LIGHT CSW 2023 Quantitative Analysis Workshop

Exercises

Alex Richards, Marc Henrion

1 March 2023

Exercise 1

Go to the course website on GitHub:

https://github.com/mlw-stats/LIGHT_CSW2

From here, download the following files:

`btTBreg.csv`
`btTBregHospitals.csv`

1. Load the `btTBreg.csv` data table into R.
2. The variables `cd41`, `cd42` and `cd41.sk`, `cd42.sk` measure the same variables (`cd4` and `cd4.sk` respectively) in the same individuals at two different time point. This means the data are in wide format. Reformat to long format.
3. Save the reformatted data into a file called `btTBregLong.tab` in such a way that
   i. Columns are tab-separated.
   ii. Column names are saved.
   iii. No row number is saved in the resulting file.
4. Load the `btTBregHospitals.csv` data table. Join the data frames storing `btTBreg.csv` and `btTBregHospitals.csv`.
5. Compute the average patient age and the proportion of male patients for each hospital.
6. Write an R function that computes the following summary statistics, then, using your custom function, compute these for the `bmi`, `cd41`, `cd42` columns:
   i. mean
   ii. median
   iii. inter quartile range
   iv. minimum
   v. maximum
   vi. number of missing values
7. Do the same now, but only for female patients. Repeat for only male patients.

Exercise 2

Using the `iris` dataset (type `?iris` to get more information about this dataset) that comes pre-loaded with R, produce the following figures:

- Produce histograms for each of `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`.

- Produce a bar plot for `Species`.
- Produce box and whisker plots for each of the 4 continuous variables. Put them all on a single, multi-panel figure.
- Repeat for just `Sepal.Length` using a violin plot, stratifying by `Species`.
- Produce a single graph (not multi-panel) that has histograms for `Sepal.Length` for each of the 3 flower species.
- There are 4 continuous variables. This means there are 6 possible pairs of these. For each such pair, produce a scatter plot of one variable against the other and highlight the different flower species by using a different colour for each species.
- For one of these 6 scatter plots: estimate the bivariate probability density and add density contour lines to the figure.

Exercise 3

Install the package `nycflights13`, then load it. This has data on flights that took off in the US during 2013. There are 5 data tables: + `airlines`, data on airlines + `airports`, data on airports + `planes`, data on planes + `weather`, hourly weather data at NYC airports for 2013 + `flights`, data on flights leaving NYC airports during 2013

- Compute the average delay by destination, then join the airports data frame to get the longitude and latitude of delays. Plot this (if you are using ggplot2, then the functions `borders()` and `coord_quickmap()` can be useful for a nicer figure).
- Construct data frames giving average delay per wind speed / temperature / precipitation / visibility. Produce scatter plots of each of these against delay and add an average trend line.

Exercise 4

Decide what design could be used to answer the following research questions:

1. What is the prevalence of HIV in urban Blantyre in 2018?
2. Do men experience higher mortality compared to women once they start ART?
3. Does smoking increase the chance of having lung cancer?
4. What is the effect of providing oral HIV self-test kits on the uptake of HIV testing?
5. What interventions may improve linkage to ART following community based HIV testing?

Exercise 5

Take the iris dataset and explain how you would, in a formal statistical way, compare the following:

1. `Petal.Width` between the flower species `virginica` and `setosa`.
2. `Sepal.Length` between all 3 flower species.

For each comparison, state which test you will use (there may be more than one valid option!), state the null and alternative hypotheses, do the test and interpret the results.

Exercise 6

In a drug trial, researchers are assessing overall in-hospital mortality as the primary outcome. The new drug is compared against the standard-of-care treatment (SOC). Patients are randomised 1:1 to the new drug and SOC. At trial conclusion, the researchers observe that out of 250 SOC patients, 61 have died and out of 250 patients on the new drug arm, 48 have died.

Perform a statistical test to conclude whether or not there is a difference between the new drug and the SOC. State the test you use, the null and alternative hypotheses, perform the test and interpret the results.

Exercise 7

Test whether the 2 variables from Table 1 below are independent or not. State the test you use, the null and alternative hypotheses, do the test and interpret the results.

Table 1: Summary of patient outcomes for different health centers.

	alive	dead
Hospital1	92	29
Hospital2	54	15
Hospital3	31	3

What when you repeat your analysis for Table 2 below?

Table 2: Summary of patient outcomes for different health centers.

	alive	dead
Hospital1	920	290
Hospital2	540	150
Hospital3	310	30

Comment on the results from your analyses for both tables.

Exercise 8

Using the `adolescent_small.csv` data, fit a linear model regressing weight (variable `a104wt`) on age (variable `a12age`).

Test if the regression coefficient of age $\beta_{age} = 0$.

Note:

- deviance = sum of squares
- residual = error

Exercise 9

Using the `adolescent_small.csv` data, fit the following GLM model:

Weight `a104wt` as a function of

- age `a12age`
- height `a103ht`
- hiv `hiv`
- sex `a13sex`

Produce:

- a residuals vs. fitted values graph
- histogram of the residuals
- a QQ plot.

Exercise 10

You are given the following data:

$$\mathbf{x} = (-6, -6, -4, -1, 0.5, 2, 8, 8, 11, 11.5)^T$$

$$\mathbf{y} = (-3.7, -4.3, -3.9, -4.6, 0.5, -6.9, 10.2, 16.1, 6, 19.5)^T$$

a. Fit a linear regression model to these data and show the model output.

b. Describe the resulting regression line:

- What is the relationship between variables X and Y?
- How much (on average) does Y change when X changes by 1?
- What value does Y take (on average) when $X = 0$?

c. Compute the coefficient of determination R^2, the adjusted R^2, the likelihood and the AIC. Which of these tell you how good your model fits the data?

d. Compute the residuals $r_i = y_i - \hat{y}_i$ and do a normal distribution QQ plot.

e. What other diagnostic check(s) could you do? Do this and explain whether you think this is a good model.

f. Re-fit the model, but now including a term for X^2: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$. Check and discuss the resulting model and compare it to the previous one. Which model would you recommend for this dataset?