fitted(model2))
names(strikes2)<-c("Country","year","logsv","fit1","fit2")
attach(strikes2)
plot(year,logsv,type="n",xlab="Year",ylab="Log(strike volume+1)",
main="Log(Strike volume+1) against year, by unidentified OECD country,
observed and fitted")
for (i in 1:length(levels(Country))) {
    temp<-strikes2[Country==levels(Country)[i],]
    points(temp$year,temp$logsv,col=i)
    matlines(temp$year,temp[,4:5],type="l",lty=c(1,2),col=i)}
legend(1972.5,9,
c("Observed","Fitted Country*year","Fitted Country*year+inflat"),
pch=c(1,-1,-1),lty=c(0,1,2),bty="n")
```

Explain in a few words what is done by each command in this code, and explain what is performed overall by the code.

For the remainder of this question, we settle on the model given by `logsv~Country*year+inflat`.
Produce plots as required to answer the following questions.

(g) Do the data appear to have constant variance? Justify your answer.

(h) Plot the residuals against the fitted values. A clear pattern appears in the lower left corner of the plot. Identify this pattern and explain its presence (briefly).

(i) Test the model for any outlier at the 5% level.

(j) Do the data appear to depart from Normality? In what manner?

(k) Test for Normality of the data.

End of Assignment 2.