Week 7: Generalised Linear Models part 2

Required R packages

Before we proceed, load all the packages needed for this week:

```
library(tidyr)
library(ggplot2)
library(moderndive)
library(sjPlot)
library(tidymodels)
library(broom)
library(performance)
library(faraway)
```

Introduction

Last week we introduced **Generalised Linear Models** (GLMs). Particularly, we looked at **logistic regression** to model outcomes of interest that take one of two categorical values (e.g. yes/no, success/failure, alive/dead). This week we will continue reviewing logistic regression to model grouped binary outcomes (e.g. number of success out of a fixed number of trials) and then we we will generalise this to situations where the response variable is categorical with more than two categories. First lets look at the framework for modelling categorical data with only two categories, i.e.

- **binary**, taking the value 1 (say success, with probability p) or 0 (failure, with probability $1-p$) or
- **binomial**, where y_i is the number of events (successes) in a given number of trials n_i, with the probability of success being p_i and the probability of failure being $1-p_i$.

In both cases the distribution of y_i is assumed to be binomial, but in the first case it is $\text{Bin}(1, p_i)$ and in the second case it is $\text{Bin}(n_i, p_i)$. The first case was covered last week, so now lets focus on the second case.

Logistic regression with grouped binary data

Suppose that our binary outcome y_i is grouped across n_i number of trials, e.g. number of times a head landed when a coin was tossed on multiple occasions or the proportion of beetles that were killed after being exposed to an insecticide.

In such cases $y_i \sim \text{Bin}(n_i, p_i)$, often referred to as proportional data, since our dependent variables are expressed as percentages or fractions of a whole. Lets look at a an example.

It is known that the incubation temperature can affect the sex determination of turtles. Data from an experiment where turtle eggs were incubated at various temperatures and the proportion of male hatchlings was recorded are used to examine the link between incubation temperature and the chance of hatchlings being a male.

The `turtle` data set within the `faraway` library contains the number of hatched male and female turtles across different temperatures, with 3 independent replicates for each temperature.

```
turtles = faraway::turtle
turtles%>% glimpse()
```

```
Rows: 15
Columns: 3
```

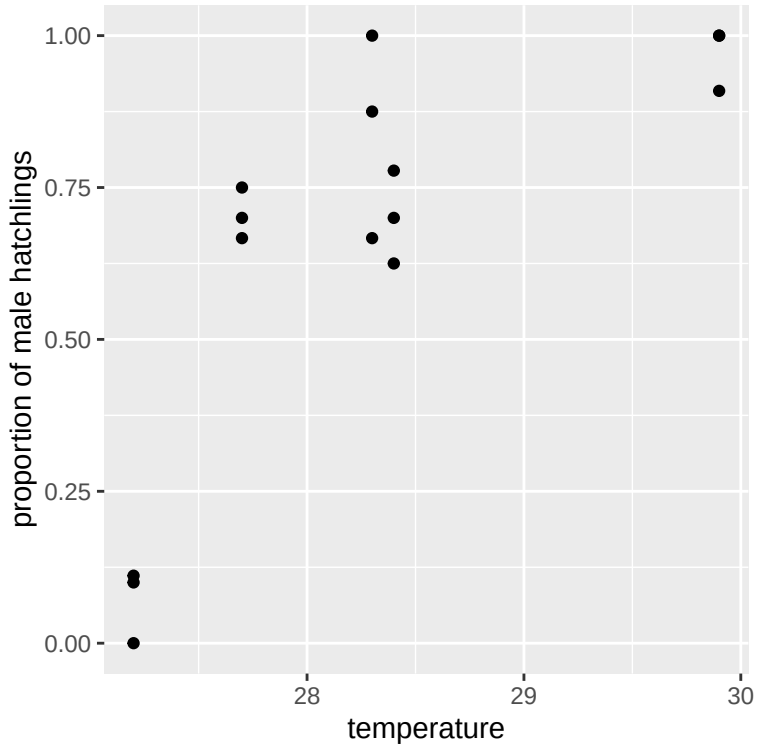
```
$ temp   <dbl> 27.2, 27.2, 27.2, 27.7, 27.7, 27.7, 28.3, 28.3, 28.3, 28.4, 28.~
$ male   <int> 1, 0, 1, 7, 4, 6, 13, 6, 7, 7, 5, 7, 10, 8, 9
$ female <int> 9, 8, 8, 3, 2, 2, 0, 3, 1, 3, 3, 2, 1, 0, 0
```

Lets investigate whether the probability of a male hatching increases or decrease with the temperature. First, we need to compute the proportion of males that hatched on each replicate per temperature. To do this, we obtain the ratio between the total number of male hatchlings and total number hatchlings (males+females):

```
turtles = turtles %>%
  mutate(totals = male+female,
         male_props = male/totals)
```

We can see on the next plot, that the proportion of males hatchling seems to increase as the temperature as the incubuation period rises.

```
ggplot(turtles,aes(y= male_props,x=temp))+geom_point()+ labs(y="proportion of male hatchlings
```



To corroborate this result, we can fit a GLM to the data.

$$y_i \sim \text{Binomial}(n_i, p_i) \tag{1}$$
$$\text{logit}(p_i) = \beta_0 + \beta_1 \times \text{temperature}. \tag{2}$$

Here, y_i denotes the number of hatched males on the ith replicate, n_i is the fixed total number of hatched eggs per replicate ($i = 1, \ldots, N$), p_i is the probability of a male turtle being hatched and β_0 and β_1 are our unknown parameters to be estimated.

Proportions can be modelled by providing an $N \times 2$ matrix of the number of success (num. of males hatchlings) and the number of failures (number of female hatchlings):

```
model_turtles <- glm(cbind(male,female) ~ temp,
                     data = turtles,
                     family = binomial)
```

or by providing the proportion of males hatchlings and `weights` totals in the function (i.e the number of trials):

```
model_turtles <- glm(male_props ~ temp,
                     data = turtles,
                     weights =  totals,
                     family = binomial)
```

These two formulation are valid and will yield to the same following result:

```
model_turtles %>% tidy(conf.int = T)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-61.32	12.02	-5.10	0	-86.83	-39.73
temp	2.21	0.43	5.13	0	1.44	3.13

The interpretation goes as follows:

- For every unit increase (celsius degrees presumably) in *Temperature,* the log-odds of a male being hatched increase by 2.21 i.e. the chances of a hatching a male increases as the temperature during the incubation increases.
- Given $p_{val} < 0.05$, we can **reject the null hypothesis** $\beta_1 = 0$ that one unit increase in temperature does not affect chances of a male being hatched.

- For every unit increase in *Temperature*, the **odds** of hatching a male are 9.13 times the odds of those with one *temperature* unit less.

Question

If an egg is incubated at a temperature of 27.5 degrees, what at the chances (odds) of a **female** being hatched.
I need a hint
Recall that $log\ Odds\ (male) = \log\left(\frac{P(male)}{1-P(male)}\right) = \hat{\beta}_0 + \hat{\beta}_1 \times temperature$, but we are interested in $\frac{P(female)}{1-P(female)}$, thus $Odds\ (female) = \exp\left[-\log\left(\frac{P(male)}{1-P(male)}\right)\right] = \exp\left(-\left[\hat{\beta}_0 + \hat{\beta}_1 \times 27.5\right]\right)$

- (A) The chances of an male being hatched are 45% greater than a female hatchling if the egg was incubated at a temperature of 27.5 degrees
- (B) The chances of an female being hatched are 45% greater than a male hatchling if the egg was incubated at a temperature of 27.5 degrees
- (C) The chances of an female being hatched are 67% greater than a male hatchling if the egg was incubated at a temperature of 27.5 degrees
- (D) The chances of an male being hatched are 67 greater than a female hatchling if the egg was incubated at a temperature of 27.5 degrees

Question

What is the probability of a turtle egg that is incubated in a temperature of 28.5 degrees to become a male?
I need a hint
Recall that $P\ (male) = \frac{Odds(male)}{1+Odds(male)} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \times temperature)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 \times temperature)} = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 \times 28.5)$

- (A) 0.45
- (B) 0.18
- (C) 0.85
- (D) 0.15

Besides our usual model checks and model evaluation metrics, when dealing with proportional data sometimes we find that the observed variability in the data is greater than the one expected by the model, i.e. $Var(Y) = n\ p\ (1 - p)$.

This excess of variance is called **overdispersion** and its an indicator that our model is missing some important variability in the data (e.g. unaccounted factors affecting the probability of an event, non-independent trials, clustering within the data, among others).

To check for overdispersion we can use the built-in `check_overdispersion()` function from the performance package (to learn more about overdispersion see [@gelman2006]):

```
check_overdispersion(model_turtles)
```

```
# Overdispersion test

 dispersion ratio = 1.250
          p-value = 0.176

No overdispersion detected.
```

In this example its seems we don't have to worry about it. But what about the binary case? well overdispersion is usually not a concern here because the variance cannot exceed the range for a binary response where each observation represents a single outcome (0 or 1) and the variance of the model is constrained since $Var(Y) = p(1 - p)$.