Week 8: Generalised Linear Models part 2

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 successes 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.

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

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

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. An experiment was conducted where turtle eggs were incubated at various temperatures and the number of male and female hatchlings was recorded. The goal of the experiment was 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
```

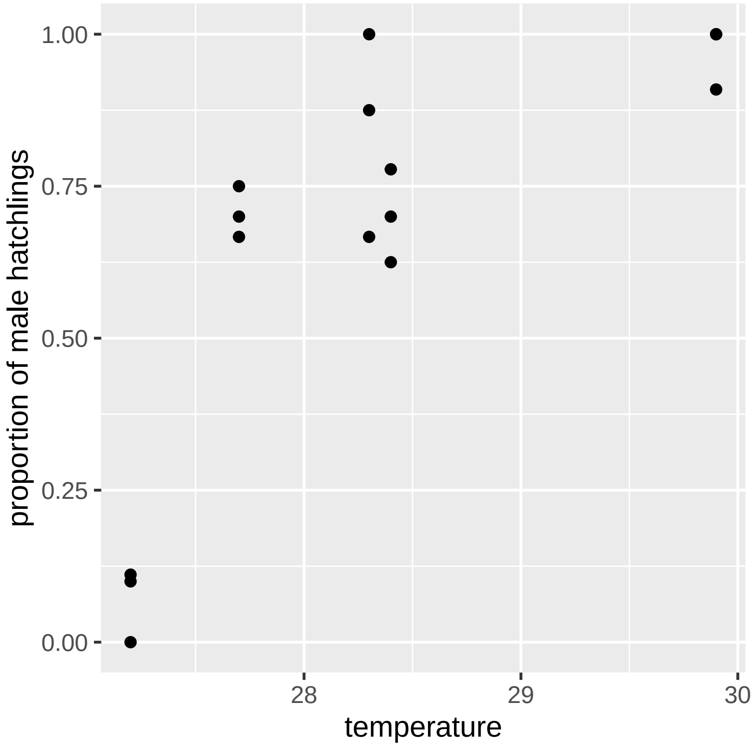
```
$ temp   <dbl> 27.2, 27.2, 27.2, 27.7, 27.7, 27.7, 28.3, 28.3, 28.3, 28.4, 28.~
$ male   <int> 1, 0, 1, 7, 4, 6, 13, 6, 7, 7, 5, 7, 10, 8, 9
$ female <int> 9, 8, 8, 3, 2, 2, 0, 3, 1, 3, 3, 2, 1, 0, 0
```

Lets investigate whether the probability of a male being hatched 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 hatchlings seems to increase as the incubation temperature rises.

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



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

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n_i, p_i) \qquad (1) \\
\text{logit}(p_i) &= \beta_0 + \beta_1 \times \text{temperature}. \qquad (2)
\end{aligned}
$$

Here, y_i denotes the number of hatched males on the ith experiment replicate, n_i is the fixed total number of hatched eggs per replicate, 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 positive events (num. of males hatchlings) and the number of negative events (number of female hatchlings):

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

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

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

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

```r
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 hatching a male increases as the incubation temperature 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 $\exp(\beta_1) = 9.13$ 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

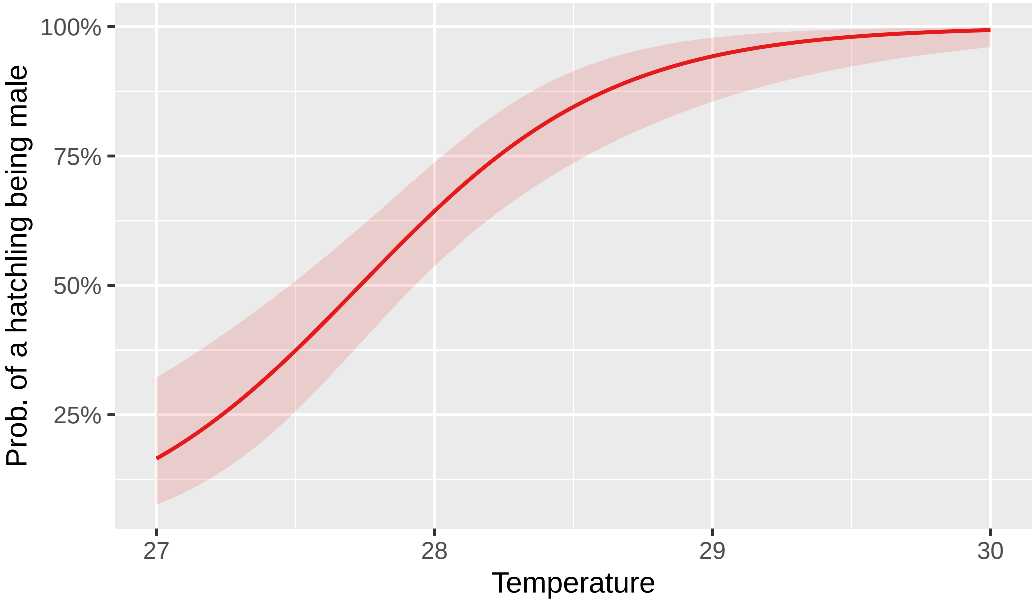
We can now plot the predicted probabilities of a hatchling being male. However, notice that the number of unique temperature values in the `turtles` data set is not very large:

```r
turtles %>% select(temp) %>% unique()
```

```
   temp
1  27.2
4  27.7
7  28.3
10 28.4
13 29.9
```

Thus, we can create a coarser grid of temperature values to make our predictions and then use the `plot_model()` function as follows:

```r
temp_pred = seq(27,30,by=0.01)
plot_model(model_turtles,
           type = "eff",
           title = "",
           terms="temp[temp_pred]",
           axis.title = c("Temperature", "Prob. of a hatchling being male"))
```



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 $P\ (male) = \dfrac{Odds(male)}{1+Odds(male)} = \dfrac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \times temperature)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 \times temperature)}$. In R, the inverse logit transformation can be achieved using the `plogis()` function.

- (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 Gelman and Hill (2006)):

```r
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 (i.e. ungrouped data)? 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 $Var(Y) = p(1-p)$.

Modelling grouped binary data and a categorical covariate

In the last section we reviewed the case when the explanatory variable was continuous., lets look now at the case when the explanatory variable is categorical.

To illustrate how the previous model works with categorical predictors we can discretized the temperature values into arbitrary categories as follows:

$$\text{temperature category} = \begin{cases} \text{temperature} > 29^{\circ}C & \text{high} \\ \text{temperature} > 28^{\circ}C & \text{medium} \\ \text{else} & \text{low} \end{cases}$$

In R we can use the `case_when()` function to accomplish this:

```r
turtles = turtles  %>% mutate(
  temp_fct = case_when(
    temp > 29 ~ "high",
    temp > 28 ~ "medium",
    .default = "low"
  ) %>% as.factor()
)
```

Now, recall that as usual, R will set the baseline category for our explanatory variable in alphabetical order, i.e. the `high` temperature level will be treated as reference for the dicretized variable. However, we already know that the chances of a male being hatched increases with higher incubation temperatures.

Thus, it makes sense to assess how the chances of a male being hatched are affected by comparing higher temperature categories against lower ones. This implies that we will set `low` to be our reference category. Luckily, we have seen in previous tasks how to do this using the `relevel()` function:

```r
turtles = turtles %>%
  mutate(temp_fct = relevel(temp_fct,ref = "low"))
```

We can now fit a logistic regression using the `low` temperature level as the reference category for our dicretized temperature covariate. The model is then given by:

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n_i p_i) && (3)\\
\text{logit}(p_i) &= \alpha + \beta_1 \times \mathbb{I}_{\text{temperature}}(\text{high}) + \beta_2 \times \mathbb{I}_{\text{temperature}}(\text{medium}). && (4)
\end{aligned}
$$

- α represent the **log-odds** of a male turtle being hatched in `low` incubation temperature.
- β_1 are the change int the **log-odds** of a male turtle being hatched given it was incubated in a `high` temperature condition compared to a `low` one.
- $\mathbb{I}_{\text{temperature}}(\text{high})$ is an indicator variable that takes the value of 1 if the ith experiment replicate was conducted on a `high` temperature.
- β_2 are the change int the **log-odds** of a male turtle being hatched given it was incubated in a `medium` condition compared to a `low` one.
- $\mathbb{I}_{\text{temperature}}(\text{medium})$ is an indicator variable that takes the value of 1 if the ith experiment replicate was conducted on a `medium` temperature.

In R, the model can be fitted as follows:

```
model_turtles_2 <- glm(cbind(male,female) ~ temp_fct,
                     data = turtles,
                     family = binomial)
```

Lets print the model estimates odds scale and 95% confidence intervals (remember we can achieve this by setting `conf.int=TRUE` and `exponentiate=TRUE` within the `tidy` function):

```
model_turtles_2 %>% broom::tidy(conf.int = T,exponentiate = T)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.59	0.29	-1.80	0.07	0.33	1.04
temp_fcthigh	45.47	1.06	3.61	0.00	8.59	843.97
temp_fctmedium	6.32	0.44	4.23	0.00	2.76	15.31

We can see that the odds of a male being hacthed if it was incubated on a `low` temperature condition are 0.59 the odds of a female being hatched if it was incubated on the same condition.

Alternatively, we could interpret this as the odds of female being hatched in a `low` temperature incubation settings being $\exp(\alpha)^{-1} = 1.68$ higher than the odds of a male being hatched under the same setting. However, there is not enough evidence to support that the change in the odds is statistically significant since the confidence interval (0.33 , 1.04) contains 1 (remember we are in the odds scale).

On the other hand, the odds of a male being hatched are 45.47! significantly higher in a `high` temperature setting compared to a `low` temperature. Likewise, the odds of a male being hatched are 6.32 higher in a `medium` temperature contidion compared to a `low` one.

What if we want to compare the odds of a male being hatched if the egg was incubate on a `high` temperture conditon against a `medium` one?

In that case, we we will be looking at the following odds ratio:

$$
\begin{aligned}
\frac{\text{Odds}(\text{male} = 1|\text{temperature} = \textit{high})}{\text{Odds}(\text{male} = 1|\text{temperature} = \textit{medium})} &= \frac{\frac{p_{\text{temperature}=high}}{1-p_{\text{temperature}=high}}}{\frac{p_{\text{temperature}=medium}}{1-p_{\text{temperature}=medium}}} && (5)\\
&= \frac{\exp(\alpha+\beta_1)}{\exp(\alpha+\beta_2)} && (6)\\
&= \exp(\alpha+\beta_1-\alpha-\beta_2) && (7)\\
&= \exp(\beta_1-\beta_2) = \frac{\exp(\beta_1)}{\exp(\beta_2)} && (8)
\end{aligned}
$$

Where β_1 and β_2 are the coefficients in the **log-odd** scale. However since we already have $\exp(\beta_1) = 45.47$ and $\exp(\beta_2) = 6.32$, then the odds of male being hatched from an egg that was incubated on a `high` temperature condition are $\frac{45.47}{6.32} = 7.2$ greater than the one that was incubate on a `medium` temperature condition.

Finally, we can calculate the probabilities of a male being hatched in each temperature condition as follows:

- $P(\text{male} = 1|\text{temperature} = low) = \frac{\exp( )}{1 + \exp(\alpha)}$ In R this is:

  ```
  plogis(coef(model_turtles_2)[1])
  ```

  ```
  (Intercept) 
     0.372549 
  ```

- $P(\text{male} = 1|\text{temperature} = medium) = \frac{\exp(\; + \; {}_1)}{1 + \exp(\alpha + \beta_1)}$. In R this is equivalent to:

  ```
  plogis(coef(model_turtles_2)[1] + coef(model_turtles_2)[3])
  ```

  ```
  (Intercept) 
    0.7894737 
  ```

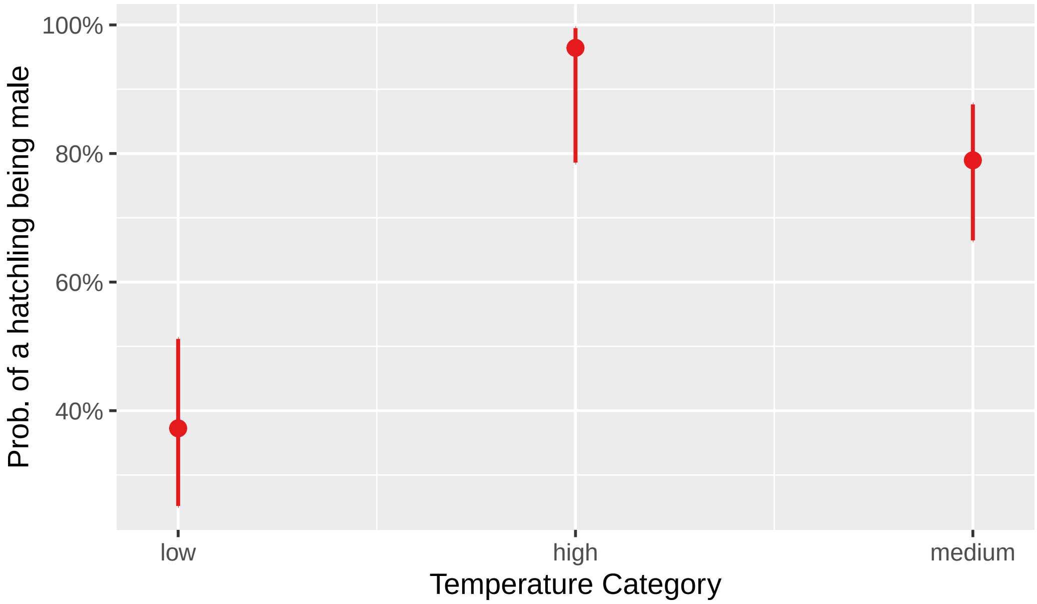
- $P(\text{male} = 1|\text{temperature} = high) = \frac{\exp(\; + \; {}_2)}{1 + \exp(\alpha + \beta_2)}$. In R this is computed as:

  ```
  plogis(coef(model_turtles_2)[1] + coef(model_turtles_2)[2])
  ```

  ```
  (Intercept) 
    0.9642857 
  ```

We can visualize this probabilities using the `plot_model()` function as follows:

```
plot_model(type = "pred",
           model_turtles_2,
           terms = "temp_fct",
           axis.title = c("Temperature Category",
                          "Prob. of a hatchling being male"),
           title = " ")
```



Gelman, Andrew, and Jennifer Hill. 2006. "Data Analysis Using Regression and Multilevel/Hierarchical Models," December. https://doi.org/10.1017/cbo9780511790942.