STA 214 Homework 9

Due: Friday, April 7, 12:00pm (noon) on Canvas.

Instructions: There are two parts to this assignment. In the first part, you will use ZIP models to model the number of fish caught by groups of campers in a state park. In the second part, you will conduct a small simulation study to compare the performance of ZIP models with regular Poisson models when there is zero-inflation in the data.

Getting started: Begin by downloading the HW9 template from the course website:

`https://sta214-s23.github.io/homework/hw_09_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.

Data Analysis

In the first part of this assignment, you will model the number of fish caught by campers in a state park. We have a sample of 250 groups of park guests who visited the state park. For each group, we record:

- `count`: the number of fish caught by the group during their stay
- `camper`: whether the group brought a camper van into the park (0 = no, 1 = yes)
- `child`: the number of children in the group
- `persons`: the total number of people in the group (including children)
- `LOS`: length of stay (in days)

You can load the data into R by

```
books <- read.csv("https://sta214-s23.github.io/homework/fish2.csv")
```

1. Let's begin by thinking about the variables available in the data.
   (a) Create a plot showing the distribution of the number of fish caught.
   (b) Using the plot, explain why a zero-inflated Poisson model will be helpful for modeling this variable. What would the latent variable represent in this data? (Recall that in the campus drinking data, the latent variable indicated whether the student ever drank; in the Framingham heart data, the latent variable indicated whether the participant was a smoker or non-smoker).
   (c) Explain why $\log(LOS)$ should be used as an offset term when modeling the number of fish caught.

2. Park rangers at the state park wonder whether groups with many children tend to catch fewer fish. They ask you to fit a model to investigate their hypothesis, and they want you to account for the total number of visitors in the group and whether the group brought a camper van (they suspect that camper vans make noise that scares away the fish).

   (a) Write down a ZIP model that allows you to investigate the researchers' hypothesis.
   - Write down the probability function, and the logistic and Poisson components
   - Without further information, we generally include the same explanatory variables in both the logistic and Poisson components of the ZIP model
   - Include the offset term from question 1(c). Offset terms are included *only* in the Poisson component.

   (b) Use the `zeroinfl` function from the `pscl` package to fit the model in R:

```
m1 <- zeroinfl(count ~ ... | ...,
               offset = ...,
               data = fish2)
```

   (c) Interpret the fitted coefficients for `child` in the logistic and Poisson components of the model.

   (d) Construct confidence intervals for the two coefficients in (c).

3. Now let's make some predictions.

   (a) What is the predicted probability that a 3-person group with 0 children and no camper van, who stayed for 3 days, caught *at least one* fish during their stay?

   (b) What is the estimated number of fish caught by a 5-person group with 3 children in a camper van, who stayed for 7 days?

4. Finally, let's use hypothesis testing to answer the research question. The park rangers want to know whether groups with more children catch fewer fish.

   (a) Are groups with more children less likely to go fishing? Carry out a hypothesis test: state the null and alternative hypotheses, report the test statistic, and make a conclusion in context.

   (b) D groups with more children, who *do* go fishing, catch fewer fish per day? Carry out a hypothesis test: state the null and alternative hypotheses, report the test statistic, and make a conclusion in context.

Simulation study

We introduced ZIP models as a method to handle zero-inflation in count data. But what happens if we just ignore the zero-inflation, and fit a simpler count model (e.g., a Poisson model) instead? The goal of this section is to use simulations to explore what happens when we ignore possible zero-inflation.

The code below generates data $(X_1, Y_1), ..., (X_n, Y_n)$ from the model

$$X_i \sim N(0, 1)$$

$$P(Y_i = y) = \begin{cases} e^{-\lambda_i}(1 - \alpha_i) + \alpha_i & y = 0 \\ \dfrac{e^{-\lambda_i}\lambda_i^y}{y!}(1 - \alpha_i) & y > 0 \end{cases}$$

$$\log\left(\frac{\alpha_i}{1 - \alpha_i}\right) = \gamma_0$$

$$\log(\lambda_i) = \beta_0 + \beta_1 X_i$$

and then fits two models: a Poisson model, and a ZIP model.

Code:

```
# simulate data
n <- 1000
x <- rnorm(n)
gamma0 <- 0
beta0 <- 1
beta1 <- 1
alpha <- exp(gamma0)/(1 + exp(gamma0))
lambda <- exp(beta0 + beta1*x)
z <- rbinom(n, 1, prob=alpha)
y <- 0*z + rpois(n, lambda)*(1 - z)

# fit Poisson and ZIP models
m_pois <- glm(y ~ x, family = poisson)
m_zip <- zeroinfl(y ~ x | x)
```

5. (a) Run the provided code, and report the estimated coefficients for the Poisson and ZIP models. Which estimates of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are closer to the truth – the Poisson estimates or the ZIP estimates?

   (b) Calculate a 95% confidence interval for β_1 using the Poisson model. Does your interval contain β_1?

   (c) Calculate a 95% confidence interval for β_1 using the ZIP model. Does your interval contain β_1?

6. Repeat question 5 many times. What fraction of your confidence intervals contain β_1 for the Poisson model? For the ZIP model?

7. In the simple ZIP model above, γ_0 controls the amount of zero-inflation. When $\gamma_0 = 0$, 50% of the data will be excess 0s. Smaller (more negative) values of γ_0 result in less zero-inflation, while larger values of γ_0 result in more zero-inflation.

(a) Repeat question 6 for different values of γ_0, and make a plot showing the coverage of Poisson and ZIP confidence intervals for different values of γ_0.

(b) At what value of γ_0 do we see a noticeable difference in coverage for the Poisson and ZIP model intervals?