MATH38172 Generalised Linear Models

Computer Lab 1

We will use the following software:

- R – a fully-featured statistical programming language (https://www.r-project.org)
- RStudio – a development environment for R, which organizes R nicely and makes it more efficient to work with (https://rstudio.com/products/rstudio/)

Both programs are available on the workstations in the Alan Turing Cluster and can also be downloaded free of charge on your own computer.

Instructions: work through the explanatory content below and attempt the excercises when ready. The explanatory material gives many R commands. Make sure you type these in to the console manually to check you can get them to work. It is a good idea to save your work in an R script file.

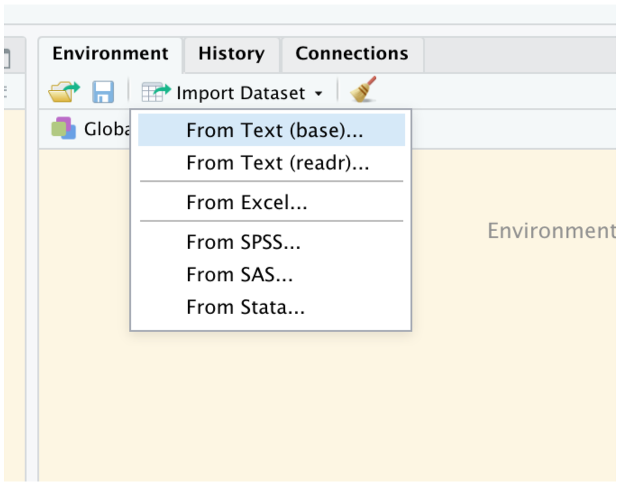
Importing and exploring data

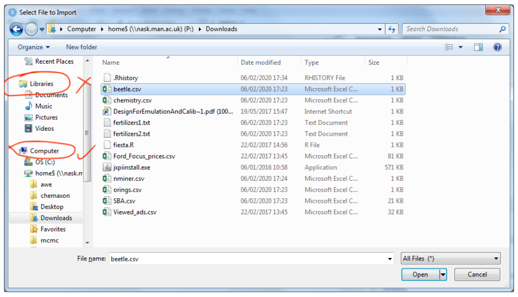
To download the data file `beetle.csv` from Blackboard:

- navigate to this week's folder and find the item `Datasets for Lab`
- right click on `beetle.csv`, and select 'Download linked file as…' / 'Save linked file as…' / 'Save target as…' (depending on your browser)
- choose an appropriate location to save the file.

Importing data

To import the data in RStudio, use the menu option below (see left figure below):





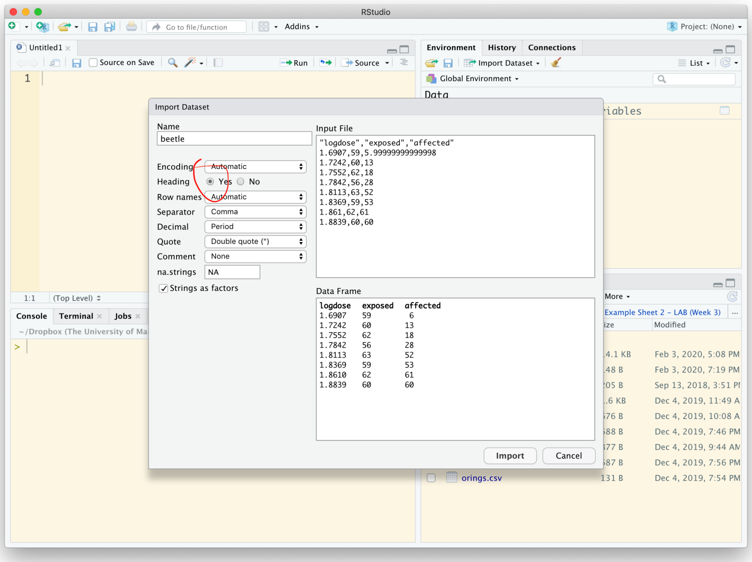
Name	Date modified	Type	Size
.Rhistory	06/02/2020 17:34	RHISTORY File	1 KB
beetle.csv	06/02/2020 17:23	Microsoft Excel C...	1 KB
chemistry.csv	06/02/2020 17:23	Microsoft Excel C...	1 KB
DesignForEmulationAndCalib~1.pdf (100...	19/05/2017 15:47	Internet Shortcut	1 KB
fertilizers1.txt	06/02/2020 17:23	Text Document	1 KB
fertilizers2.txt	06/02/2020 17:23	Text Document	1 KB
fiesta.R	22/02/2017 14:56	R File	1 KB
Ford_Focus_prices.csv	22/02/2017 13:45	Microsoft Excel C...	81 KB
jxpiinstall.exe	06/01/2016 10:58	Application	571 KB
nminer.csv	06/02/2020 17:24	Microsoft Excel C...	1 KB
orings.csv	06/02/2020 17:23	Microsoft Excel C...	1 KB
SBA.csv	06/02/2020 17:23	Microsoft Excel C...	21 KB
Viewed_ads.csv	22/02/2017 13:45	Microsoft Excel C...	32 KB

File name: beetle.csv

All Files (*)

This allows us to browse to the appropriate data file on the hard drive, or use a web address. If using the computers in the Alan Turing cluster, make sure to access files through 'Computer' rather than 'Libraries' (see right figure above), as the latter will not work.

Once we have chosen a file, a menu appears which allows us to select the appropriate file format. E.g. below we select Heading = ‘yes’ so that R reads the variable names from the data.



Import Dataset

Name: beetle

Encoding: Automatic
Heading: [✓ Yes] No
Row names: Automatic
Separator: Comma
Decimal: Period
Quote: Double quote (")
Comment: None
na.strings: NA
✓ Strings as factors

Input File

```
"logdose","exposed","affected"
1.6907,59,5.99999999999998
1.7242,60,13
1.7552,62,18
1.7842,56,28
1.8113,63,52
1.8369,59,53
1.861,62,61
1.8839,60,60
```

Data Frame

logdose	exposed	affected
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Clicking ‘Import’ puts the data in a data frame called `beetle`. We can inspect the structure of the data frame using

```r
str(beetle)
```

```
## 'data.frame':    8 obs. of  3 variables:
##  $ logdose : num  1.69 1.72 1.76 1.78 1.81 ...
##  $ exposed : int  59 60 62 56 63 59 62 60
##  $ affected: int  6 13 18 28 52 53 61 60
```

This shows the names of the different variables and their types, e.g. `logdose` is numeric, while `exposed` and `affected` are integers.

Regression for Binomial data

We will use R to fit a simple logistic regression model to the beetle data, with $\log_{10} x_i$ as the explanatory variable i.e.

$$m_i Y_i \sim \text{Bi}(m_i, \mu_i)$$
$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \beta_0 + \beta_1 \log_{10} x_i$$

where

- x_i denotes the ith dose
- m_i denotes the number of beetles exposed to the ith dose
- Y_i denotes the *proportion* of beetles that are affected at dose x_i
- m_iY_i is the total number of beetles affected at dose x_i

Computing the response

First note that we can compute the response Y as follows

```r
y <- beetle$affected / beetle$exposed
y
```

```
## [1] 0.1016949 0.2166667 0.2903226 0.5000000 0.8253968 0.8983051 0.9838710
## [8] 1.0000000
```

Note that for vectors `a`, `b`, `a/b` performs elementwise division, i.e. the ith element of `a/b` is |`a[i]`|/|`b[i]`|.

Fitting the model

Now to fit the model use the `glm` function as follows:

```r
fit1 <- glm(y~logdose, family=binomial, weights=exposed, data=beetle)
```

- `fit1 <-` tells R to store the model fit in an object called `fit1` (you can use any name, so long as it is not already taken)
- the formula `y~logdose` tells R to fit a model with response `y` (the proportion of affected beetles) using `logdose` as an explanatory variable.

  Note that you do not need to explicitly specify the intercept or the parameters in the formula: R will include them automatically.
- `family=binomial` tells R to use a binomial response distribution – R will then automatically fit a logistic regression model.
- `weights=exposed` tells R that the numbers of binomial trials for each observation are given in the vector `exposed`.
- `data=beetle` tells R to look for the vector `logdose` in the data frame `beetle`.

Multiple explanatory variables

If we had multiple explanatory variables, say `x1` and `x2` we could include these in the model by replacing the formula `y~logdose` with `y~x1+x2`.

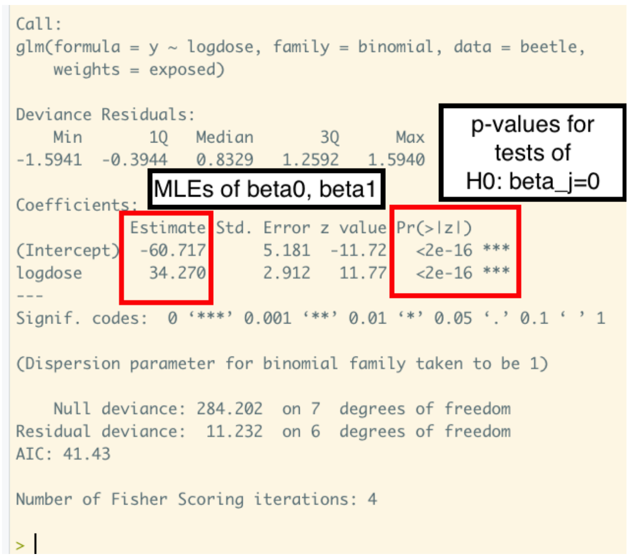
Inspecting the results

The coefficients can be extracted as follows:

```r
coef(fit1)
```

```
## (Intercept)     logdose
##   -60.71745    34.27033
```

More detailed results can be inspected using the command `summary(fit1)`. This gives the output shown overleaf.



```
Call:
glm(formula = y ~ logdose, family = binomial, data = beetle,
    weights = exposed)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
logdose       34.270      2.912   11.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4

> |
```

Alternative response specification

Another way of specifying the response is as a matrix where the first column gives the number of binomial successes and the second column gives the number of binomial failures, e.g.

```
fit2 <- glm(cbind(affected, exposed-affected)~logdose, family=binomial, data=beetle)
```

In this case the argument `weights` is not needed.

Note that the `cbind` command used above joins two vectors together as columns to create a matrix, e.g.

```
a <- c(1,2,3,4,5,6)
a
```

```
## [1] 1 2 3 4 5 6
```

```
b <- rep(7,6)
b
```

```
## [1] 7 7 7 7 7 7
```

```
cbind(a,b)
```

```
##      a b
## [1,] 1 7
## [2,] 2 7
## [3,] 3 7
## [4,] 4 7
## [5,] 5 7
## [6,] 6 7
```

An alternative way of specifying the vector `a` is `a <- 1:6`.

Exercises part 1

1 For the beetle data, check that the two different ways of specifying the response variable give the same results.

2 The file `orings.csv` contains the Challenger data.

a) Import the data into R and inspect them.

b) Fit a simple logistic regression model for the probability of O-ring failure using temperature as the explanatory variable. Try both ways of specifying the response.

c) Check that the results from both ways agree with each other and with what was given in lectures.

3 (Credit risk example). The file `SBA.csv` contains data about loans made to 2102 small businesses in the USA. The file contains four variables:

- `Default` – whether the business defaulted on the loan (1 – default; 0 – did not default)
- `Portion` – the proportion of the loan gross amount that is guaranteed to the bank by the US Small Business Administration (SBA) (a real number in (0,1))
- `Recession` – whether the loan was active during the Great Recession (December 2007 to June 2009) (1– active; 0 – not active)
- `RealEstate` – whether the loan is backed by real estate (1 – backed by real estate; 0 – not backed by real estate)

a) Fit a logistic regression model to predict the probability of loan default using `Portion`, `Recession`, and `RealEstate` as explanatory variables.

b) Interpret the parameters of this model.

c) A small business applies to the bank for a loan with no SBA guarantee, and no real estate backing. What is the estimated probability that they default on the loan?

Poisson regression

Recall that the simple Poisson regression model (with a log link) is

$$\begin{aligned} Y_i &\sim \text{Po}(\mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

This model can be fitted using `glm(y~x,family=poisson,data=dataset)`. Note that the argument `weights` is not needed here.

Exercises part 2

4 The data `nminer.csv` contains the Noisy miner data.

i) Import the data into R and inspect them.
ii) Fit a Poisson regression model and verify that your results agree with what was given in lectures.

Fitting linear models

To fit a linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

use the R command

```
model1 <- lm(y~x1+x2)
```

(assuming the data are stored in R vectors `y`, `x1`, `x2` – they may have different names in R). Similar to the `glm` command, to inspect the fitted model use `summary(model1)`.

Qualitative explanatory variables

It is straightforward to include a qualitative/categorical variable as an explanatory variable in all of the above models. If the variable is stored as a character type – e.g. letters or text strings for the different levels – then R will automatically create appropriate dummy variables.

If the levels of the categorical variable are coded numerically, then it should first be converted to a `factor` type (e.g `lm(y~factor(x))`). E.g. if we have a numerical vector

```
a<- c(1,1,2,2)
a
```

```
## [1] 1 1 2 2
```

this can be converted to a categorical variable with two levels:

```
factor(a)
```

```
## [1] 1 1 2 2
## Levels: 1 2
```

Exercises part 3

5 The file `chemistry.csv` contains data on reagent concentration (x) and product yield (Y) for 15 different runs of an experiment. We wish to fit the quadratic model

$$Y = \alpha + \beta x + \gamma x^2 + \epsilon\,.$$

i) Import the data, create a new variable `z=chemistry$x^2` and fit the above model using `lm()`.
ii) What model is fitted if you use the command `lm(y~x+x^2,data=chemistry)`? What model is fitted if you use the command `lm(y~x+I(x^2),data=chemistry)`?

6 The file `fertilizers1.txt` contains experimental data on plant heights for ten different types of fertilizer, with the fertilizers coded as A, B, ..., J. The file `fertilizers2.txt` contain the same data but with the data coded as 1, 2, ..., 10.

i) Import both datasets and inspect their structure.
ii) For each of the two datasets, fit a linear model to predict the plant height from the fertilizer. Write down the fitted models and interpret the parameters.
iii) Why aren't the results the same for the two datasets? How can we make them the same?