Tutorial 6 - Generalised Linear Models (Solutions).

Christopher A Oldnall

Welcome to the sixth tutorial of the Regression and Simulation methods module. This is the next script in developing your skills in R, whilst learning about how to implement GLMs. Throughout this notebook we will start to consider how to pick a model for the situation you are in.

Exercise 0

Throughout remember we will need tidyverse, go ahead and do this as your first task.

Your Answer:

```
library(tidyverse)
```

Exercise 1

Before we explore some exercises this week, we are going to load in the data. This week to GLMs in context we will be working on a data set, that is derived from Public Health Scotland's monitoring of cancer waiting times within Scotland. More information about the cancer waiting times data collected by Public Health Scotland can be found here (with an example report from PHS): https://publichealthscotland.scot/publications/cancer-waiting-times/cancer-waiting-times-1-april-to-30-june-2023/

For our purposes we will be using a 'cleaned' set of this data. It is called, '31days.csv' Go ahead and load this in below.

```
ThirtyOneDays <- read_csv("/Users/chrisoldnall/Library/Mobile Documents/com~apple~CloudDocs/Teaching/SMS
```

```
## New names:
## Rows: 4999 Columns: 12
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (3): Quarter, HB, CancerType dbl (9): ...1,
## NumberOfEligibleReferrals31DayStandard, NumberOfEligibleRefer...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

Exercise 2

Explore the following code. The comments have been delibretly taken out. Add in accurate comments where possible, to make the data handling most efficient and clean.

```
HealthBoards <- c('S08000020','S08000022', 'S08000025', 'S08000026', 'S08000030', 'S08000028', 'S080000
CancerAreas <- c('NCA', 'NCA', 'NCA', 'NCA', 'NCA', 'NCA', 'SCAN', 'SCAN', 'SCAN', 'SCAN', 'WOSCAN', 'W
```

```
df_mapping <- data.frame(HealthBoards = HealthBoards, CancerAreas = CancerAreas)

ThirtyOneDays <- ThirtyOneDays %>%
  left_join(df_mapping, by=c("HB" = "HealthBoards"))

ThirtyOneDays_Aggregated <- aggregate(cbind(NumberOfEligibleReferrals31DayStandard=ThirtyOneDays$Number
```

Exercise 3

It is always important when you receive data to explore it and ask questions! Write a small commentary on what is in 'ThirtyOneDays_Aggregated'. Once you have done that ask Clara any questions you have following this exploration. Update your notes in a new paragraph.

Your Answer: Within the data there are 8 variables, and 39 observations. There is a quarter character representing the time period the data comes from. Following this the Date period follows this along ranging from 0 to 38. Three categories Pandemic, JustPandemic and Pandemic_Cat which all relate to indicating the periods of the pandemic. We then have the number of referals and then the number of referals who started treatment within 31 days. Finally the population size is included.

Exercise 4

We want to build a GLM for the purposes of predicting future number of eligible referrals for the 31 day standard. Consider what type family of models (normal, possion, binomial etc.) may be used for this. Write an explanation as to why you think this and further more set the parameter family1 to this family.

```
family1 = "poisson"
```

Exercise 5

Now we want to create the GLM. Consider from the data which factors you might include. Use the below framework to create a model, replacing _____ with factors.

```
model_1 <- glm(NumberOfEligibleReferrals31DayStandard ~ Date + Pandemic, data = ThirtyOneDays_Aggregate
summary(model_1)
```

```
##
## Call:
## glm(formula = NumberOfEligibleReferrals31DayStandard ~ Date +
##     Pandemic, family = family1, data = ThirtyOneDays_Aggregated)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  9.3037970  0.0029998 3101.44   <2e-16 ***
## Date         0.0037539  0.0001407   26.68   <2e-16 ***
## Pandemic    -0.0670814  0.0028602  -23.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.27  on 38  degrees of freedom
## Residual deviance:  867.54  on 36  degrees of freedom
## AIC: 1310.4
##
## Number of Fisher Scoring iterations: 4
```

Exercise 6

Following creating a GLM we want to use it to predict. Using the predict function and your model with the type 'response' to get a model prediction.

```
model_1_predictions <- predict(model_1, type="response")
summary(model_1_predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10239   11293   11659   11683   12083   12663
```

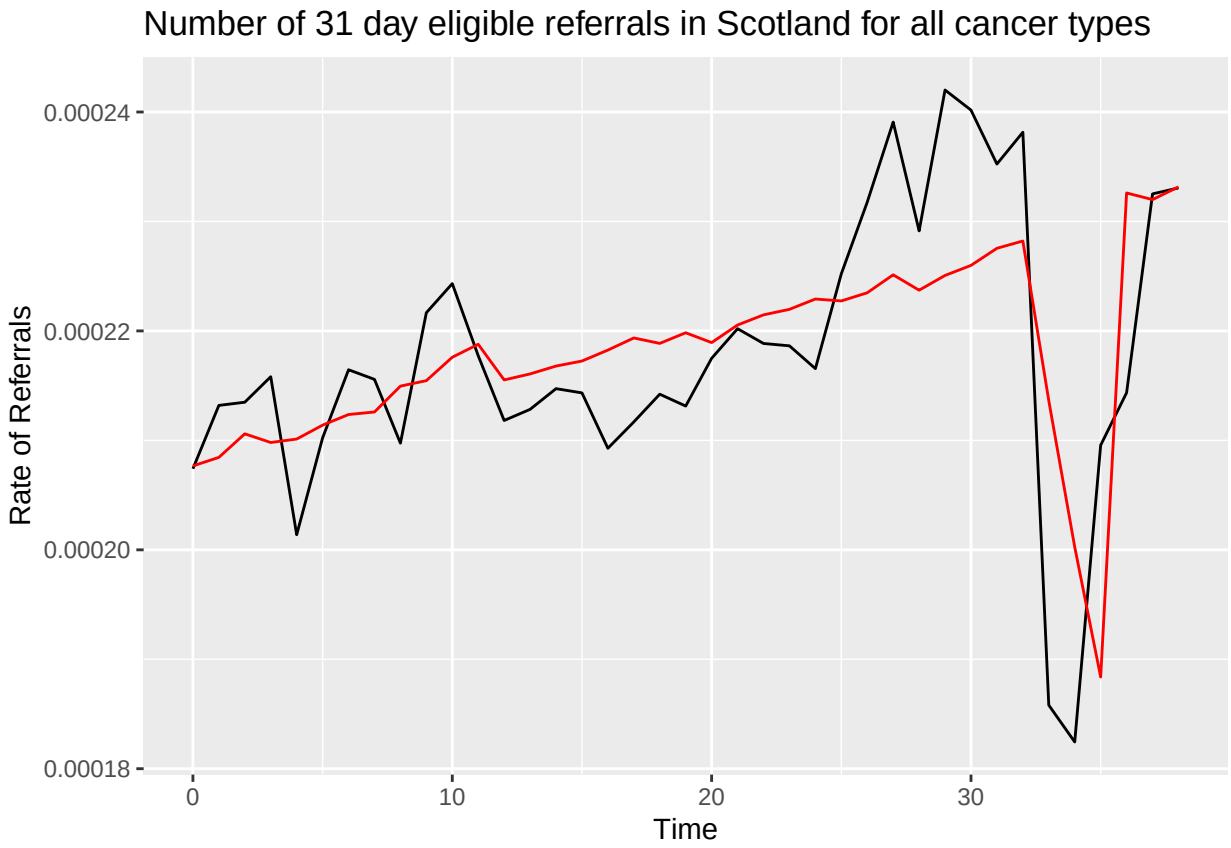
Exercise 7

Now we look to plot our GLM and see visually whether it fits the data well. Use the code below to do this. You should spend some time commenting the code, in particular comment on the division by population size throughout.

```
ThirtyOneDays_Aggregated <- ThirtyOneDays_Aggregated %>%
  mutate('fitted1' = model_1_predictions/PopSize, 'rates' = NumberOfEligibleReferrals31DayStandard/PopS

ThirtyOneDays_Plot <- ThirtyOneDays_Aggregated %>%
  ggplot(aes(x=Date)) +
  labs(x = 'Time', y = 'Rate of Referrals', title = 'Number of 31 day eligible referrals in Scotland fo
  geom_line(aes(y = rates), color = "black") +
  geom_line(aes(y = fitted1), color = "red")

ThirtyOneDays_Plot
```


Number of 31 day eligible referrals in Scotland for all cancer types
Rate of Referrals
0.00024
0.00022
0.00020
0.00018
0
10
20
30
Time