Applied Statistics

Sheet 2 — MT 2023

This sheet is split into three sections:

- Section A: introductory question(s)
- Section B: core questions
- Section C: further question(s) – maybe some of these are slightly harder but all of them are on material that is part of the course

Undergraduates: only answers to Section B should be handed in for marking.

Solutions will be provided for all questions.

Neil Laws: neil.laws@stats.ox.ac.uk

Section A

1. Consider the linear model $y = X\beta + \epsilon$, with y an $n \times 1$ vector, X an $n \times p$ matrix of rank p (where $p < n$), β a $p \times 1$ vector, and ϵ an $n \times 1$ multivariate normal random vector $\epsilon \sim N(0, \sigma^2 I_n)$.

   The vector of residuals $e = (e_1, \ldots, e_n)^T$ is given by $e = y - X\widehat{\beta}$.

   Suppose the first column of X corresponds to an intercept term, so that $x_{i1} = 1$ for $i = 1, \ldots, n$. Show that $\sum_{i=1}^{n} e_i = 0$.

2. Three separate samples, one from each of four different suspensions of bacteria A, B, C and D (the 4 'treatments') were prepared. Technician I examined under a microscope one sample from each suspension in random order. Similarly the other samples were tested by Technicians II and III (the 3 'blocks').

   The recorded number of organisms from each sample are summarised in the following table. Interest principally centres on how suspensions A, B, C and D affect the recorded number of organisms.

		Suspensions			
		A	B	C	D
	I	67	84	77	60
Technician	II	71	77	74	70
	III	54	67	65	56

   (a) What is this type of experimental design called? Explain how to set up the data for analysis using the `lm()` function in R, treating the block and treatment label for each measurement as a categorical variable.

   (b) Use R to construct the analysis of variance table to test the hypothesis that there is no difference between the suspensions. Perform the test and state your conclusions.

   (c) In (ii) you were asked to do a two-way analysis of variance. Suppose the scientists had not recorded which technician made which count – so the data are now:

Suspensions	A	B	C	D
Counts	67 71 54	84 77 67	77 74 65	60 70 56

   Would you reach the same conclusion if you did a one-way analysis of variance? (i.e. ignoring the technician indicators).

Section B

3. Five treatments A, B, C, D, E, and a “no-treatment” treatment “S”, were applied to 24 apple trees to investigate the effect on apple yield. Crop sizes varied from tree to tree before any treatments were applied, so an extra variable, `History`, was included as a covariate; this represents the average volume of the crop from each tree over a 4-year period before the experiment started. Consider the normal linear model

$$y \sim 1 \ + \ \texttt{History} \ + \ \texttt{Treatment},$$

where y is the yield, `History` is a continuous variable and `Treatment` is categorical with six levels. In the regression “S” was used as the baseline level for `Treatment`.

Fitting the normal linear model above yielded the following results for the values, correlations and standard errors $s\sqrt{(X^TX)^{-1}_{j,j}}$ (notation as lectures) of the coefficient estimates $\widehat{\beta}_j$, $j = 1, \dots, 7$.

		Coefficient	Standard Error
Intercept	$\widehat{\beta}_1$	-26.94	44.74
History	$\widehat{\beta}_2$	32.95	4.50
A	$\widehat{\beta}_3$	33.01	22.78
B	$\widehat{\beta}_4$	19.55	22.86
C	$\widehat{\beta}_5$	27.05	22.86
D	$\widehat{\beta}_6$	36.06	23.29
E	$\widehat{\beta}_7$	57.89	23.91

```
Correlations of "beta-hat" coefficients
 Intercept  A     B     C     D     E     Hist
Int   1.00
A    -0.40  1.00
B    -0.42  0.52  1.00
C    -0.42  0.52  0.52  1.00
D    -0.49  0.52  0.52  0.52  1.00
E    -0.56  0.52  0.53  0.53  0.54  1.00
Hist -0.93  0.17  0.19  0.19  0.27  0.34  1.00
```

The total sum of squares (the RSS for the model $y \sim 1$) is 72,034. The sum of squares due to the covariate History (the difference in RSS's between the models $y \sim 1$ and $y \sim 1 + \text{History}$) is 48,413. The residual sum of squares for the full model above is 17,143.

(a) Carry out a test (F- or t-test) for the effect of History. Do you think its inclusion was necessary and, if so, why?

(b) Test the hypothesis that treatment E produces a higher yield than treatment S.

(c) Carry out a test of whether there is a difference between treatments D and E.
Hint: the entries in the correlation matrix $\rho_{i,j}$ are estimated by

$$\widehat{\rho}_{i,j} = s^2(X^TX)^{-1}_{ij} \Big/ \sqrt{s^2(X^TX)^{-1}_{ii} s^2(X^TX)^{-1}_{jj}}$$

(d) Carry out a test for an effect due to `Treatment`.

```
# Answer the questions above "by hand".
# Here is some R if you want to check.
a <- read.table("http://www.stats.ox.ac.uk/~laws/SB1/data/pearce.apple.txt",
                header = TRUE, stringsAsFactors = TRUE)
str(a)
head(a)
a$treatment <- relevel(a$treatment, ref = "S")
str(a)
a.lm <- lm(yield ~ history + treatment, data = a)
summary(a.lm)
anova(a.lm)
```

4. In a study investigating a new method of measuring body composition Mazess, R.B., Peppler, W.W. and Gibbons, M. (1984) gave the body fat percentage, age and sex for 18 normal adults aged between 23 and 61 years.

% fat	Age	Sex	% fat	Age	Sex	% fat	Age	Sex
9.5	23	M	25.9	41	F	32.5	56	F
7.8	27	M	25.2	49	F	30.3	57	F
17.8	27	M	31.1	50	F	33.0	58	F
27.4	45	M	34.7	53	F	33.8	58	F
27.9	23	F	42.0	53	F	41.1	60	F
31.4	39	F	29.1	54	F	34.5	61	F

The data are in the file `www.stats.ox.ac.uk/~laws/SB1/data/bodyfat.txt`.

Write down a normal linear model where the response variable is %fat and the explanatory variables are 'Age' and 'Sex'. Your model show allow the yearly change in %fat to differ for women and men. Give full mathematical details.

Does (a) body fat percentage increase as a function of age for women, and does (b) the yearly change in body fat percentage differ for men and women? Use R to answer these questions. State your conclusions with supporting hypothesis tests, and include your R code, with brief comments explaining what each line achieves. (Include your R code in your answer to this question, please don't email it in.)

5. When the k'th observation $(y_k, \mathbf{x}_k)$ is removed from a normal linear model $Y = X\beta + \epsilon$, the MLE parameter estimate $\widehat{\beta}_{-k}$ based on the reduced data is related to the MLE $\widehat{\beta} = (X^T X)^{-1} X^T y$ computed from the full data by

$$\widehat{\beta}_{-k} = \widehat{\beta} - (X^T X)^{-1} \mathbf{x}_k \frac{e_k}{1 - h_{kk}}$$

where $e_k = y_k - \widehat{y}_k$, $h_{kk} = \mathbf{x}_k^T (X^T X)^{-1} \mathbf{x}_k$ is the k'th leverage component, and $\widehat{y}_k = \mathbf{x}_k^T \widehat{\beta}$. [*You may assume this, you are not being asked to show it.*]

(a) Show that $y_k - \mathbf{x}_k^T \widehat{\beta}_{-k} = e_k/(1 - h_{kk})$.

(b) Show that $\mathrm{var}(y_k - \mathbf{x}_k^T \widehat{\beta}_{-k}) = \sigma^2/(1 - h_{kk})$.

(c) Let s_{-k} be the residual standard error in the analysis with $(y_k, \mathbf{x}_k)$ omitted. Define the studentised residuals r'_k, for $k = 1, \ldots, n$, and show they are given by

$$r'_k = (1 - h_{kk})^{1/2} \left(\frac{y_k - \mathbf{x}_k^T \widehat{\beta}_{-k}}{s_{-k}} \right).$$

(d) Show that $r'_k \sim t(n - p - 1)$ (assume without proof that $\widehat{\beta}$ and s^2 are independent in the primary fit). We often check normal qqplots for r'. Why do we compare r' to the order statistics of the $N(0, 1)$ distribution?

(e) We plot r' against $\widehat{y}$ and look for evidence of correlation. We have seen that $e = y - \widehat{y}$ and $\widehat{y}$ are independent. Show that under the normal linear model r' and $\widehat{y}$ are independent.

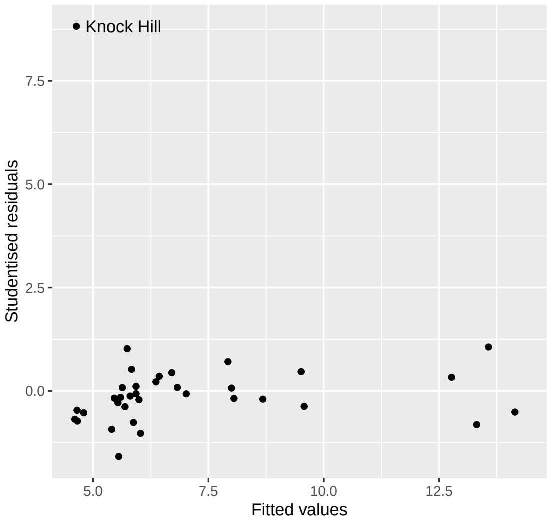
Section C

6. Atkinson (1990) reported on a data set giving *record times* (minutes) for 35 Scottish hill races, along with the *distance* run (miles) and the total *height* gained (feet). Let $y = \sqrt{record\ time}$, $x_1 = distance$, $x_2 = height/1000$. Regression of y on x_1, x_2 (with an intercept) produces the diagnostic plots A, B, C given on the next page and the four largest values of the diagonal entries of the hat matrix are given below.

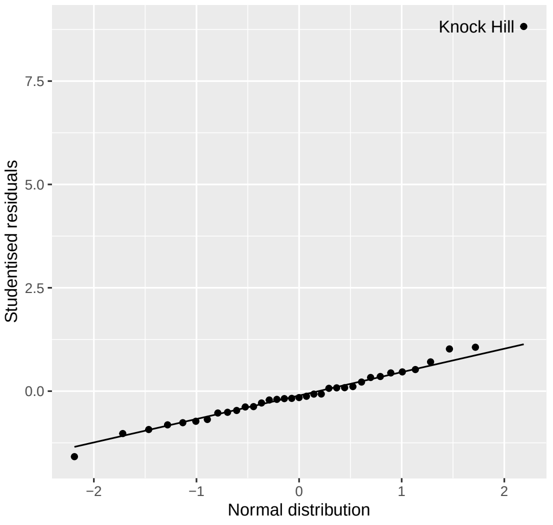
Location	h_{ii}
Lairig Ghru	0.690
Bens of Jura	0.420
Moffat Chase	0.191
Two Breweries	0.172

Comment on the diagnostics for this model, paying particular attention to issues of leverage, influence and suitability of the model itself. Comment briefly on diagnostic plots D, E, F, which were obtained by re-fitting the model with *Knock Hill* omitted.

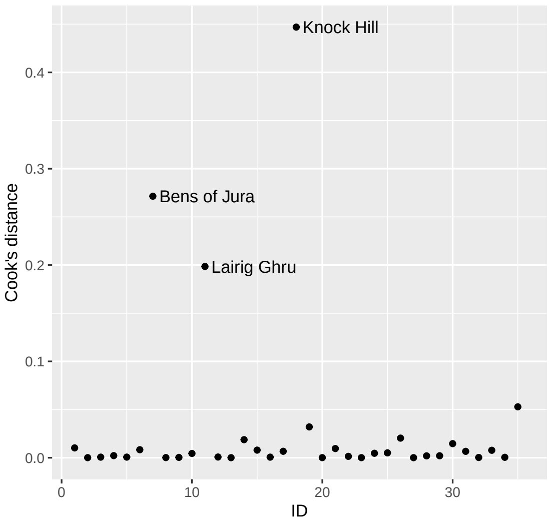
A: Studentised residuals against fitted values



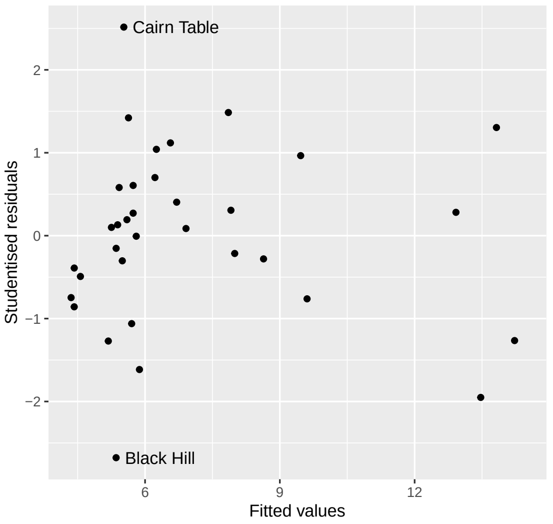
B: Normal Q−Q plot of Studentised residuals



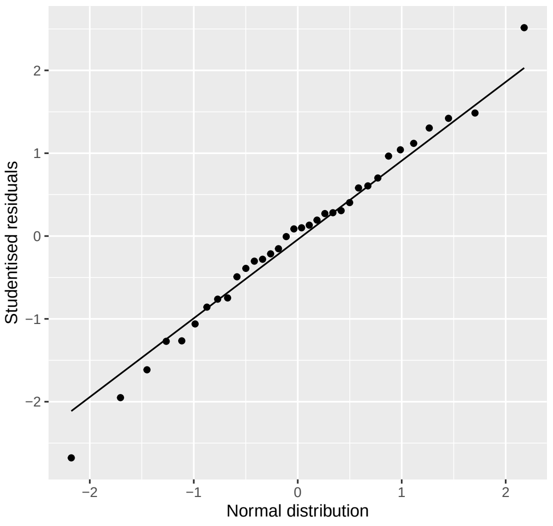
C: Cook's distance against ID



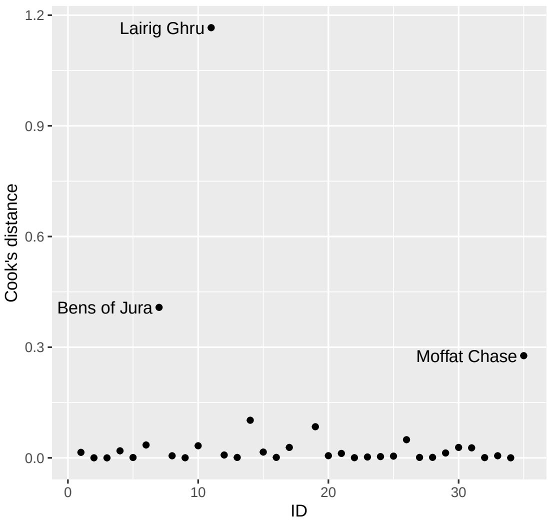
D: Studentised residuals against fitted values



E: Normal Q−Q plot of Studentised residuals



F: Cook's distance against ID


Lairig Ghru
Bens of Jura
Moffat Chase
Cook's distance
ID