STA 214 Homework 7

Due: Friday, March 24, 12:00pm (noon) on Canvas.

Instructions: In this assignment, you will explore Poisson and quasi-Poisson regression with data on air quality in Chicago.

Getting started: Begin by downloading the HW7 template from the course website:

`https://sta214-s23.github.io/homework/hw_07_template.Rmd`

Save this template file to your computer, then open it in RStudio. As you complete the assignment, you will write down your answers to all questions in the R Markdown file, and include all R code in code chunks. *If a question requires code, you will not receive credit if no code is provided.* Refer to the R Markdown instructions on the course website (`https://sta214-s23.github.io/resources/rmarkdown_instructions/`) if you have issues getting started.

Submission: When you have completed the assignment, knit your homework to HTML and submit on Canvas.

1 Poisson regression: Chicago air quality

In this homework, you will look at the relationship between temperature, air pollution, and deaths in Chicago, between January 1, 1987 and December 31, 2000. There are 5114 rows in the data, with each row representing one day in that time period. Variables include:

- death: totals deaths on that day
- pm10median: the median density over the city of large pollutant particles
- pm25median: the median density of smaller pollutant particles
- o3median: median concentration of ozone
- so2median: median concentration of sulphur dioxide
- time: time in days (since the beginning of the study)
- tmpd: average temperature that day (Fahrenheit)

Downloading the data

First, install the `gamair` package in R. Then run the following, which will load the `chicago` dataset.

```
library(gamair)
data("chicago")
```

Questions

1. We will begin by looking at the relationship between temperature and deaths. We would like to fit a Poisson regression model, with the number of deaths on each day as the response, since the number of deaths is a count variable. Recall that our Poisson regression model makes the following assumptions:

- Poisson distribution (the response can be modeled by a Poisson distribution)
- Independence (the observations are independent)
- Shape (the shape of our regression model is correct)

The independence assumption is probably violated, since we have time series data here (i.e., observations observed sequentially over time). We'll ignore that issue in this assignment, since there isn't anything we can do about it in this class. We'll focus on the Poisson and shape assumptions.

(a) Create a histogram showing the distribution of the number of deaths. Do you think it is reasonable to use a Poisson distribution for this response?

(b) For a Poisson variable, the mean and variance are the same. Calculate the mean and variance of the number of deaths. Is it reasonable to assume that the mean and variance are the same?

(c) Now let's check the shape assumption with empirical log means plots. Here is a function to make one of these plots. It works the same as the `logodds_plot` function for logistic regression, but the response is a count variable instead of a binary variable.

```
logmean_plot <- function(data, num_bins, bin_method,
                         x, y, grouping = NULL, reg_formula = y ~ x){

  if(is.null(grouping)){
    dat <- data.frame(x = data[,x],
                      y = data[,y],
                      group = 1)
  } else {
    dat <- data.frame(x = data[,x],
                      y = data[,y],
                      group = data[,grouping])
  }

  if(bin_method == "equal_size"){
    log_table <- dat %>%
      drop_na() %>%
      arrange(group, x) %>%
      group_by(group) %>%
      mutate(obs = y,
             bin = rep(1:num_bins,
                       each=ceiling(n()/num_bins))[1:n()]) %>%
      group_by(bin, group) %>%
      summarize(mean_x = mean(x),
                mean_y = mean(obs),
                num_obs = n()) %>%
      ungroup() %>%
      mutate(log_mean = log(mean_y))
  } else {
    log_table <- dat %>%
      drop_na() %>%
      group_by(group) %>%
      mutate(obs = y,
             bin = cut(x,
                       breaks = num_bins,
                       labels = F)) %>%
      group_by(bin, group) %>%
      summarize(mean_x = mean(x),
                mean_y = mean(obs),
                num_obs = n()) %>%
      ungroup() %>%
      mutate(log_mean = log(mean_y))
  }

  if(is.null(grouping)){
    log_table %>%
      ggplot(aes(x = mean_x,
                 y = log_mean)) +
      geom_point(size=2.5) +
      geom_smooth(se=F, method="lm", formula = reg_formula) +
      theme_bw() +
      labs(x = x,
           y = "Empirical log mean count") +
      theme(text = element_text(size=25))
  } else {
    log_table %>%
      ggplot(aes(x = mean_x,
                 y = log_mean,
                 color = group,
                 shape = group)) +
      geom_point(size=2.5) +
      geom_smooth(se=F, method="lm", formula = reg_formula) +
      theme_bw() +
      labs(x = x,
           y = "Empirical log mean count",
           color = grouping,
           shape = grouping) +
      theme(text = element_text(size=25))
  }

}
```

Create an empirical log mean plot for the relationship between temperature and deaths. Does a linear function of temperature seem reasonable, or does it look like we need a transformation of temperature?

2. Now we want to fit a Poisson regression model.

   (a) Should we include an offset in our model? Explain your reasoning. If so, what should our offset be, and is this information available in the data?

   (b) Based on your EDA, write down a population Poisson regression model for the relationship between temperature and deaths, which can be fit using the `chicago` data.

   (c) Fit your model in R, and report the equation of the fitted model. What is the estimated change in the mean number of deaths for a one degree increase in temperature?

3. Next, let's check model diagnostics

   (a) Perform a χ^2 goodness-of-fit test to test whether the Poisson regression model is a good fit to our data.

   (b) Create a quantile residual plot for the fitted model to check the shape and Poisson assumptions. Do you see any issues with the fitted model?

4. To handle over-dispersion, we can use a quasi-Poisson model.

   (a) Fit a quasi-Poisson model (`family = quasipoisson`), and report the estimated dispersion parameter $\widehat{\phi}$. What is the relationship between the standard errors for the quasi-Poisson fit and the standard errors for the Poisson fit?

   (b) Are you estimated coefficients $\widehat{\beta}$ different for the quasi-Poisson fit vs. the Poisson fit?

   (c) Using the quasi-Poisson fit, calculate a confidence interval for the change in the average number of deaths associated with a one degree increase in temperature, holding pollution constant.