MACQUARIE University
SYDNEY·AUSTRALIA

STAT2170 AND STAT6180
APPLIED STATISTICS

SGTA Week 9: Multiple Linear Regression 2023

Part 1

Question 1

Data for 100 elite female athletes at the Australian Institute of Sport are contained in the csv file **aisfemales.csv**. The variables are `Ht`=height in cm, `Wt`=weight in kg, `LBM`=lean body mass in kg, `RCC`=red blood cell count 1012/l, and `Hg`=haemoglobin concentration in g per decilitre. The aim is to construct a model relating lean body mass to the other variables. Load the data to the object names **aisfemales**

a. First, produce a scatterplot matrix including all variables, and comment on the relationships.

```
plot(aisfemales)
```

NB: Also can use `pairs(aisfemales, panel = panel.smooth)` to see matrix scatterplot with trend estimate in each pair.

b. Fit a regression using all predictors, including graphs for residuals and residuals vs predictors.

- **Hint**: You can conduct the analysis with `ais.reg = lm(LBM ~ Ht + Wt + RCC + Hg, data = aisfemales)`
- Normal QQ plot of the residuals: `qqnorm(ais.reg$residuals)`
- Residuals vs fitted: `plot(ais.reg$fitted, ais.reg$residuals))`. Are the assumptions satisfied? Explain.

c. RCC is not significant, so remove it from the model. Fit again, and check assumptions.

```
ais.reg <- lm(LBM ~ Ht + Wt + Hg, data = aisfemales)
```

Comment on any major change in the significance level of the coefficients, and why.

d. Using the sum of squares from the two models fitted construct an ANOVA table for RCC. You should be able to show this using the regression SS for each model. Conduct an F-test for the RCC regression term and find the P-Value.

e. Estimate the increase in LBM with 95% confidence for a 1kg increase in weight (other predictors held constant). If it was claimed that the increase in LBM was 0.5 kg for 1 kg increase in weight, would you accept or reject this at the 5% level of significance. Explain.

Question 2

The file **lifeexp.txt** contains data on life expectancy in years for a number of countries and data on the population per doctor and TV.

LifeExp	The life expectancy in years

People.per.TV	The average number of people per TV (this is a measure of affluence rather than directly affecting life expectancy)
People.per.Dr	The average number of people per physician

Read the data using a read.table command

```
lifeexp <- read.table("lifeexp.txt", header = TRUE)
```

a. Produce a plot of the relationships with a smoother and comment.
b. Log transform People per TV and Doctor and produce plot of the relationship using the log transformed variables. R code to achieve this below:
   - `lifeexp$logpptv = log(lifeexp$People.per.TV)`
   - `lifeexp$logppd = log(lifeexp$People.per.Dr)`
   - Check the log transformed data appears as two new columns by typing `head(lifeexp)`
   - Produce a new matrix scatterplot with the log data and comment again.
c. Fit a regression for life expectancy in terms of the log transformed variables with all graphs, and comment on the assumptions.
d. Interpret the effect a 10% increase in each predictor has on the life expectancy with a 95% confidence interval. Again, hold the other predictor constant in each case. (Hint: This will be an decrease in years per percentage increase in each variable, e.g. replace `People.per.TV` with $1.1 \times$ `People.per.TV` for a 10% increase and see the effect on life expectancy).

Part 2: Previous exam questions

Question 1

A study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, is conducted to examine the possible relationship between taste and the chemical composition of the cheese. Overall taste scores were obtained by combining the scores from several tasters. The following data were collected and analysed in `R` below.

`taste`	Aggregated score across the taste testers
`Acetic`	Natural log of concentration of acetic acid
`H2S`	Natural log of concentration of hydrogen sulphide
`Lactic`	Concentration of lactic acid

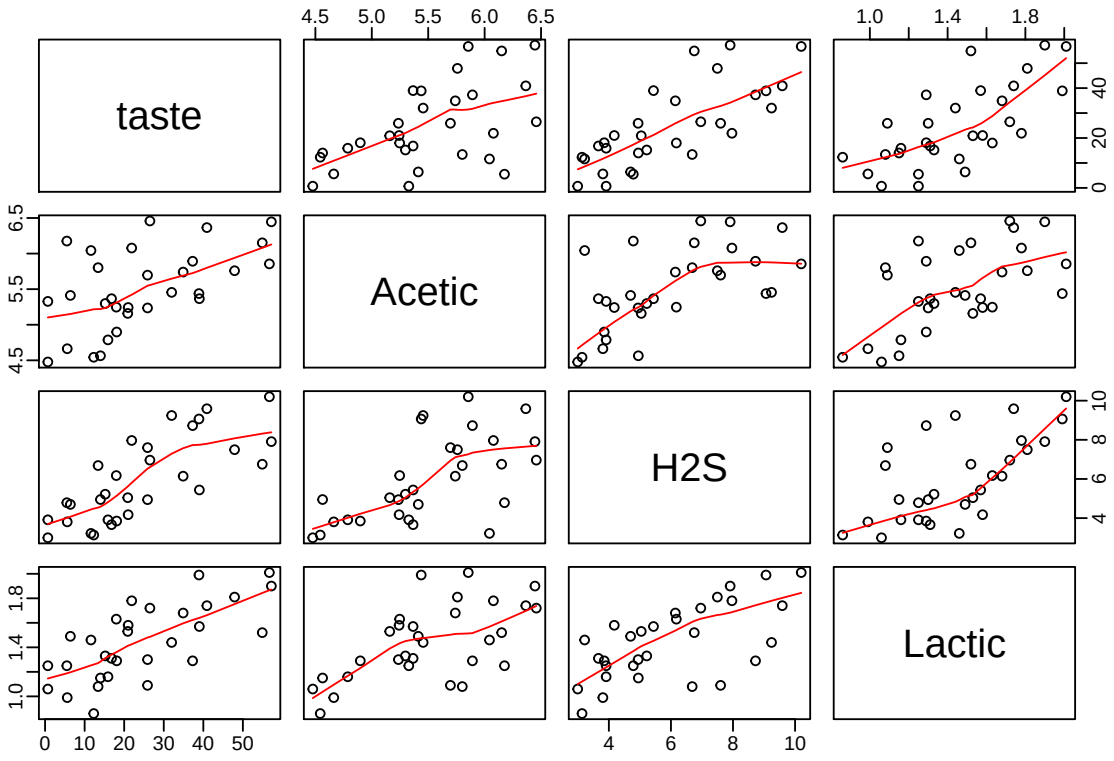

Figure 1: Matrix scatterplot of the LaTrobe Valley Cheese Data

```
n = nrow(cheese)
(n - 1) * cov(cheese)
```

```
#          taste   Acetic     H2S   Lactic
# taste  7662.89 147.8896 757.735 100.7530
# Acetic  147.89   9.4512  21.759   3.0337
# H2S     757.74  21.7591 131.185  12.0703
# Lactic  100.75   3.0337  12.070   2.6711
```

```
cor(cheese)
```

```
#         taste Acetic    H2S Lactic
# taste  1.0000 0.5495 0.7558 0.7042
# Acetic 0.5495 1.0000 0.6180 0.6038
# H2S    0.7558 0.6180 1.0000 0.6448
# Lactic 0.7042 0.6038 0.6448 1.0000
```

```
cheese.1 <- lm(taste ~ Acetic + H2S + Lactic, data = cheese)
summary(cheese.1)
```

```
#
# Call:
# lm(formula = taste ~ Acetic + H2S + Lactic, data = cheese)
#
# Residuals:
#    Min     1Q Median     3Q    Max
# -17.39  -6.61  -1.01   4.91  25.45
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -28.877     19.735   -1.46   0.1554
# Acetic         0.328      4.460    0.07   0.9420
# H2S            3.912      1.248    3.13   0.0042 **
# Lactic        19.671      8.629    2.28   0.0311 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 10.1 on 26 degrees of freedom
# Multiple R-squared:  0.652,   Adjusted R-squared:  0.612
# F-statistic: 16.2 on 3 and 26 DF,  p-value: 3.81e-06
```

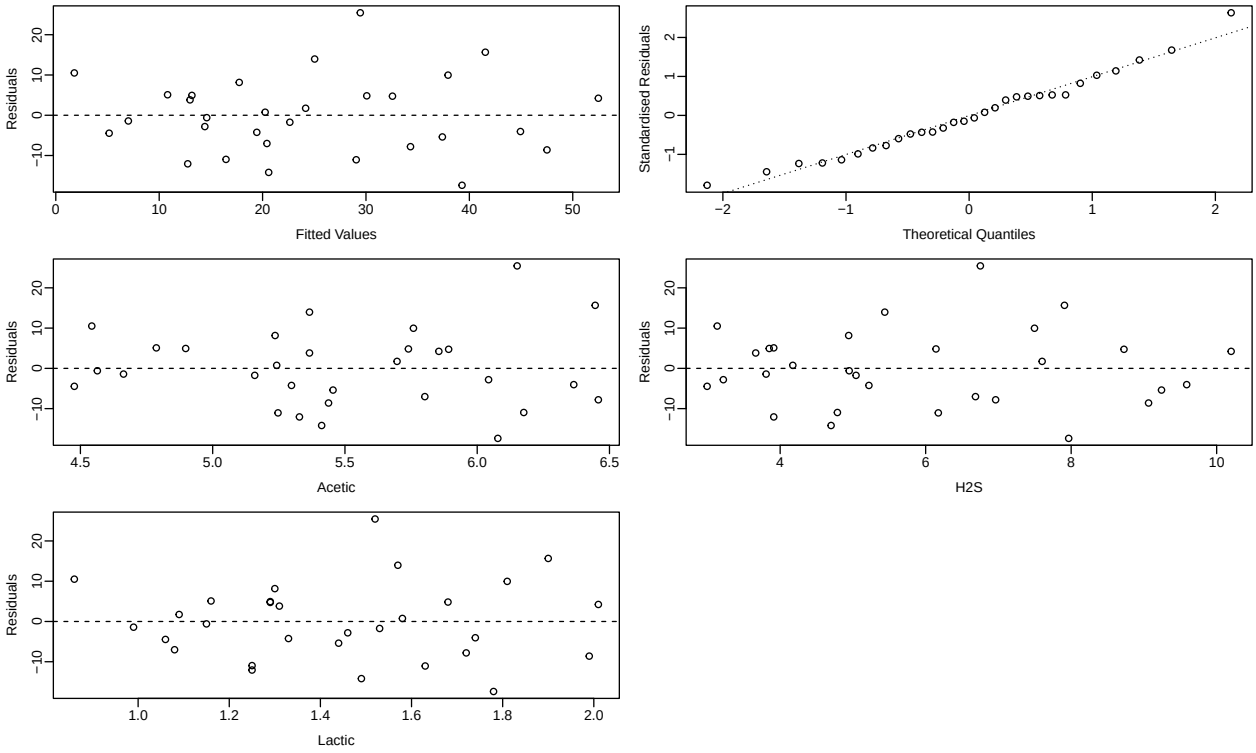

Figure 2: Diagnostic plots for the multiple regression model

a) Write down the linear model for the taste response, explained by the three chemical composition predictors. In your answer, carefully define all the necessary parameters.

b) Write down the fitted model for this dataset.

c) What are the assumptions required for a multiple regression analysis? If possible, validate those assumptions for this dataset with reference to the given output.

d) Determine R^2 and the adjusted R^2 from the output. Also, intuitively, what is R^2? What is the adjusted R^2? Why do we need this adjusted version of R^2?

e) Although the Acetic predictor in the scatter plot seems to have a linear relationship with the Taste, it is insignificant in the existing model. Give some explanation why it is insignificant; describe how you can improve the existing model.

Part 3

In `Microsoft Word`, one immediate see the what one wrote. On the other hand, in `RMarkdown`, one first includes the format information, such as header, list and font, in the `RMarkdown` file, and then see what it is like in the output file. Below we discuss more about the formatting for `RMarkdown`.

- Hint: When writing and formatting a `RMarkdown` file, one may consider knitting the file often so that you know which line(s) of code is(are) giving you the problem. This is not so dissimilar to when you work with the console that you only run a line at a time to identify the issue.

Header

In `RMarkdown`, we could include headers with `#`. For example, to produce a level-1 header, one could write in the `RMarkdown` file:

```
# Header 1
```

To generate a level-2 header, one could write in the `RMarkdown` file

```
## Header 1
```

Always put a space between the number signs (`#`) and the heading name.

List

To generate an un-ordered list, we could make use of `-`. For example, the codes below will lead to a list with two levels:

```
- Level 1
    - Level 2
```

and the output will be

- Level 1
    - Level 2

Notice that the indent was created by tabbing twice (or 4 whitespaces). Tab and whitespace have a special meaning in Markdown so if your formatting code wasn't interpreted correctly it may caused by having too many or not enough whitespaces in your code.

To generate an ordered list, one could try the codes below

```
1. Level 1
    a) Level 2
```

and the output will be

1. Level 1
    a) Level 2

Font

One could write texts in **bold** with **, for example:

```
**Text**
```

One could write texts in *Italian* with *, for example:

```
*Text*
```

To learn more about how to format your content in RMarkdown, here are some resources to get you started.

- Markdown tutorial - 10 minutes tutorial: https://www.markdowntutorial.com.

Question 1

Now try to create a pdf file with the output below using `RMarkdown`. Notice that both **Unordered list** and **Ordered list** below are Level 4 headers.

Unordered list

- **Bold**
- *Italic*
  - *Italic 1*
  - *Italic 2*
    - *Italic 2-1*
    - *Italic 2-2*

Ordered list

1. Paragraph 1
2. Paragraph 2
   a) Paragraph 2-1
   b) Paragraph 2-2