RMarkdown demo

Rudolf Schlechter

2024-05-16

## First things first

### Why do we care about R Markdown?

Scripts can be messy and difficult to follow when not enough information is provided. However, *too much* text in a script can be distracting!

R Markdown is a simple formatting syntax. The main purpose is to connect data analysis with documentation. For that, the most simple way is to combine **markdown** (md) with **R** syntax.

.Rmd file -> dynamic document -> reproducible research

## Markdown

# Header 1

## Header 2

### Header 3

#### Header 4

##### Header …

Plain text.

You can use *italics* and **bold**.

superscripts and subscripts.

~~strikethrough~~

* Unordered list
* Itemised
  + Item 1
  + Item 2

1. Ordered list

verbatim code

block quote

equation:

Table

[link](https://rmarkdown.rstudio.com)

| Component | Amount | Unit |
| --- | --- | --- |
| Buffer | 10 | L |
| Primer Fwd | 1.5 | L |
| Primer Fwd | 1.5 | L |
| Water | 6 | L |
| DNA | 1.0 | L |

## Workflow

1. Open new .Rmd file
2. Embed code
3. Write text
4. Set output format (header)
5. Save and render
6. Share!

## Let’s work with a dataset

To insert a code chunk, try alt + ctrl + I or option + cmd + I.

message <- c(print('Hello world!'))

## [1] "Hello world!"

I want to write Hello world!.

Let’s install the [penguins data set](https://allisonhorst.github.io/palmerpenguins/)

install.packages("palmerpenguins")

## The following package(s) will be installed:  
## - palmerpenguins [0.1.1]  
## These packages will be installed into "~/repo/rmarkdown\_demo/renv/library/R-4.3/aarch64-apple-darwin20".  
##   
## # Installing packages --------------------------------------------------------  
## - Installing palmerpenguins ... OK [linked from cache]  
## Successfully installed 1 package in 29 milliseconds.

Now we load the package and dataset

library(palmerpenguins)  
data(package = 'palmerpenguins')

### Data summary

str(penguins)

## tibble [344 × 8] (S3: tbl\_df/tbl/data.frame)  
## $ species : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ island : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ bill\_length\_mm : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...  
## $ bill\_depth\_mm : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...  
## $ flipper\_length\_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...  
## $ body\_mass\_g : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...  
## $ year : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...

Looking at the structure of the data, we have 3 species of penguins: Adelie, Chinstrap, Gentoo, found in the islands of Biscoe, Dream, Torgersen.

head(penguins)

## # A tibble: 6 × 8  
## species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## <fct> <fct> <dbl> <dbl> <int> <int>  
## 1 Adelie Torgersen 39.1 18.7 181 3750  
## 2 Adelie Torgersen 39.5 17.4 186 3800  
## 3 Adelie Torgersen 40.3 18 195 3250  
## 4 Adelie Torgersen NA NA NA NA  
## 5 Adelie Torgersen 36.7 19.3 193 3450  
## 6 Adelie Torgersen 39.3 20.6 190 3650  
## # ℹ 2 more variables: sex <fct>, year <int>

The dataset contains data of 344 penguins, and each entry has phenotypic data such as bill length (mm), bill depth (mm), flipper length (mm), body mass (g), sex (male/female), and year of entry.

# Determine the mean value of every parameter from each penguins, grouped by species, island and sex  
penguins %>%   
 na.exclude %>%   
 group\_by(species, island, sex) %>%   
 summarise(bill\_length = mean(bill\_length\_mm),  
 bill\_depth = mean(bill\_depth\_mm),  
 flipper\_length = mean(flipper\_length\_mm),  
 body\_mass = mean(body\_mass\_g))

## # A tibble: 10 × 7  
## # Groups: species, island [5]  
## species island sex bill\_length bill\_depth flipper\_length body\_mass  
## <fct> <fct> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 Adelie Biscoe female 37.4 17.7 187. 3369.  
## 2 Adelie Biscoe male 40.6 19.0 190. 4050   
## 3 Adelie Dream female 36.9 17.6 188. 3344.  
## 4 Adelie Dream male 40.1 18.8 192. 4046.  
## 5 Adelie Torgersen female 37.6 17.6 188. 3396.  
## 6 Adelie Torgersen male 40.6 19.4 195. 4035.  
## 7 Chinstrap Dream female 46.6 17.6 192. 3527.  
## 8 Chinstrap Dream male 51.1 19.3 200. 3939.  
## 9 Gentoo Biscoe female 45.6 14.2 213. 4680.  
## 10 Gentoo Biscoe male 49.5 15.7 222. 5485.

We can make a nicer table using kable from the knitr package. Also we can get rid of the message with local options (set message = FALSE) or move it to the global options.

penguins %>%   
 na.exclude %>%   
 group\_by(species, island, sex) %>%   
 summarise(bill\_length = mean(bill\_length\_mm),  
 bill\_depth = mean(bill\_depth\_mm),  
 flipper\_length = mean(flipper\_length\_mm),  
 body\_mass = mean(body\_mass\_g)) %>%   
 knitr::kable()

| species | island | sex | bill\_length | bill\_depth | flipper\_length | body\_mass |
| --- | --- | --- | --- | --- | --- | --- |
| Adelie | Biscoe | female | 37 | 18 | 187 | 3369 |
| Adelie | Biscoe | male | 41 | 19 | 190 | 4050 |
| Adelie | Dream | female | 37 | 18 | 188 | 3344 |
| Adelie | Dream | male | 40 | 19 | 192 | 4046 |
| Adelie | Torgersen | female | 38 | 18 | 188 | 3396 |
| Adelie | Torgersen | male | 41 | 19 | 195 | 4035 |
| Chinstrap | Dream | female | 47 | 18 | 192 | 3527 |
| Chinstrap | Dream | male | 51 | 19 | 200 | 3939 |
| Gentoo | Biscoe | female | 46 | 14 | 213 | 4680 |
| Gentoo | Biscoe | male | 49 | 16 | 222 | 5485 |

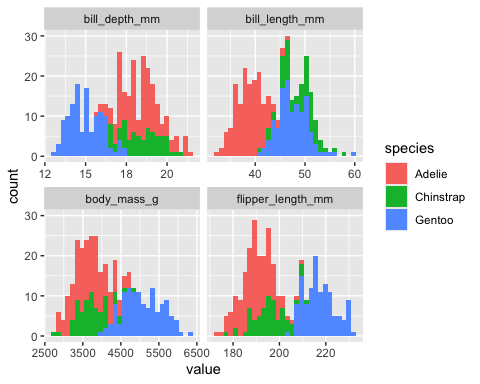
## Exploratory analysis

R Markdown is useful to add those extra steps that might not end up in your final manuscript, such as your exploratory analysis!

Let’s evaluate if the data is normally distributed by looking at histograms

# Histogram  
penguins %>%   
 pivot\_longer(cols = c(bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm, body\_mass\_g)) %>%   
 ggplot(aes(x = value, fill = species)) +  
 facet\_wrap(~name, scales = "free\_x")+  
 geom\_histogram()

## Warning: Removed 8 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



We can analyse a bit further the body mass of the different species of penguins.

## Linear model

A linear model is a regression analysis in which a response variable can be described as a function of one or more predictors. In R, we express it as y ~ x.

In our case, we will use body mass as our response variable and we want to see how the different species contribute to the variation in weight among penguins.

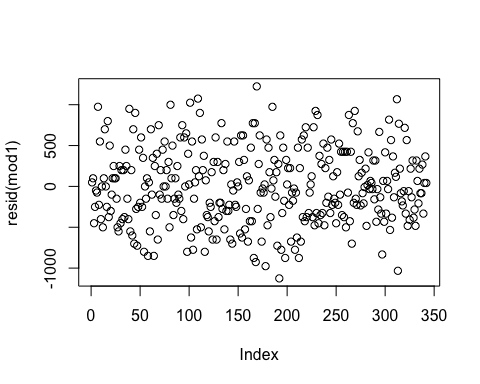
# Model definition  
mod0 = lm(body\_mass\_g ~ 1, data = penguins) # Null model  
mod1 = lm(body\_mass\_g ~ species, data = penguins)  
  
# Is our model better than the null model?  
anova(mod0, mod1)

## Analysis of Variance Table  
##   
## Model 1: body\_mass\_g ~ 1  
## Model 2: body\_mass\_g ~ species  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 341 219307697   
## 2 339 72443483 2 146864214 344 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Model Summary  
summary(mod1)

##   
## Call:  
## lm(formula = body\_mass\_g ~ species, data = penguins)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1126.0 -333.1 -33.1 316.9 1224.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3700.7 37.6 98.37 <2e-16 \*\*\*  
## speciesChinstrap 32.4 67.5 0.48 0.63   
## speciesGentoo 1375.4 56.1 24.50 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 462 on 339 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.67, Adjusted R-squared: 0.668   
## F-statistic: 344 on 2 and 339 DF, p-value: <2e-16

# Residuals  
plot(resid(mod1))

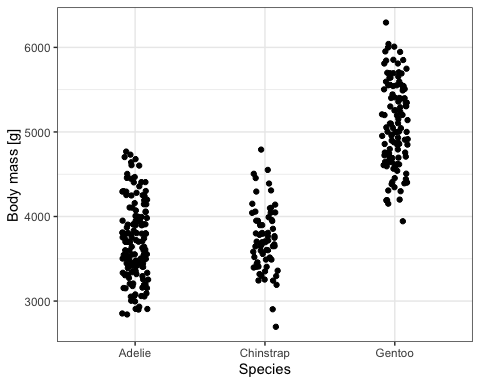


# Extract some the data from the model  
df.mod1 <- tidy(mod1) %>%   
 mutate(species = levels(penguins$species),  
 mean\_body\_mass = ifelse(term == "(Intercept)",   
 estimate,   
 estimate[species == "Adelie"] + estimate))  
  
species.adelie.model <- df.mod1$species[1]

We can extract information from our model and use inline code. For example, we have now learned that the weight of species Adelie is lower (mean = 3700.66 g) than that of species Gentoo (mean = 5076.02 g), with a *p*-value = 5.42^{-77} and an adjusted *R*2 = 66.77%.

Let’s plot our results:

penguins %>%   
 na.exclude %>%   
 ggplot(aes(x = species, y = body\_mass\_g))+  
 geom\_jitter(width = 0.1)+  
 theme\_bw()+  
 labs(x = "Species", y = "Body mass [g]")



We can include the means extracted from the model:

penguins %>%   
 na.exclude %>%   
 ggplot(aes(x = species, y = body\_mass\_g))+  
 geom\_jitter(width = 0.1)+  
 # Add a red dot representing the mean of each group  
 geom\_point(data = df.mod1, aes(x = species, y = mean\_body\_mass),  
 color = "red",  
 size = 3)+  
 theme\_bw()+  
 labs(x = "Species", y = "Body mass [g]")

