STA 442/2101F Homework 1. *due October 13 before 4 pm*

When answering questions requiring numerical work, the results are to be reported in a narrative summary, in your own words. Tables and Figures may be included, but must be formatted along with the text. DO NOT include in this summary printouts of computer code with the relevant selections highlighted.

All computer code used to obtain the results summarized in the response should be provided as an appendix. In this appendix you may highlight the relevant results.

1. **STA 442** Table 8.1 on p.355 of Davison, the `cement` data, gives a set of measurements taken on 13 samples of cement. The response is the heat per gram of cement, and the four explanatory variables relate to the chemical composition of the samples. The data is available on my web page, as `cement.dat`, which can be read into `R` using `read.table`.

   (a) Fit a linear regression model with y as the response variable and the 4 explanatory variables x_1, x_2, x_3 and x_4. Give a table of estimated coefficients and their estimated standard errors, and an analysis of variance table.

   (b) Construct a 95% confidence interval for β_1, and for $\beta_3 - \beta_2$.

   (c) Continue with Exercise 8.5.1 on p.385.

   (d) Exercise 8.3.1 suggests fitting a linear regression omitting x_4: why is this suggestion reasonable?

2. **STA 442 and 2101** Set 4 on p.170-1 of Cox & Snell, also available from my web page as the file `set4.R`,[1] gives data on several socio-economic variables in 47 states of the USA; the source given is Vandaele (1978). The variables in the data set are described in Table 4(b) on p. 171, which is reproduced on the last page.

   (a) Fit a linear regression model with crime rate as the response variable and the remaining 13 variables as explanatory variables. Use either backward or stepwise regression to eliminate variables, and present the estimated coefficients and standard errors for your final model.

   (b) What assumptions are implicit in the analysis in part (a)? Present three plots that provide information on the validity or otherwise of these assumptions.

   (c) If two explanatory variables are very highly correlated, then estimates of their corresponding coefficients can be ill-determined. Which pairs of explanatory variables might be expected to be highly correlated? Construct a `pairs` plot in `R` and summarize the information obtained.

[1] Which can be read into `R` using the `source` command: see an example in the Sept 15 code.

(d) Is there any evidence that an *increase* in per capita police expenditure between 1959 and 1960 had an effect on the crime rate?

(e) Provide a non-technical summary of at most one paragraph, suitable for publication in, for example, *The Varsity*, to explain the main findings of this study.

3. **STA 442 and 2101** Suppose we have n measurements on a a response y, two explanatory variables x and z, and we assume a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i, \quad i = 1, \ldots, n.$$

Further, assume that x is a continuous variate, but $z_i = \pm 1$; for example, x might be age, and z might be gender. We will also make the usual assumptions that the observations are independent, and that $\epsilon_i \sim (0, \sigma^2)$.

(a) Express this model as $y = X\beta + \epsilon$, where y and ϵ are $n \times 1$ vectors, making explicit the entries of the X matrix.

(b) Show that $E(y_i \mid x_i, z_i = -1)$ and $E(y_i \mid x_i, z_i = +1)$ are lines, and give their slopes and intercepts.

(c) What conclusions are implied by the hypothesis $\beta_3 = 0$? What conclusions are implied by the hypothesis $\beta_2 = 0$? What conclusions are implied by the hypothesis $\beta_3 = \beta_2 = 0$?

(d) Using the general result that the least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T y$, and the answer to (a), give explicit expressions for $\hat{\beta}_1$ and $\hat{\beta}_3$.

4. **STA 442 and 2101** In the general linear regression model $y = X\beta + \epsilon$,

(a) Show that under the assumption that ϵ has mean 0 and covariance matrix $\sigma^2 I$ where I is the $n \times n$ identity matrix, that the least squares estimator of β, $\hat{\beta} = (X^T X)^{-1} X^T y$ is unbiased for β and has variance-covariance matrix given by $\sigma^2 (X^T X)^{-1}$.

(b) Show that if instead ϵ has mean 0 and covariance matrix V, a positive definite symmetric $n \times n$ matrix, that $\hat{\beta}$ is still unbiased, and find an expression for its variance-covariance matrix.

5. **STA 2101**

(a) Still in the linear model (1), show that the hat matrix $H = X(X^T X)^{-1} X^T$ is symmetric and idempotent, i.e. $H^T = H$ and $H^2 = H$, and that $\mathrm{tr}(H) = p$. Show that $I - H$ is also symmetric and idempotent, and verify that

$$\mathrm{cov}(\hat{\epsilon}) = \mathrm{cov}(y - \hat{y}) = (I - H)\sigma^2.$$

(b) The estimator $\hat{\beta}_{-i}$ is the least squares estimator when the ith observation is omitted. Show that

$$\hat{\beta}_{-i} = \hat{\beta} - (1 - h_{ii})^{-1} (X^T X)^{-1} x_i (y_i - \hat{y}_i).$$

Writing $\hat{y}_{-i}$ for the predicted value of y_i based on $\hat{\beta}_{-i}$ show that

$$\frac{(\hat{y} - \hat{y}_{-i})^T(\hat{y} - \hat{y}_{-i})}{ps^2} = \frac{r_i^2 h_{ii}}{p(1 - h_{ii})}.$$

This quantity is usually denoted C_i and called *Cook's distance*; see §8.6.3 of Davison.

6. **C: STA 2101**[2] The data in Table 1 below gives the weight of 15 experimental animals over a five week period. A plot of the data for each rat is given in Figure 1. The goal of the data collection was to understand the typical growth, and the variability in growth, of this species, during their first five weeks of life. The animals are a random sample of a much larger group of newborns, all bred to parents supplied to the scientists from the same source.

	Week				
	1	2	3	4	5
1	151	199	246	283	320
2	145	199	249	293	354
3	147	214	263	312	328
4	155	200	237	272	297
5	135	188	230	280	323
6	159	210	252	298	331
7	141	189	231	275	305
8	159	201	248	297	338
9	177	236	285	340	376
11	160	208	261	313	352
12	143	188	220	273	314
13	154	200	244	289	325
14	171	221	270	326	358
15	163	216	242	281	312

Table 1: Weights of experimental animals, by week

(a) Denote by y_{jt} the response for animal j at week t. A simple starting model for this data is

$$y_{jt} = \beta_0 + \beta_1 t + \epsilon_{jt}, \quad t = 1, \ldots, 5; j = 1, \ldots, 15, \tag{1}$$

where we assume $\epsilon_{jt} \sim N(0, \sigma^2)$, and that the ϵ's are all independent. The least squares estimates of β_0 and β_1 are 112.63 (5.01) and 44.07 (1.51), respectively, where the estimated standard errors are given in parentheses. The residual sum of squares from this model is 24945, on 73 degrees of freedom.

It is suggested that it would be preferable to fit a model with a separate slope and intercept for each animal, to see if this fits the data better. The residual sum of squares after this exercise is 1949, on 45 degrees of freedom.

[2]Comp question from 2009

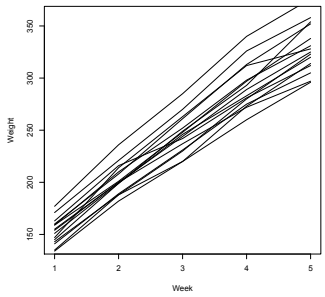

Figure 1: Plot of data in Table 1

i. Write down a formula similar to (1) for the model with separate slopes and intercepts.

ii. Explain why the residual degrees of freedom are 73, and 45, respectively, for the two model fits. Explain how to carry out a formal test of the significance comparing the two models.

iii. If you average the 15 slope estimates from the fit of the second model, will you get the slope estimate of 44.07 from the fit of (1)?

(b) Another way to model the data is to assume that the intercepts and slopes for each animal are random, and distributed around population average values. This could be modelled as

$$y_{jt} = \beta_0 + b_{j0} + (\beta_1 + b_{j1})t + \epsilon_{jt}, \quad t = 1, \ldots, 5; j = 1, \ldots, 15, \tag{2}$$

where β_0 and β_1 are as before, but now it is assumed that

$$b_{j0} \sim N(0, \sigma_0^2), \quad b_{j1} \sim N(0, \sigma_1^2),$$

and the b's are independently distributed and independent of the ϵs. When fit to this data, the maximum likelihood estimates of σ^2, σ_0^2 and σ_1^2 are $(6.58)^2$, $(10.92)^2$, and $(3.96)^2$, respectively. The estimates of β_0 and β_1 are the same as for (1).

Write a non-technical sentence explaining what these variance estimates mean in the context of this data.

(c) Model assessment:

i. How would you assess the appropriateness of a linear model (with or without random intercepts and slopes) for this data?

ii. How would you assess the appropriateness of the normality assumption for the within group errors ϵ_{jt}?

Table 4(b). Variables listed in Table 4(a)

	The source is the *Uniform Crime Report* of the Federal Bureau of Investigation. All the data relate to calendar year 1960 except when explicitly stated otherwise.
R:	Crime rate: the number of offences known to the police per 1000 000 population.
Age:	Age distribution: the number of males aged 14–24 per 1000 of total state population.
S:	Dummy variable distinguishing place of occurrence of the crime (south = 1). The southern states are: Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia.
Ed:	Educational level: the mean number of years of schooling $\times 10$ of the population, 25 years old and over.
Ex_0, Ex_1:	Police expenditure: the per capita expenditure on police protection by state and local government in 1960 and 1959, respectively. Sources used are *Governmental Finances in 1960* and *Governmental Finances in 1959*, published by the US Bureau of the Census.
LF:	Labour force participation rate per 1000 of civilian urban males in the age-group 14–24.
M:	The number of males per 1000 females.
N:	State population size in hundred thousands.
NW:	Nonwhites: the number of nonwhites per 1000.
U_1:	Unemployment rate of urban males per 1000 in the age-group 14–24, as measured by census estimate.
U_2:	Unemployment rate of urban males per 1000 in the age-group 35–39.
W:	Wealth as measured by the median value of transferable goods and assets or family income (unit 10 dollars).
X:	Income inequality: the number of families per 1000 earning below one-half of the median income.