MACQUARIE University
SYDNEY·AUSTRALIA

STAT2170 AND STAT6180
APPLIED STATISTICS

SGTA Week 8 : Simple Linear Regression 2023

Part 1

Question 1

Two methods were used to determine the level of E. Coli. The aim was to evaluate the agreement between HEC (in $\log_{10}$ CFU/ml) and a more complex test HGMF (in $\log_{10}$ CFU/ml). Portions of beef were contaminated with E. Coli and then each piece was tested by both methods. Data is contained in the file `ecoli.txt`

HEC	HGMF
0.50	0.42
0.06	0.20
0.20	0.42
0.61	0.33
0.20	0.42
0.56	0.64
-0.82	-0.82
0.67	1.06
1.02	1.21
1.20	1.25
0.93	0.83
2.27	2.37
2.02	2.21
2.32	2.44
2.14	2.28
2.09	2.69
2.30	2.43
-0.10	1.07

a) Inspect the scatterplot of the data.
b) Fit a model with HEC as the response
c) Check if the model is appropriate and comment on the required assumptions for the data.
d) Identify the outlier, which was found to be as a result of incorrect procedures.
e) Make the necessary change to the data (remove the outlier) and refit the model. Then record your conclusions and comment on assumptions. Also obtain confidence and prediction intervals for a HGMF value of 1.

Hint: After loading the data (e.g. `ecoli = read.table("ecoli.txt", header = TRUE)`), the model can be fit with the code `mylm = lm(HEC ~ HGMF, data = ecoli)`.

```
# Warning: Be careful when copying and pasting code between PDF and the R window.
# The tilde character on PDF font is different to the ~ font character in R
```

Question 2

a) Construct 95% confidence interval for the slope.
b) Using this confidence interval test the hypothesis that the slope is 1.

Question 3 (to be done by hand)

Suppose `cleanEcoli` is the R object that contains the data without the outlier.

```
x = cleanEcoli$HGMF
xbar = mean(x)
n = length(x)
Sxx = (n - 1) * var(x)
c(n, xbar, Sxx)
```

```
# [1] 17.000000  1.198824 16.961176
```

Using the above summary statistics and previous regression output, answer the questions below:

a) Calculate the confidence and prediction intervals for a HGMF value of 1. Compare these to the values obtained from R.
b) Explain why the prediction interval is wider.
c) If we had a single measurement of HGMF which is the appropriate interval for the variability in an estimate of HEC.

Question 4

Compute the Analysis of Variance for your clean regression object using the `anova` command. Show that the relationship between the t statistic from the table of coefficients and the F-test is as expected.

Hint: Suppose you typed `mycleanlm = lm(HEC ~ HGMF, data = cleanEcoli)` and `cleanEcoli` is the ecoli data with the outlier removed. Your clean linear regression analysis was named `mycleanlm`, to compute the simple linear regression then you can compute the ANOVA by typing `anova(mycleanlm)` to see the ANOVA table for that regression analysis.

Question 5:

Why is it not appropriate to rearrange the regression equation to give HGMF as a response? The model definition may be helpful.

Part 2

As you may have already known, `RMarkdown` is a file format that allows for the usage of both the `R` language, a language for statistics programming, and the `Tex` language, a language for mathematical typesetting.

Question 1

Complete the installations if you haven't done it before. Specifically, to install the `RMarkdown` components, you could type in the `R` console

```
install.packages("rmarkdown")
```

- For Mac users, you may be asked to install `Xcode` (a rather big installation). We only need a small piece of it called the `command-line tools`. Run the following line: `xcode-select --install` in the `Terminal` to continue. You should be able to find the `Termainl` tab next to the `Console` tab in `RStudio`.

Further, to incorporate the `Tex` language, you could type in the `R` console

```
install.packages("tinytex")
tinytex::install_tinytex()
```

Question 2

`RMarkdown` can turn a combination of math formula, `R` output, and text into a `HTML` webpage. To have a try,

a) Create a new sample `RMarkdown` file using `File > New File > RMarkdown` on the `RStudio` menu; for the `Default Output Format` option, choose `HTML`.

b) Click on the new tab that just popped up, which by default has a name of `Untitled1`. This tab include a `RMarkdown` file. Knit this `RMarkdown` file by one of the ways below. After knitting, in the pop-up box, name the file `Test1` and put it to your working directory.

- clicking on the `Knit` button right below the tabs
- following the path of `File > Knit Document` on the `RStudio` menu
- use a keyboard shortcut: `Ctrl+Shift+K` for Window & Linux and `Shift+Command+K` for Mac.

(When knitting, `RStudio` will predominantly take the current `RMarkdown` file as input and execute the code in a **new, separate R session** and therefore largely unaffected from whatever have been previously run in the `R` console, such as reading in a dataset or setting up a working directory.)

c) Now there should be two files with the name of `Test1` in your working directory. The extension of one file should be `html`. What's the extension of the second file? Try to open this second file with `RStudio`.

Question 3

The `tinytex` package we have installed allows `RMarkdown` to export a collection of math formula, `R` output, and text into a `PDF` file. To see how,

a) Create a new sample `RMarkdown` file using `File > New File > RMarkdown` on the `RStudio` menu; for the `Default Output Format` option, choose `PDF`.

b) Click on the new tab that just popped up. What is there after the `output:` on the third line of this `RMarkdown` file ? Is this the same as in the third line of the `RMarkdown` file in the previous question?

c) Knit this `RMarkdown` file as in the previous question. After knitting, in the pop-up box, name the file `Test2` and put it to your working directory. Then try to find the file `Test2.pdf` in your working directory.