MATH 523B Assignment #1
Due Monday 10 February 2003 in class

Note: This Assignment comprises 4 questions on 6 pages.

1. Consider the following data associated with the life cycle

$$\text{egg} \Rightarrow \text{caterpillar} \Rightarrow \text{pupa} \Rightarrow \text{adult} \Rightarrow \text{eggs laid by female}$$

of the Gypsy Moth *Lymantria dispar*.

Diet	a	a	b	b	e	e	a	b	h
Pupal weight (g)	1.23	1.31	1.03	1.07	2.38	2.45	1.40	1.01	0.41
Number of eggs	442	343	284	271	858	941	322	271	46
Diet	h	p	p	p	b	p	p	p	
Pupal weight (g)	0.43	1.42	1.70	1.12	0.863	1.307	1.007	1.461	
Number of eggs	71	400	677	427	244	511	385	543	

The "Pupal weight (g)" the female weight, in grams, at pupation. The diets fed to the caterpillars are coded as below

a	=	Willow Oak	h	=	Red Maple
b	=	Black Oak	p	=	Pine
e	=	Sweet Gum			

(a) Plot the Number of eggs against pupal weight.

(b) Fit a straight line $\mathbb{E}[Y_i] = \beta_1 + \beta_2 x_i$ to the data, where Y_i and x_i are Number of eggs and Pupal weight, respectively. Draw the line in your plot in **(a)**.

(c) Compute the residual sum of squares *SSE* and the coefficient of determination R^2 from part **(b)**.

(d) Write down formally the null and alternative hypotheses implied by the following questions. (Introduce notation as needed to do so.) Compute t- or F-statistics to test these alternatives in each case, and state your conclusions. Be sure to fully specify the distribution of the test statistic under the null hypothesis. Consider Number of eggs to be the response variable in all cases.

i. Disregarding the diet, is Number of eggs positively associated with Pupal weight?

ii. Does a Pine diet provide an additive effect to Pupal weight in determining the Number of eggs, with respect to any other diet?

iii. Are the slope and intercept of the line relating Number of eggs to Pupal weight the same under a Pine diet and any other diet?

(e) What assumptions about the Number of eggs should be made in order for the statistics in **(d)** to follow a t- or an F-distribution?

(f) Which t-statistic in **(d) i.** can be used directly to assess the magnitude of R^2? Write down a simple formula relating this t-statistics to R^2.

2. Consider the models

$$H_0 : \mathbb{E}\left[\mathbf{Y}_{n\times 1}\right] = \mathbf{Z}_{n\times q}\boldsymbol{\gamma}_{q\times 1}$$

and

$$H_1 : \mathbb{E}\left[\mathbf{Y}_{n\times 1}\right] = \mathbf{Z}_{n\times q}\boldsymbol{\gamma}_{q\times 1} + \mathbf{x}_{n\times 1}\beta_{1\times 1}$$

where the $\mathbf{x}$ is a column vector and β is a scalar. Let ρ be the uncentered correlation coefficient between $\mathbf{x}$ and $\mathbf{Y}$ with the $\mathbf{Z}$-effect removed, i.e.

$$\rho = \frac{\mathbf{x}^{*\prime}\mathbf{Y}^{*}}{\sqrt{(\mathbf{x}^{*\prime}\mathbf{x}^{*})(\mathbf{Y}^{*\prime}\mathbf{Y}^{*})}}$$

where

$$\begin{aligned} \mathbf{x}^{*} &= \mathbf{R}_{\mathbf{Z}}\mathbf{x} \\ \mathbf{Y}^{*} &= \mathbf{R}_{\mathbf{Z}}\mathbf{Y} \text{ with} \\ \mathbf{R}_{\mathbf{Z}} &= \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \end{aligned}$$

Show that

$$\rho^2 = R^2,$$

where

$$R^2 = \frac{\mathrm{SSE}_0 - \mathrm{SSE}_1}{\mathrm{SSE}_0}.$$

Here, SSE_i is the error sum of squares for H_i, $i = 0, 1$.

3. The file *accidoutre.dput*, available from the 189-523B Web site, contains monthly data collected between 1976 and 1978 on accidents in Outremont. The file contains the following variables:

Col. 1:	`N`=	Number of accidents during the month
Col. 2:	`maxtmp`=	Maximum daily temperature
Col. 3:	`mintmp`=	Minimum daily temperature
Col. 4:	`avgtmp`=	Average daily temperature
Col. 5:	`dew`=	Dew point (temperature at which condensation forms)
Col. 6:	`sun`=	Hours of sunshine during the month
Col. 7:	`rain`=	Millimetres of rain fallen during the month
Col. 8:	`snow`=	Centimetres of snow fallen during the month
Col. 9:	`days`=	Number of days in the month

We wish to explain the number of accidents as a function of the weather variables.

Read the data in R using the commands

```
accid<-dget("accidoutre.dput")
attach(accid)
```

`accid` is called a *data frame*. Models can now be fitted using the variable names, e.g. `mymodel<-glm(N~maxtmp+dew)`, etc.

(a) Fit a model with the number of accidents as the random variable and each of the weather variables as regressors, ignoring the effect of the others. Carry out suitable hypothesis tests. What do you conclude?

(b) Now test for the effect of each weather variable (in turn), allowing for an effect of the others. What do you conclude?

(c) Test for a simultaneous effect of all weather variables on the number of accidents. What do you conclude?

(d) Explain the discrepancies you see in your answers to **(a)**, **(b)** and **(c)**

4. The file *airq.dput*, available on the MATH 523B Web site, contains measurements relating to air quality taken close to New York City from May to September 1973. The variables considered are listed in the table below.

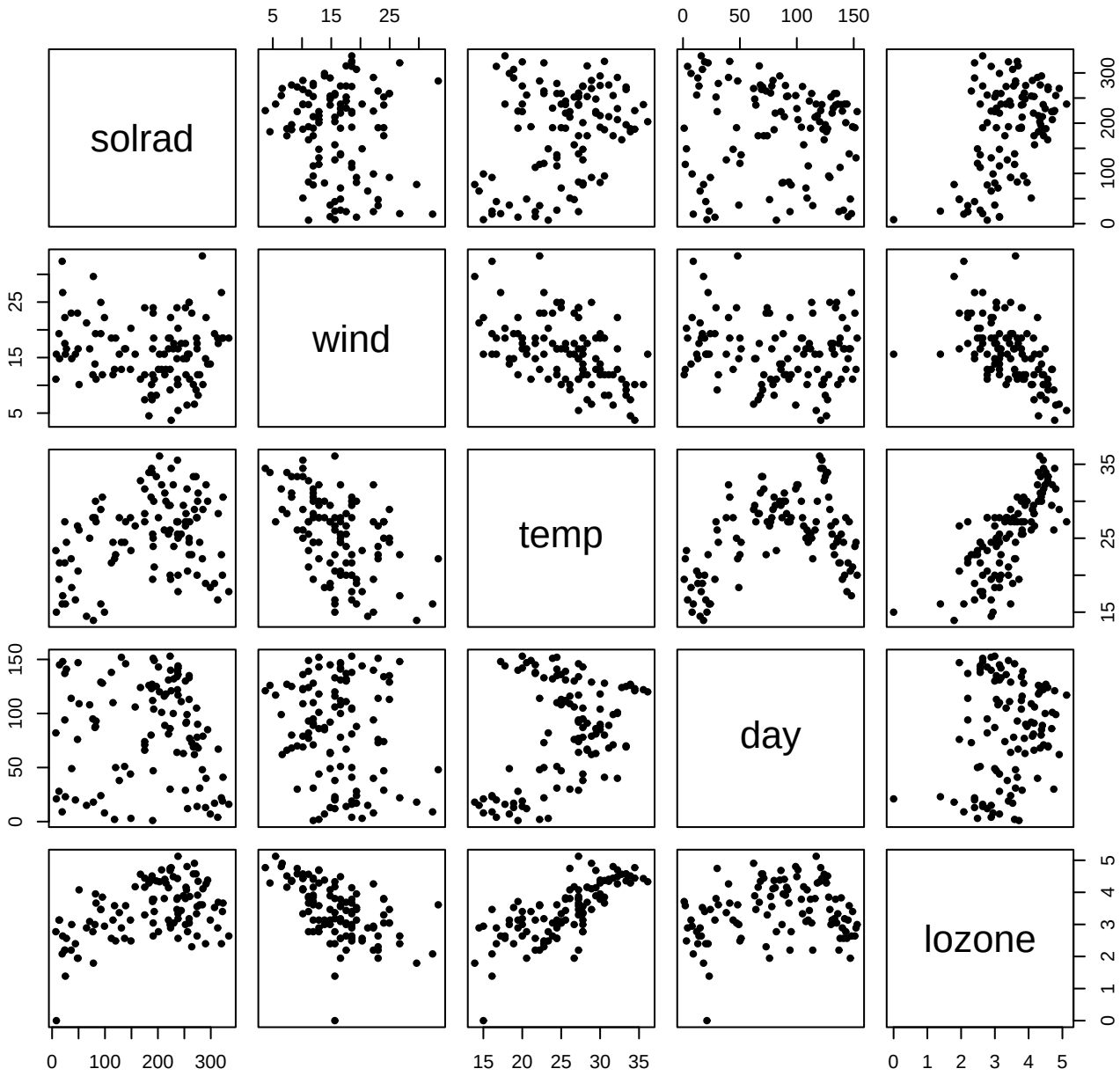
Variable	**Variable values**	**Description**
`lozone`	Continuous, ≥ 0	log(mean daily ozone ppb+1)
`day`	integer, > 1	day number, where 1=May 1 1973
`temp`	continuous	maximum daily Celsius temperature
`wind`	continuous, > 0	average daily wind speed in km/h
`solrad`	continuous, > 0	daily solar radiation in Langleys

SOURCE: New York State Department of Conservation and United States National Weather Service.

ppb: parts per billion

We analyze only the complete observations: there are 111 of them. We are interested in linear modelling of `lozone` using the other measurements. We assume that `lozone` is normally distributed with common variance and a mean determined by the models specified below.

Figure L007−1: New York City air quality pairwise scatterplots



(a) Consider the following models:

$$\begin{aligned} M_1 : \mathbb{E}\left[\texttt{lozone}_i\right] &= \alpha + \beta_{\mathrm{T}}\texttt{temp}_i \\ M_2 : \mathbb{E}\left[\texttt{lozone}_i\right] &= \alpha + \beta_{\mathrm{T}}\texttt{temp}_i + \beta_{\mathrm{W}}\texttt{wind}_i \end{aligned}$$

i. Obtain least-squares estimates for the unknown parameters for both models.

ii. Figure L007-1 displays pairwise scatterplots of all the variables indicated in the table above. Relate appropriate scatterplots from Figure L007-1 to the least-squares estimates of β_{T} from M_1 and from M_2. Explain the change in the estimate of β_{T} between M_1 and M_2 in terms of the scatterplots.

(b) Test the effect of `day` and $\texttt{day}^2$ simultaneously as explanatory variables for `lozone`.

Figure L007-2 enlarges the scatterplot of `lozone` against `day` of Figure L007-1. Consider this graph in answering parts **(c)** and **(d)**.

(c) Consider the models

$$\begin{aligned} M_3 : \mathbb{E}\left[\texttt{lozone}_i\right] &= \alpha + \beta_{\mathrm{D}}\texttt{day}_i \\ M_4 : \mathbb{E}\left[\texttt{lozone}_i\right] &= \alpha + \beta_{\mathrm{D2}}\texttt{day}_i^2 \end{aligned}$$

Find the least-squares estimates of β_{D} in M_3 and of β_{D2} in M_4. Explain in a single sentence both following phenomena evidenced by these fitted models (invoke features of the scatterplot of Figure L007-2):

i. Covariate `day` is only marginally significant by itself.

ii. Covariate $\texttt{day}^2$ is not significant by itself.

(d) The mean value of `day` is 84. Consider the following transformed covariate `cday=day-84`. Consider also the following models:

$$\begin{aligned} M_5 : \mathbb{E}\left[\texttt{lozone}_i\right] &= \alpha + \beta_{\mathrm{D}}\texttt{day}_i + \beta_{\mathrm{D2}}\texttt{day}_i^2 \\ M_6 : \mathbb{E}\left[\texttt{lozone}_i\right] &= \alpha + \beta_{\mathrm{CD}}\texttt{cday}_i + \beta_{\mathrm{CD2}}\texttt{cday}_i^2 \end{aligned}$$

Fit these models.

i. Explain why the residual sum of squares (residual deviance) is the same for models M_5 and M_6.

ii. Explain in 5 lines or less the changes in all three parameter estimates and in significance observed between models M_5 and M_6.

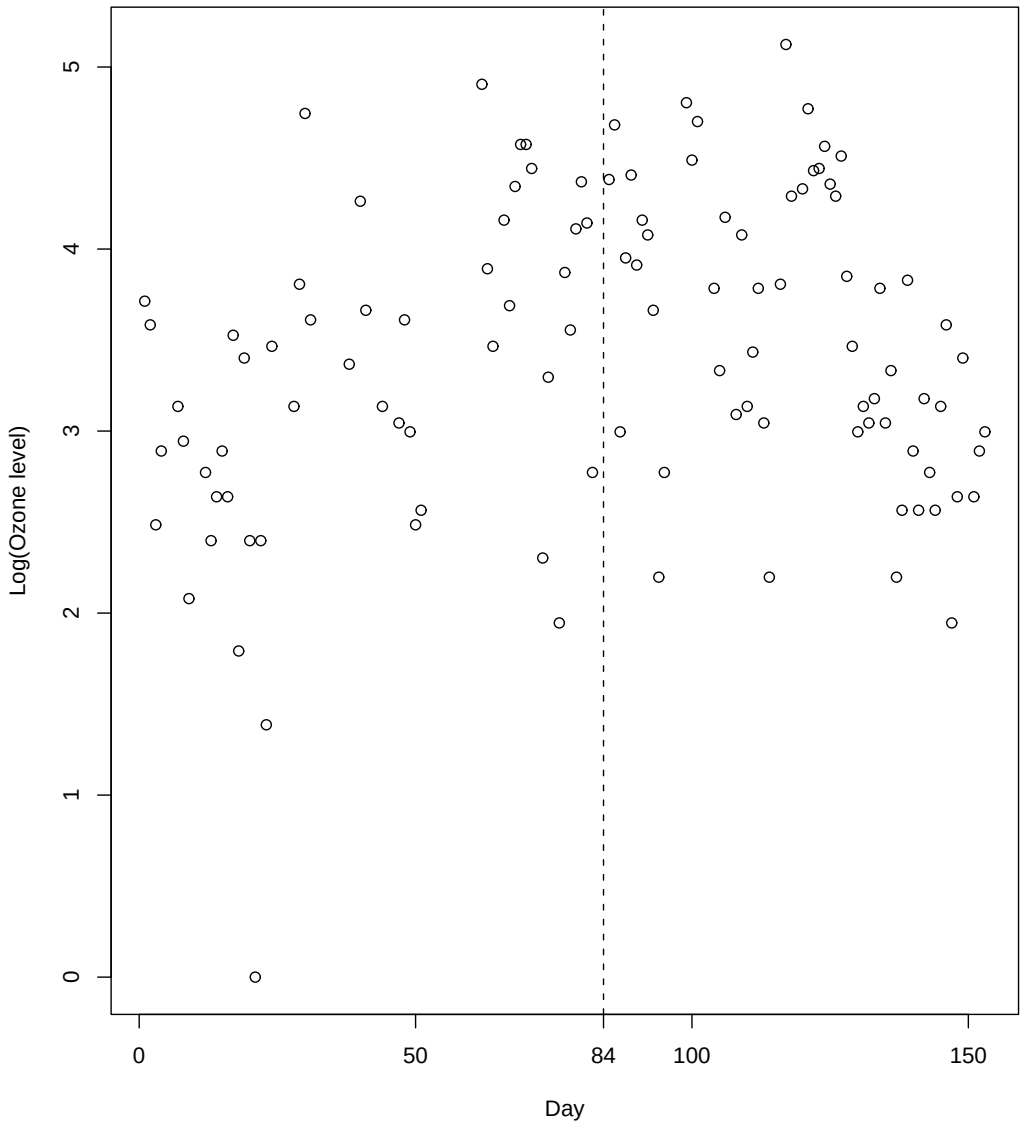
iii. Is the fact that 84 is the mean of covariate `day` relevant to these changes? Explain.

(e) We now compare models

$$\begin{aligned} M_7 : \mathbb{E}\left[\texttt{lozone}_i\right] &= \alpha + \beta_{\mathrm{S}}\texttt{solrad}_i + \beta_{\mathrm{W}}\texttt{wind}_i + \beta_{\mathrm{T}}\texttt{temp}_i \\ M_8 : \mathbb{E}\left[\texttt{lozone}_i\right] &= \alpha + \beta_{\mathrm{S}}\texttt{solrad}_i + \beta_{\mathrm{W}}\texttt{wind}_i + \beta_{\mathrm{T}}\texttt{temp}_i + \beta_{\mathrm{D}}\texttt{day}_i + \beta_{\mathrm{D2}}\texttt{day}_i^2 \end{aligned}$$

Explain the lack of significance of covariates `day` and `day`^{2} in the presence of the other terms in light of the results from model M_5. Refer to Figure L007-1 and to the nature of the data being modelled in your explanation.

Figure L007−2: NYC Log(Ozone) against Day



End of Assignment 1.