ST 518 - HW2

Nora Quick

```
library(arm)
library(Sleuth3)
library(tidyverse)
library(vcdExtra)
library(magrittr)
library(MASS)
library(pscl)
```

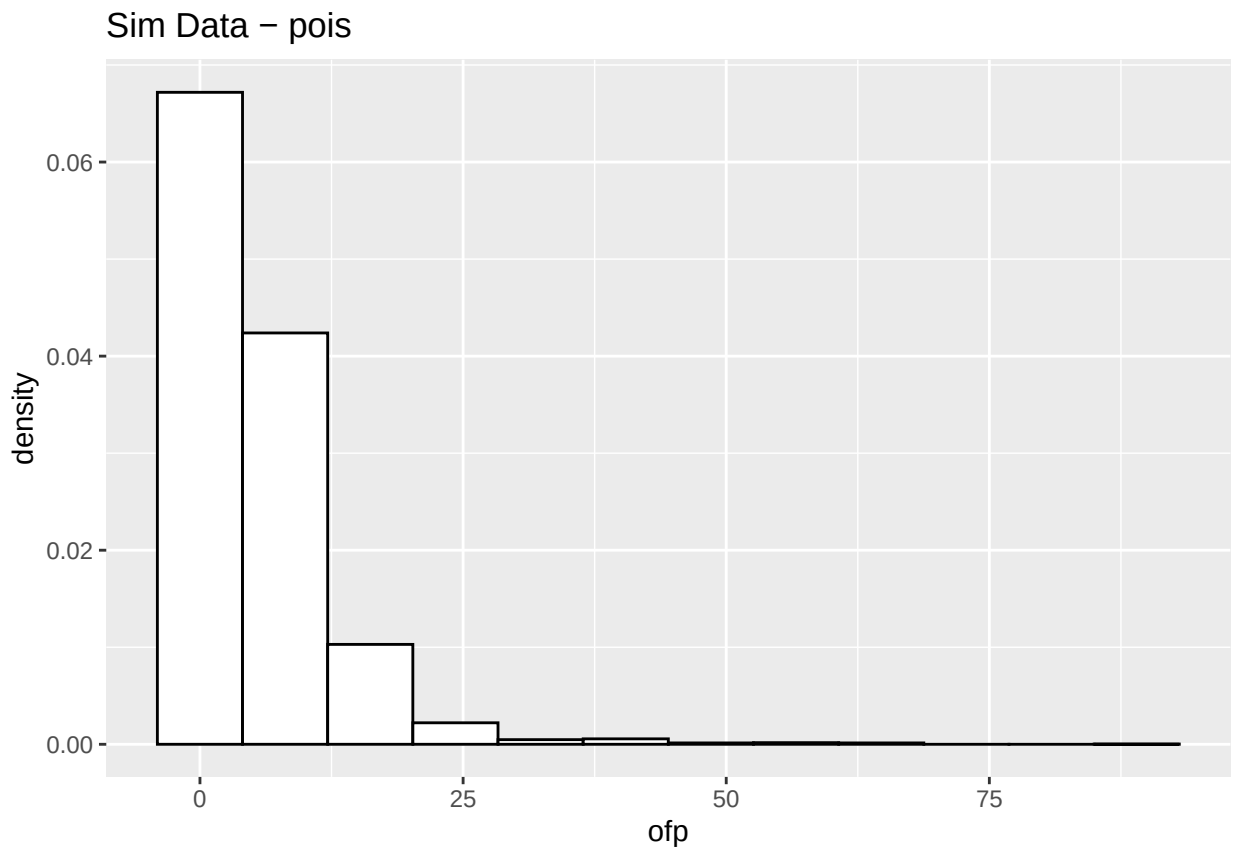
```
load("DT.rda")
#dt
```

R Questions

1.

a. Produce a histogram of the dependent variable. What types of statistical models (.e.g, Poisson regression, hurdle Poisson model, etc.) should you consider for these data? Give a brief justification for your answer.

```
ggplot(data = dt, aes(x = ofp, y = ..density..)) +
    geom_histogram(bins = 12, colour = "black", fill = "white") +
    ggtitle("Sim Data - pois")
```



I think that the zero-inflated negative binomial model due to the massive skewing to the right and the large number of zeros that appear to be in the data.

If I were to choose another I would like to compare it to a hurdle model to see the difference between the two models.

b. Fit the models you indicated in part (a), using the identical set of explanatory variables in each, and report the AIC for each.

```
mod_zinb <- zeroinfl(ofp ~ hosp + health + numchron + gender + school + privins,
  dist = "negbin", data = dt,)
#summary(mod_zinb)
AIC(mod_zinb)
```

```
## [1] 24215.29
```

```
mod_hurdle <- hurdle(ofp ~ hosp + health + numchron + gender + school + privins,
  dist = "poisson", data = dt,)
#summary(mod_hurdle)
AIC(mod_hurdle)
```

```
## [1] 32300.9
```

For the zero-inflation negative binomial model the AIC is 24215 and for the hurdle model the AIC is 32301.

c. Based on parts (a) and (b) and considering interpretation of the models in the context of the data, write a sentence summarizing your findings in terms of which model seems most appropriate for these data.

Based on AIC alone I believe that the hudle model is better as it has the smaller (but still large) value and on top of that it indicated most of the variables as significant which I find more reasonable than only the socioeconomic variables that the zero-infaltion negative binomial model shows.

Conceptual Questions

##2.

a. School administrators study the attendance behavior of high school juniors and take the number of days absent as the response variable.

I believe that a zero-inflated model would be best here because we could justify an independent reason for the excess zeros. For example, transfer students could have a lack of data or the occasional miss click/mark of a student being absent when they weren't. (There is the assumption there will be excess zeros = zero-inflated).

b. Wildlife biologists want to model how many fish are being caught by fishermen at a state park. All park visitors are asked how many fish they caught.

I believe that a zero-inflated model wold be best here because there is surely going to incorrect and inconsistent data because of asking the visitors how many they caught. Some may light, others might forget, and some may never answer the question. (There is the assumption there will be excess zeros = zero-inflated).

c. Researchers are interested in creating a stock trading model for investors. The response is trades per week made by each investor.

I believe the hurdle model would be best here because I believe this would be mostly an excess of sampling zeros and not the addition of structural zeros being introduced. In other words, there aren't outside veriables that might be a reason.

d. Researchers want to create a model for loan defaults. They take the number of outstanding payments that exist for each of a random sample of loans.

I believe a zero-inflated model would be best here because there are outside factors that would contribute to an outstanding payment. Income or residential area could be factors that account for excess zeros. (There is the assumption there will be excess zeros = zero-inflated).