Final Exam

Jakob Orel

05/12/2021

Instructions

- Please submit any files used to a Final Exam folder on on Google Drive (please make a folder)
- You must submit your exam by Wednesday May 12th at 3pm
- This is an open book and note exam. You *cannot* use another students notes or scripts, talk to others about the exam, or use the internet (other than to get the exam/data)
- If you have questions email me right away.
- Your solution should be written as a report. This means that I should be able to read through it and see your evidence that you are discussing. *I should not need to look at your R code to follow your work.*
- Label your tables and plots. You can quickly add "Figure a" under a plot to reference it. This is most easily done through either word (copying plots or tables from R) or RMarkdown (with a working knitted file).
- You need to also need to submit your code but I will be evaluating the report, not the code used to make the tables/plots.
- I have included the RMarkdown file to create this exam that you can choose to use or not.
- Make sure to conduct a thorough EDA and justify each of your decisions throughout the process.
- Convince me you know the topics and write report in a way someone not in our class could understand and be convinced your model and interpretations are legitimate
- Since some students expressed the Midterm taking 6-8 hours, I have not made this exam longer than the midterm.

More Fish

The number of fish caught (*count*), persons in the party (*persons*), the number of children in the party (*child*), whether or not they brought a camper into the park (*camper*), and the length of stay (*LOS*) were recorded for 250 camping parties. The data can be found in **fish2.csv** (source: UCLA Statistical Consulting Group (2018)). Create and assess a model for the number of fish caught.

Consider all the types of models we discussed this block. This does not mean to fit all those models but to discuss their appropriateness or lack of appropriateness for your data. Also make sure to interpret your final chosen model at the end.

Exploratory Data Analysis (EDA)

Single Variable

We should investigate the response variable in question.

```
ggplot(fish, aes(x=count))+
  geom_histogram(fill='white', color='black')+
  labs(x="Number of fish", y="Frequency", title="Figure 1")
```

Figure 1

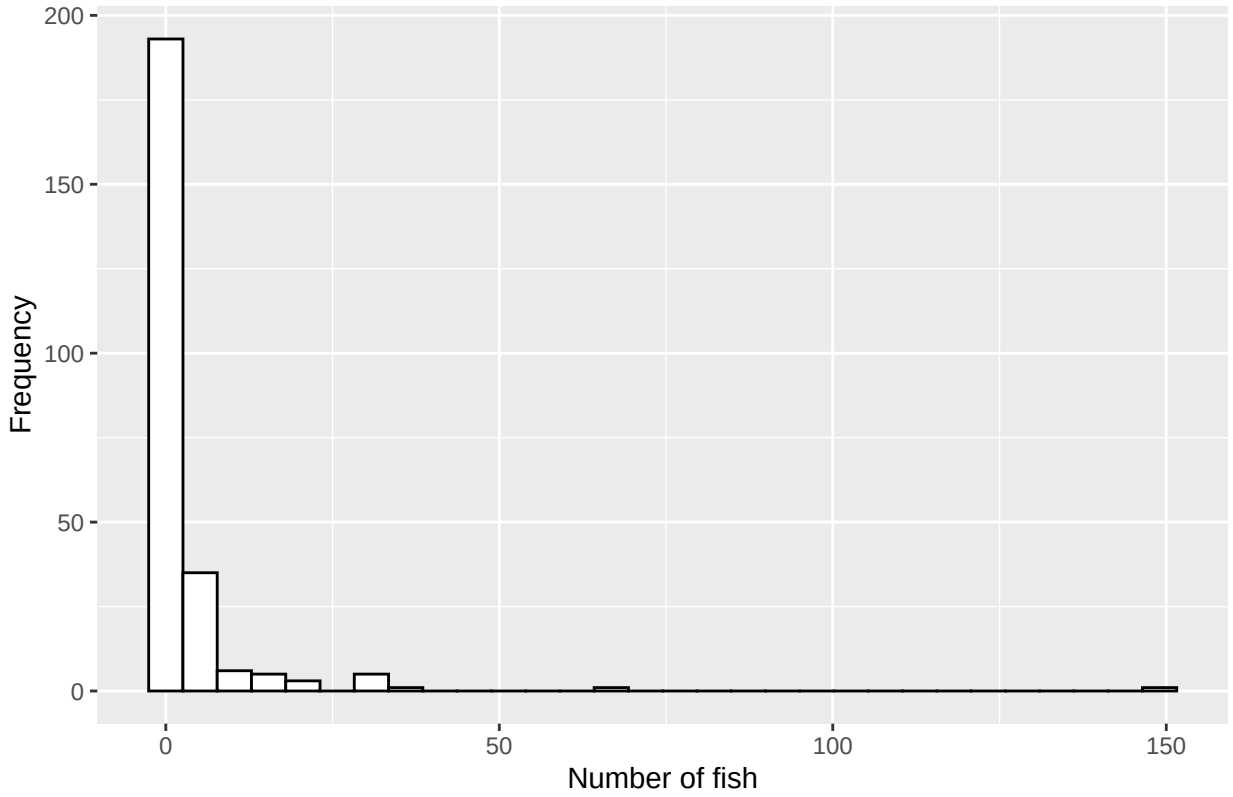


Figure 1 indicates there is a very high proportion of zero counts in our response variable. The histogram is right skewed and the variable is a count over time which indicates a Poisson response.

```
obs.table <- tally(group_by(fish, count)) %>% mutate(prop = round(n/sum(n),3))

g.obs <- obs.table %>% ggplot(aes(x=count,y=prop))+
  geom_bar(stat='identity')+
  labs(x="Number of Fish", y="Proportion", title="a) Observed")+
  coord_cartesian(ylim=c(0,.6))

sum1 <- fish %>%
  summarise(lambda = mean(count),
            maxCount = max(count))

possible.values <- with(sum1, 0:maxCount)

# Simulate the probability of counts using Poisson distribution and observed mean
model.prob <- with(sum1, dpois(possible.values,
                               lambda))

pois.model <- data.frame(possible.values, model.prob)

g.model <- ggplot(pois.model, aes(x=possible.values, y=model.prob))+
  geom_bar(stat="identity")+
  labs(x="Number of Fish", y="Probability", title="b) Poisson Model")+
  coord_cartesian(ylim=c(0,.6))
```

```
grid.arrange(g.obs, g.model, top="Figure 2")
```

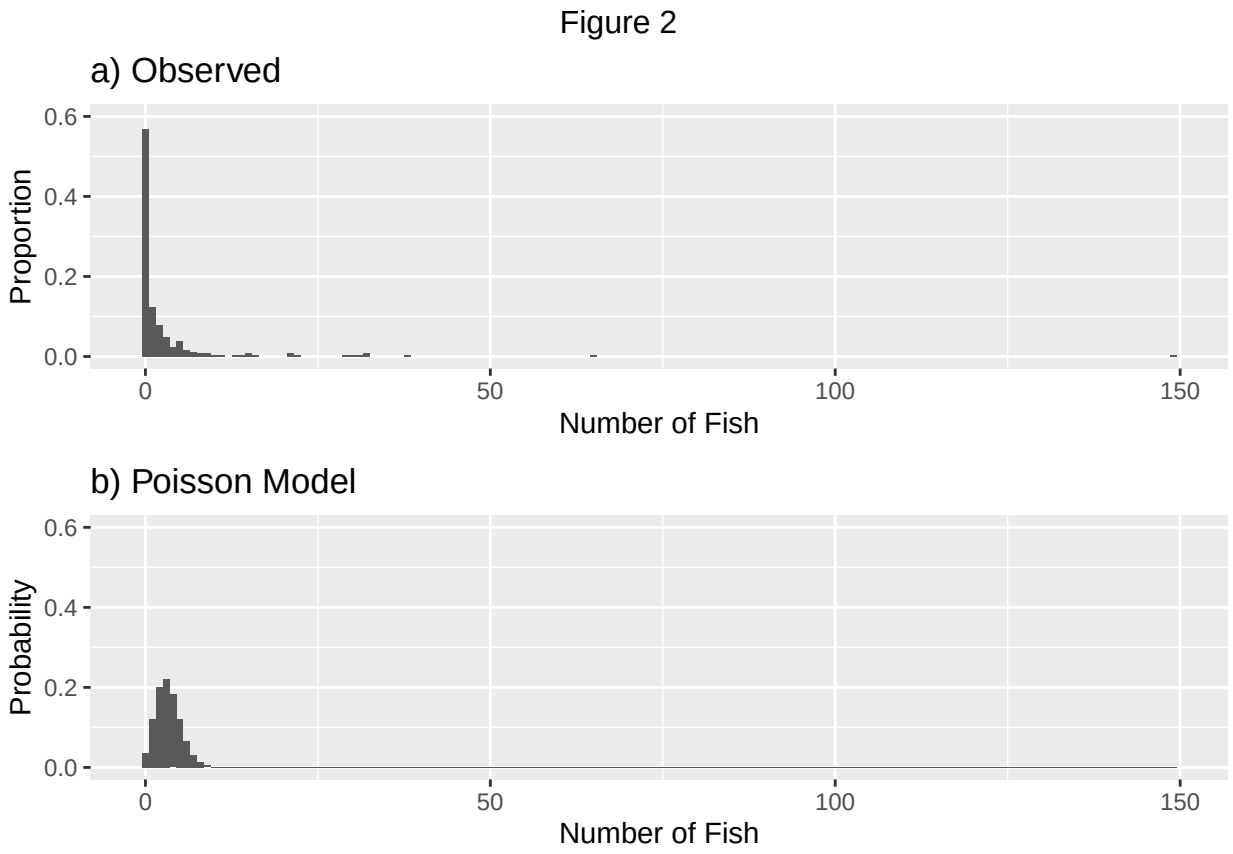


Figure 2 shows what a normal Poisson distribution would look like for the count of number fish based on the empirical mean of the data. You can see the observed distribution is quite different from the Poisson model. There is a high proportion of zeros and there is also quite a bit of variance with what appears as an outlier at 149.

This high proportion of zeros leads me to believe a zero-inflated model may be appropriate. There may be fishers who catch zero fish or non-fishers who do not fish at all. This would lead to a higher number of zeros that could be explained by a logistic model along with a Poisson regression model. We should investigate other variables that would be useful in explaining the response variable.

```
# Frequency table
table(fish$persons)
```

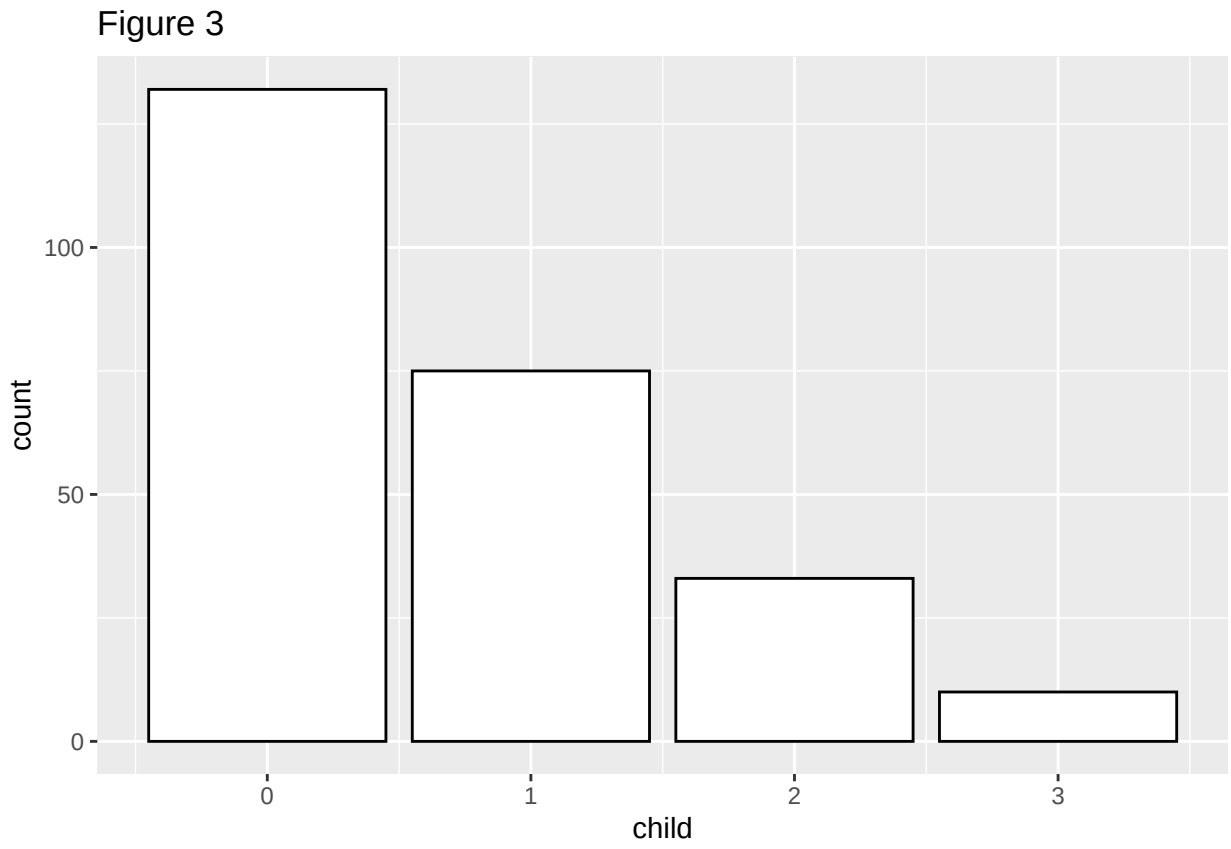
```
##
##  1  2  3  4
## 57 70 57 66
```

```
# Relative frequency table
table(fish$persons)/nrow(fish)
```

```
##
##     1     2     3     4
## 0.228 0.280 0.228 0.264
```

We can see that camping parties had a roughly equally distributed number of adults on their stay ranging from 1 to 4. The most common being 2 adults (likely couples). I would assume that having more adults would mean the party would catch more fish. This will have to be explored further.

```
ggplot(fish, aes(x=child))+
  geom_bar(fill='white', color='black')+labs(title="Figure 3")
```



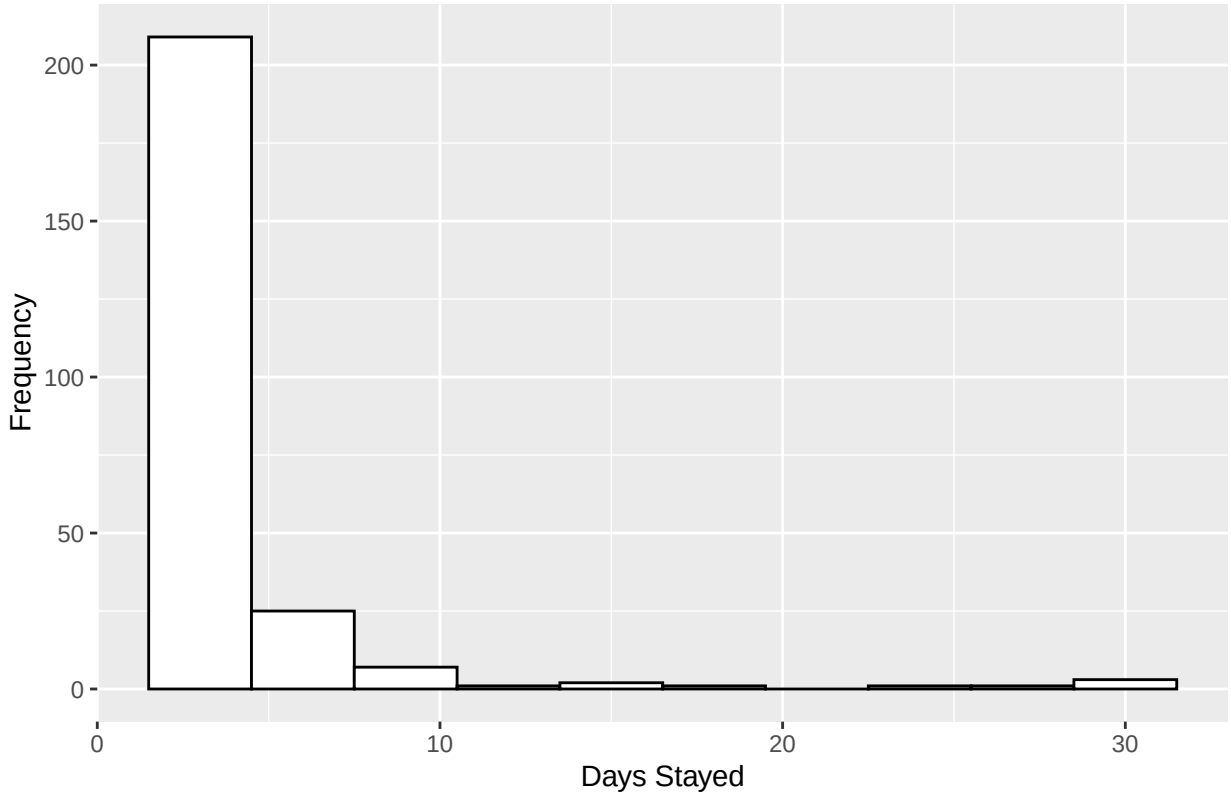

```
# Relative frequency table
table(fish$child)/nrow(fish)
```

```
##
##     0     1     2     3
## 0.528 0.300 0.132 0.040
```

We can see the majority of camping parties do not have children (over 50%) and usually only 1 is present if at all in Figure 3. It may be interesting to see if taking children would have any effect on the number of fish that are caught. It should be noted that children under this variable are not counted in the persons variable for party size.

```
ggplot(fish, aes(x=LOS))+
  geom_histogram(binwidth=3, fill='white', color='black')+
  labs(x="Days Stayed", y="Frequency", title="Figure 4")
```

Figure 4



The histogram in Figure 4 shows that the length of stay (days) is usually less than 5 days. One might expect that the number of fish caught would increase as the length of stay increases because they have more time to catch fish. This may be useful to use as an offset in our model. This should be explored further.

```
table(fish$camper)
```

```
##
##   0   1
## 103 147
```

There are more parties who bring campers to the park, but I am unsure if this will have an effect on the number of fish caught.

Comparing Variables

I will now compare variables to explore the significance of variables in predicting the number of fish caught.

```
ggplot(fish, aes(x=factor(persons), y=count))+
  geom_boxplot() + labs(x="Persons", y="Number of Fish Caught", title="Figure 5")
```

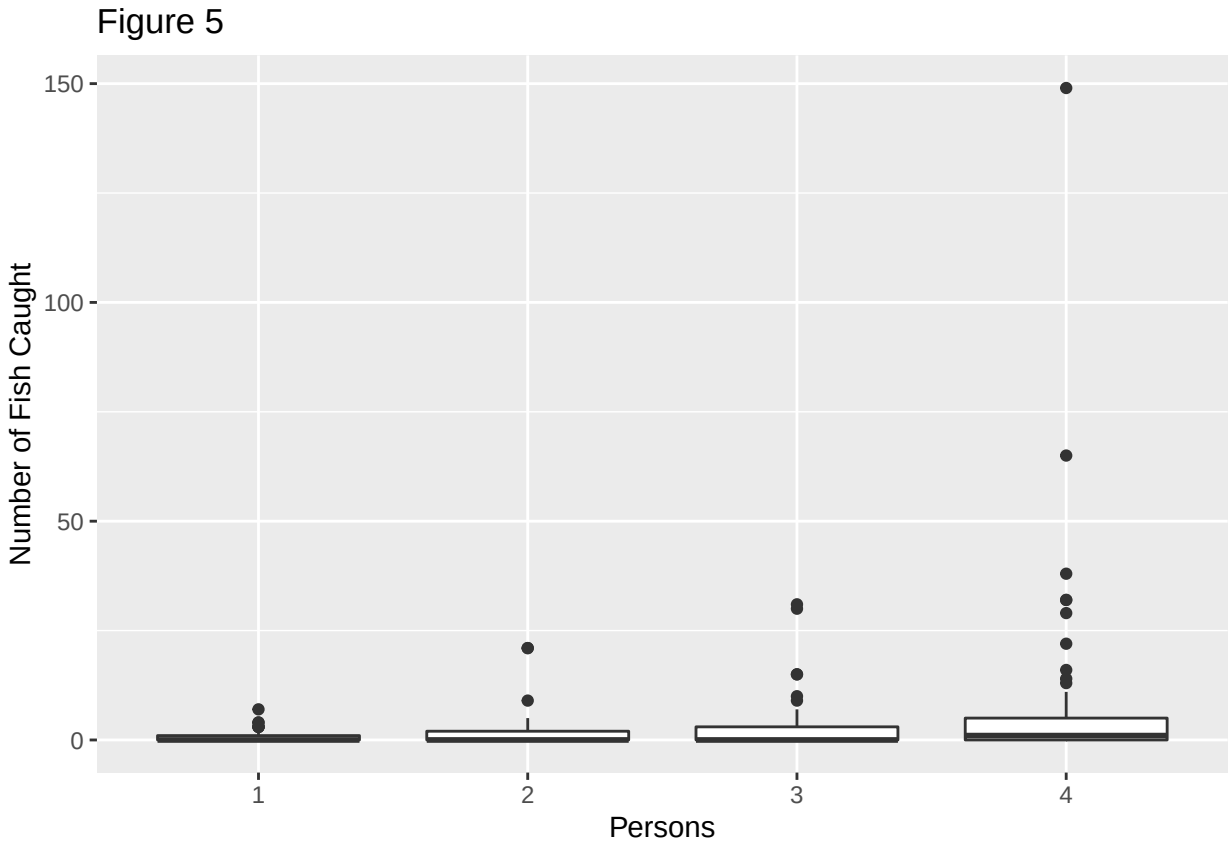
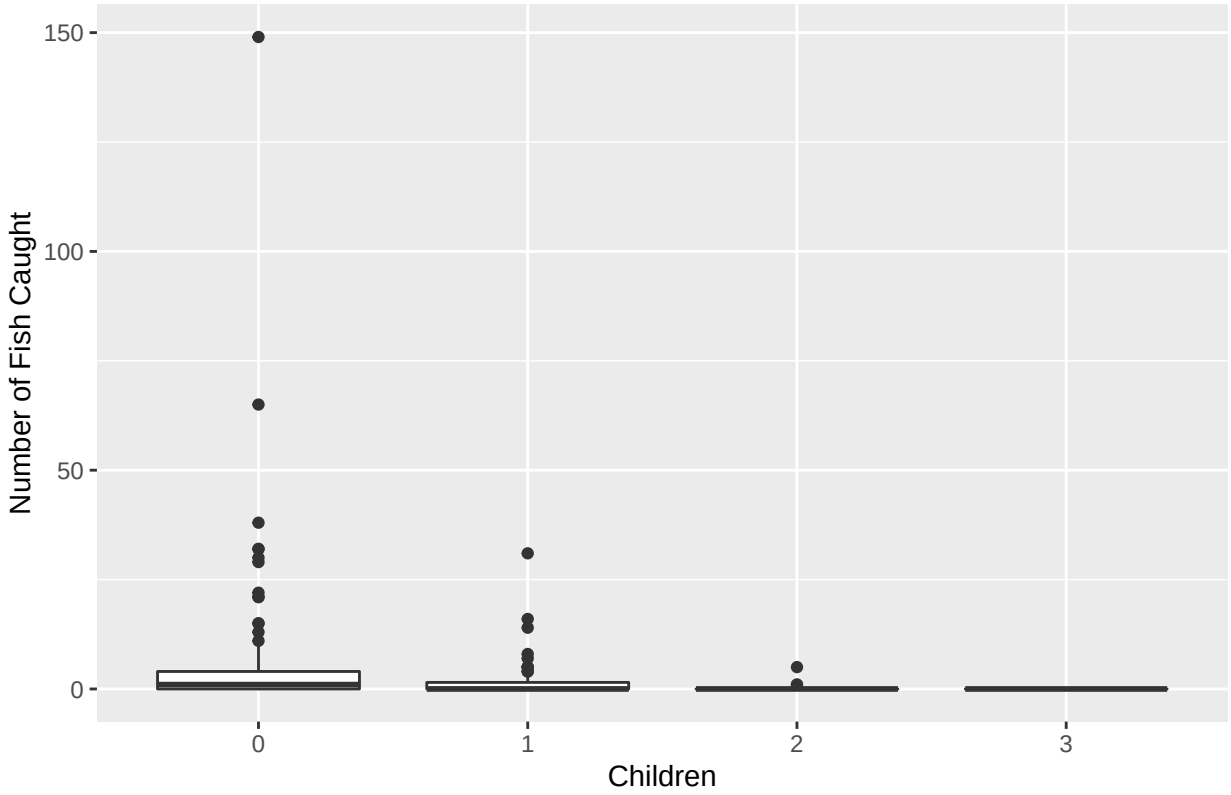


Figure 5 appears to show that as the number of adults in the camping party increase the number of fish caught tends to generally increase. There also appears to be a relatively strong outlier near 150 that could affect our modeling as it is a relatively extreme distance from the whisker of the fourth box.

```
ggplot(fish, aes(x=factor(child), y=count))+
  geom_boxplot() + labs(x="Children", y="Number of Fish Caught", title="Figure 6")
```

Figure 6

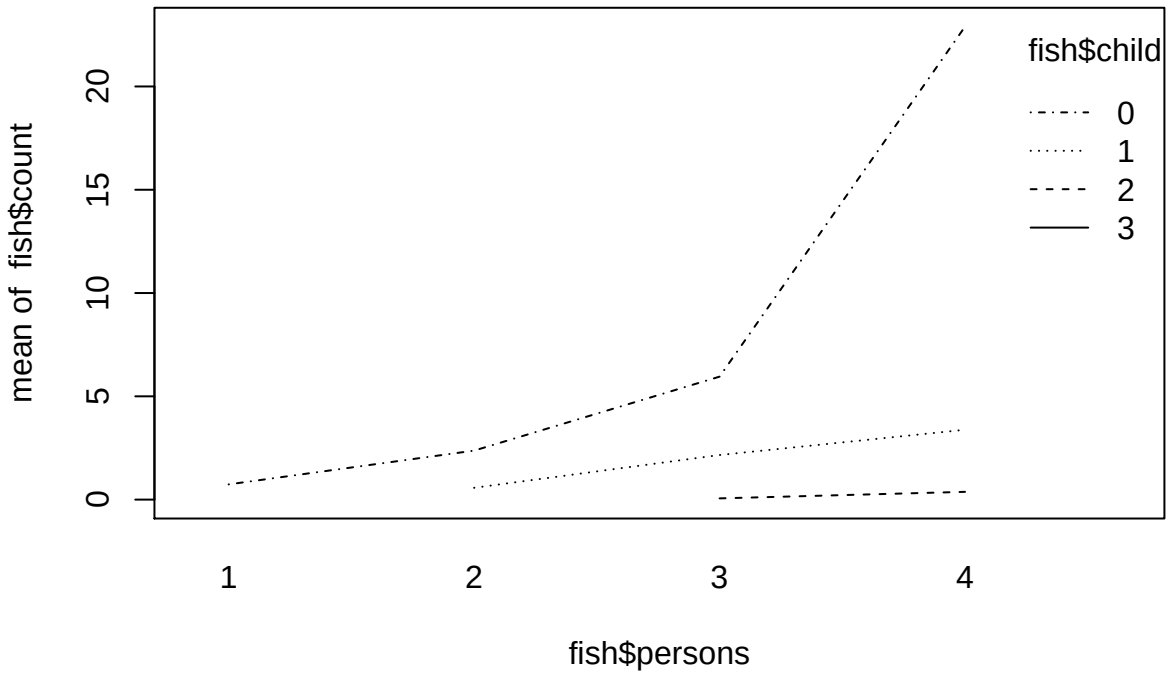



```
fish %>% group_by(child) %>%
  summarise(mean= mean(count),
            sd = sd(count))
```

```
## # A tibble: 4 x 3
##   child  mean     sd
##   <int> <dbl>  <dbl>
## 1     0 5.19  15.4
## 2     1 1.76   4.46
## 3     2 0.212  0.893
## 4     3 0      0
```

Figure 6 and the following table appear to indicate the inverse trend seen in Figure 5. It appears that when there are more children in the party the number of fish decreases. There is also quite a bit of deviance, so this trend may not be significant.

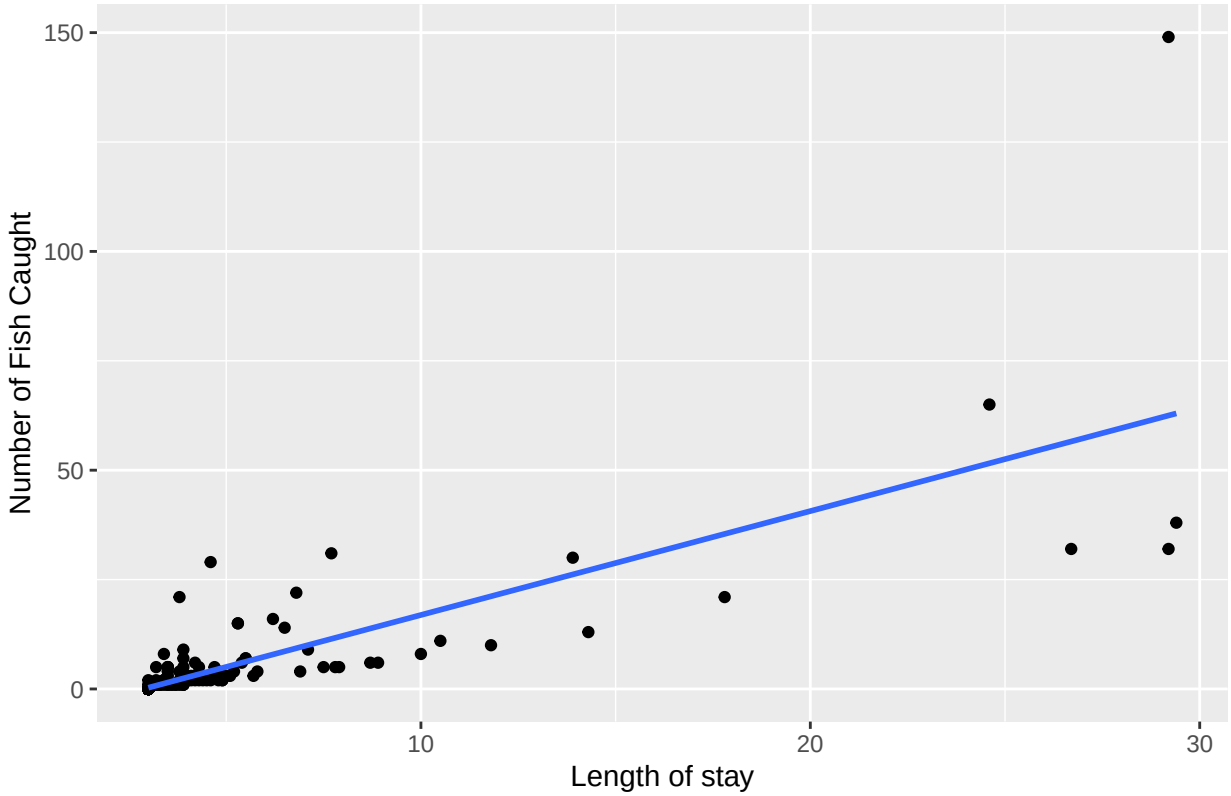
```
interaction.plot(fish$persons, fish$child, fish$count)
```



The interaction plot above has different slopes for the different number of children and different number of persons in the party. This inverse trend and differing slope may indicate an interaction may be useful to test in our modeling.

```
ggplot(fish, aes(x=LOS, y=count))+
  geom_point() +
  geom_smooth(method="lm", se=F)+
  labs(x="Length of stay", y="Number of Fish Caught", title="Figure 7")
```

Figure 7




```
cor(fish$count, fish$LOS)
```

```
## [1] 0.7860501
```

Figure 7 indicates there may be a linear relationship (r=.786) with the length of stay and the number of fish caught. This is expected and may be useful to apply as an offset in our model. This would then change our interpretation into a rate for the number of fish caught per day.

```
fish %>% group_by(camper) %>% summarise(mean = mean(count),
                                        var = var(count),
                                        median = median(count),
                                        n= n())
```

```
## # A tibble: 2 x 5
##   camper  mean   var median     n
##    <int> <dbl> <dbl>  <int> <int>
## 1      0  1.52  21.1      0   103
## 2      1  4.54 212.       1   147
```

The above table shows that those with campers may have a slightly higher number of fish caught, but there is a lot of variance that may affect the significance of this variable in a model.

Checking Assumptions

I will check the assumptions for a Poisson regression model as this will be the most appropriate model for this response and research question. As we will also explore a zero-inflated model, I will explore the assumptions for the logistic part of that model.

Poisson Response and Independence

We have already determined based on Figure 1 and the provided context of the scenario that the number of fish caught is a Poisson response. I will assume that each observation is independent of one another (different parties and unrelated to each other) based on the information provided from the context.

Mean=Variance

```
fish %>% mutate(LOSgroups = cut(LOS, breaks=10)) %>%
  group_by(LOSgroups) %>%
  summarise(mean= mean(count),
            var = var(count),
            n = n())
```

```
## # A tibble: 8 x 4
##   LOSgroups    mean     var     n
##   <fct>       <dbl>   <dbl> <int>
## 1 (2.97,5.64]  1.26    9.71   226
## 2 (5.64,8.28] 10.7    82.8     11
## 3 (8.28,10.9]  7.75    5.58     4
## 4 (10.9,13.6] 10      NA        1
## 5 (13.6,16.2] 21.5   144.       2
## 6 (16.2,18.8] 21      NA        1
## 7 (24.1,26.8] 48.5   544.       2
## 8 (26.8,29.4] 73    4341        3
```

The variance is not roughly equal to the mean of number of fish caught for the different bins of length of stay, but this is difficult to examine if it will be an issue as there are very unequal number of observations in each group and the outlier is heavily affecting the variance of the last group. I will continue assuming this is satisfied and will explore further if models do not fit the data appropriately because of overdispersion.

```
fish %>% group_by(persons) %>%
  summarise(mean= mean(count),
            var = var(count),
            n = n())
```

```
## # A tibble: 4 x 4
##   persons  mean    var     n
##     <int> <dbl>  <dbl> <int>
## 1       1 0.737   1.88    57
## 2       2 1.47   13.6     70
## 3       3 2.93   40.2     57
## 4       4 7.76  438.      66
```

The variance greatly increases as the number of persons increase and there appears to be an issue with overdispersion. The variance appears to be larger than the mean of number of fish caught.

```
fish %>% group_by(child) %>%
  summarise(mean= mean(count),
            var = var(count),
            n = n())
```

```
## # A tibble: 4 x 4
##   child  mean     var     n
##   <int> <dbl>   <dbl> <int>
## 1     0 5.19  238.      132
## 2     1 1.76   19.9      75
```

```
## 3     2 0.212   0.797    33
## 4     3 0       0        10
```

There may be similar effects of overdispersion seen in the child variable when looking at mean vs variance of counts of fish caught, but again the distribution of observations is very unequal.

```
fish %>% group_by(camper) %>%
  summarise(mean= mean(count),
            var = var(count),
            n = n())
```

```
## # A tibble: 2 x 4
##   camper  mean   var     n
##    <int> <dbl> <dbl> <int>
## 1      0  1.52  21.1   103
## 2      1  4.54 212.    147
```

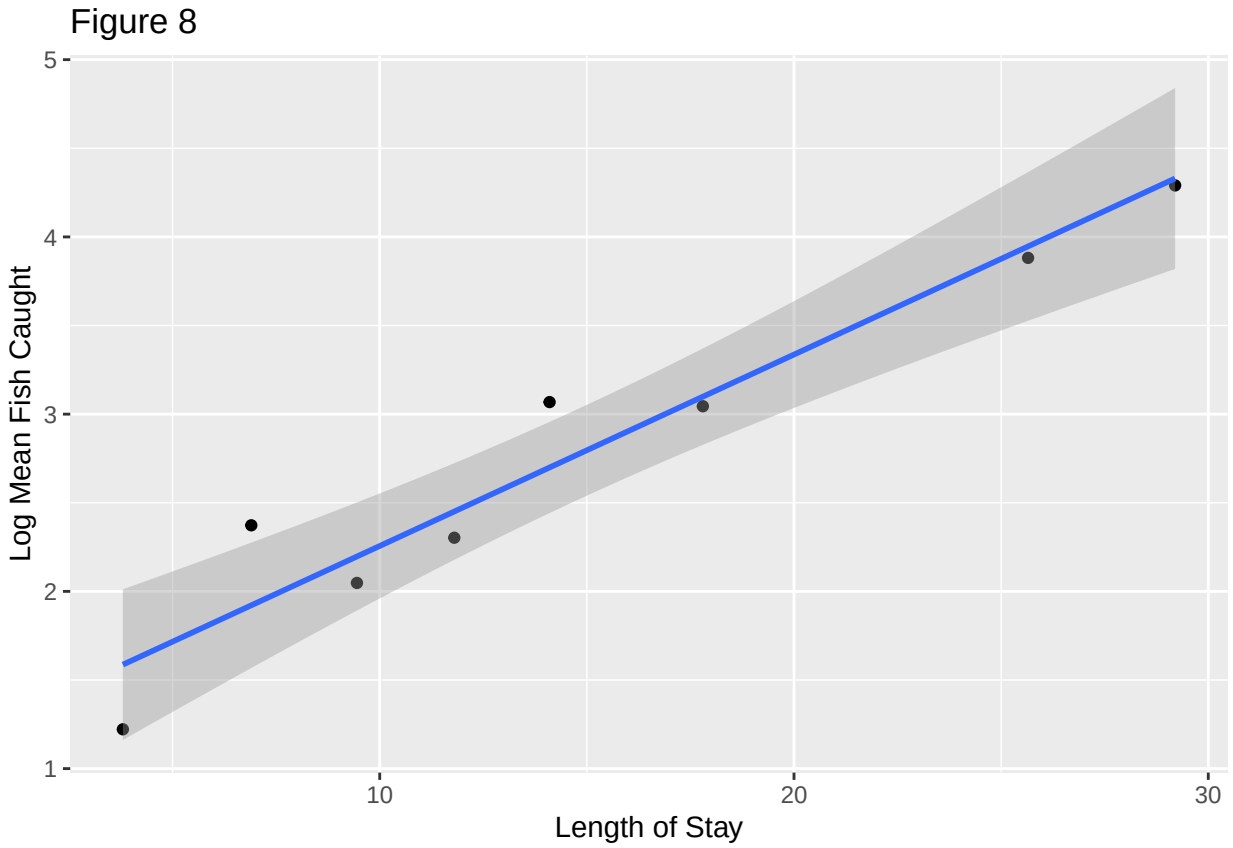
The variance of the group with campers is greater than the mean and may be an indicator of overdispersion. We will deal with these issues later in the modeling stage.

Linearity

We should determine that the log of the mean rate has a linear relationship with the predictors for a Poisson regression model.

```
logmean_table <- fish %>%
  filter(count > 0) %>%
  mutate(LOSgroups = cut(LOS, breaks=10)) %>%
  group_by(LOSgroups) %>%
  summarise(logmean = log(mean(count)),
            median_los = median(LOS))

ggplot(data = logmean_table, aes(x = median_los, y = logmean)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x="Length of Stay", y="Log Mean Fish Caught", title="Figure 8")
```



There is a clear linear relationship between the length of stay and the log of mean fish caught which satisfies the linearity assumption for the Poisson part of the zero-inflated model (count > 0).

```
logmean_table <- fish %>%
  filter(count > 0) %>%
  group_by(persons) %>%
  summarise(logmean = log(mean(count)),
            median_persons = median(persons))

ggplot(data = logmean_table, aes(x = persons, y = logmean)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x="Persons", y="Log Mean Fish Caught", title="Figure 9")
```

Figure 9

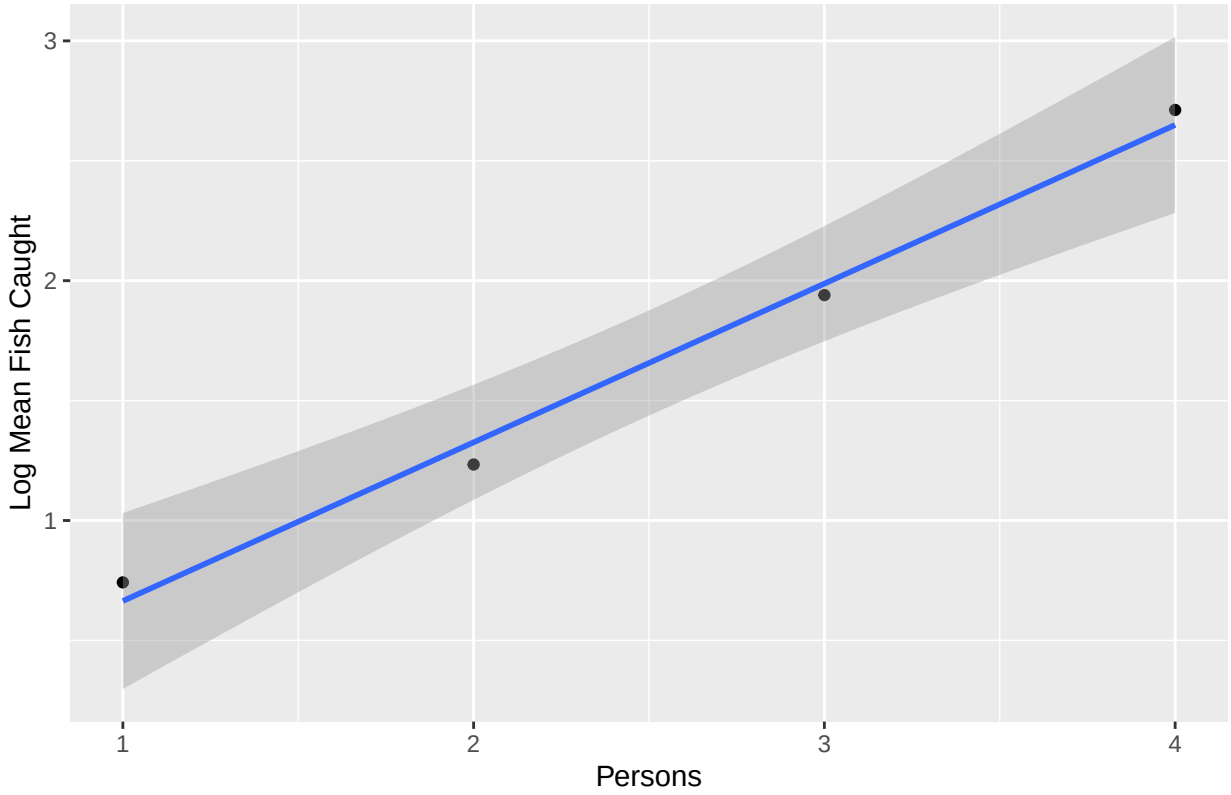
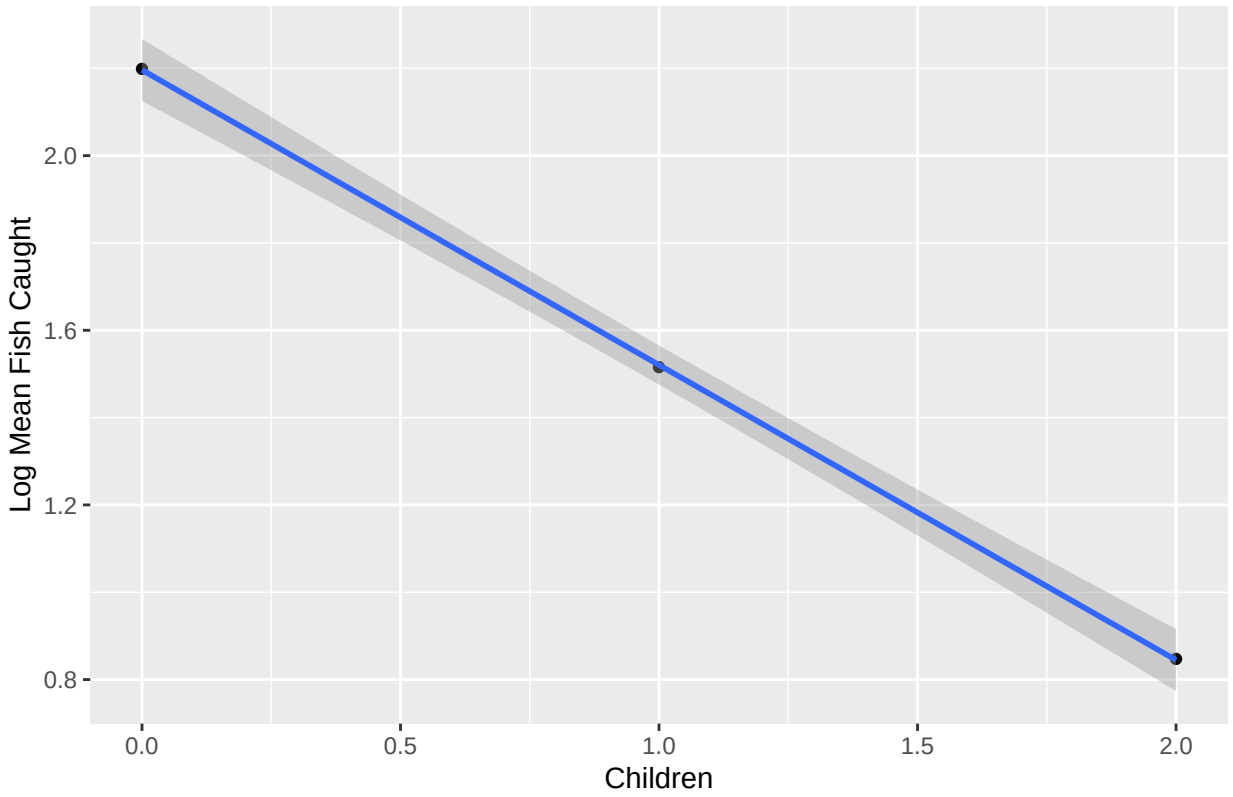


Figure 9 shows a linear relationship with log mean count and the number of persons in the party. This also satisfies the linearity assumption. Persons could be considered as a continuous variable, but it only contains 4 numbers so it can also easily be considered as a factor.

```
logmean_table <- fish %>%
  filter(count > 0) %>%
  group_by(child) %>%
  summarise(logmean = log(mean(count)),
            median_child = median(child))

ggplot(data = logmean_table, aes(x = child, y = logmean)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x="Children", y="Log Mean Fish Caught", title="Figure 10")
```

Figure 10



This again satisfies the linearity assumption, but there are only 3 levels for children so this is not really useful to use as a check for linearity.

Modeling

We have discussed many models, but there are only a few that are appropriate in this scenario. Linear Least Squares Regression would not be appropriate because there is not constant variance and the response is not normally distributed. This would cause several issues in the assumptions of the model which will affect the stability and accuracy of the model. This scenario is not appropriate for Logistic Regression or Likelihood models because we are trying to model/explain the number of fish caught which is not a probability or likelihood of an event occurring. We may include a part of logistic regression in a zero-inflated Poisson regression model.

This scenario would be appropriate for a Poisson regression model as it is representing a count of an event over a time interval. Furthermore, this would be best suited by a zero-inflated Poisson regression model because there is a high proportion of zeros in the count. There is a possibility of groups being fishers who simply caught zero fish and there could also be groups that caught zero fish because they are not fishers (did not actively fish during their stay). The zero-inflated model can estimate the proportion of non-fishers based on the reported zeros. A hurdle model would be used if there was only one group of true zeros (zero fish or at least one fish). This is not the case in the provided context.

In the EDA, I noticed there could be a factor of overdispersion in our data. This could be explored by using a quasipoisson model to see if our predictors are truly significant.

I will first start by fitting a regular Poisson model and then move on to zero-inflated (ZIP) models to see the improvement.

```r
model1 <- glm(count~LOS+persons+child, family=poisson, data=fish)
summary(model1)
```

```
##
## Call:
## glm(formula = count ~ LOS + persons + child, family = poisson,
##     data = fish)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8994  -1.4231  -1.1244  -0.0337  10.8103
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.894270   0.132591  -6.745 1.53e-11 ***
## LOS          0.094289   0.004102  22.984  < 2e-16 ***
## persons      0.623956   0.046559  13.401  < 2e-16 ***
## child       -1.095243   0.085415 -12.823  < 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: 2958.37  on 249  degrees of freedom
## Residual deviance:  910.12  on 246  degrees of freedom
## AIC: 1255.2
##
## Number of Fisher Scoring iterations: 6
```

From the EDA, I assumed that LOS, persons, and child would be significant in predicting the number of fish caught because of their relationship to count.

```r
model2 <- glm(count~LOS+persons+child+camper, family=poisson, data=fish)
summary(model2)
```

```
##
## Call:
## glm(formula = count ~ LOS + persons + child + camper, family = poisson,
##     data = fish)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.8205  -1.4761  -1.0308   0.0215  11.3139
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.066757   0.147775  -7.219 5.25e-13 ***
## LOS          0.089125   0.004451  20.024  < 2e-16 ***
## persons      0.632011   0.046649  13.548  < 2e-16 ***
## child       -1.097120   0.085427 -12.843  < 2e-16 ***
## camper       0.280888   0.100099   2.806  0.00501 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 2958.37  on 249  degrees of freedom
## Residual deviance:  902.06  on 245  degrees of freedom
## AIC: 1249.1
##
## Number of Fisher Scoring iterations: 6
```

```
anova(model2, model1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: count ~ LOS + persons + child + camper
## Model 2: count ~ LOS + persons + child
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       245     902.06
## 2       246     910.12 -1  -8.0627 0.004519 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In model2, all predictors are significant at the 0.05 significance level according to the Wald test in the coefficients summary. The drop in deviance test from the significant model1 to model2 (adding camper), indicates the drop in deviance is significant (p=0.004519). This indicates camper is also a significant variable.

```
cat("Goodness of Fit Test p-value= ", pchisq(model2$deviance, model2$df.residual, lower.tail=F))
```

```
## Goodness of Fit Test p-value=  6.190874e-76
```

The regular Poisson model with all predictors rejects the null hypothesis of the goodness of fit test indicating that it does not explain the data sufficiently.

```
model3 <- glm(count~persons+child+camper, offset=log(LOS), family=poisson, data=fish)
summary(model3)
```

```
##
## Call:
## glm(formula = count ~ persons + child + camper, family = poisson,
##     data = fish, offset = log(LOS))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2727  -1.3924  -0.9814  -0.0017  10.2987
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.03873    0.14014 -14.548  < 2e-16 ***
## persons      0.63079    0.03791  16.637  < 2e-16 ***
## child       -1.10014    0.07839 -14.035  < 2e-16 ***
## camper       0.27822    0.09159   3.038  0.00238 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1432.11  on 249  degrees of freedom
## Residual deviance:  788.45  on 246  degrees of freedom
## AIC: 1133.5
##
```

```
## Number of Fisher Scoring iterations: 6
```

```
cat("Goodness of Fit Test p-value= ", pchisq(model3$deviance, model3$df.residual, lower.tail=F))
```

```
## Goodness of Fit Test p-value=  4.335495e-58
```

Using length of stay as an offset improves the model. The AIC has decreased, but the goodness of fit test is still rejecting. This is an indication that we should use a different type of model.

```
model3.5 <- glm(count~persons+child+camper+persons:child, offset=log(LOS), family=poisson, data=fish)
summary(model3.5)
```

```
##
## Call:
## glm(formula = count ~ persons + child + camper + persons:child,
##     family = poisson, data = fish, offset = log(LOS))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2890  -1.3819  -0.9912   0.0134  10.2288
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.05607    0.14586 -14.096  < 2e-16 ***
## persons        0.63657    0.04017  15.847  < 2e-16 ***
## child         -0.92880    0.38728  -2.398  0.01647 *
## camper         0.27468    0.09192   2.988  0.00281 **
## persons:child -0.04881    0.10841  -0.450  0.65256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1432.11  on 249  degrees of freedom
## Residual deviance:  788.25  on 245  degrees of freedom
## AIC: 1135.3
##
## Number of Fisher Scoring iterations: 6
```

Adding the interaction between persons and child is not significant according to the Wald test. The AIC does not practically change.

```
model4 <- zeroinfl(count~persons+child+camper|persons+child+camper, offset=log(LOS), data=fish)
summary(model4)
```

```
##
## Call:
## zeroinfl(formula = count ~ persons + child + camper | persons + child +
##     camper, data = fish, offset = log(LOS))
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.073183 -0.740560 -0.358079 -0.001489 10.738599
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.12953    0.16110  -7.011 2.36e-12 ***
```

```
## persons      0.45097    0.04378  10.302  < 2e-16 ***
## child       -0.56379    0.09158  -6.156 7.44e-10 ***
## camper       0.01754    0.09563   0.183    0.855
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8052     0.5224   3.456 0.000549 ***
## persons      -0.9764     0.2128  -4.588 4.48e-06 ***
## child         2.0813     0.3465   6.007 1.90e-09 ***
## camper       -1.1632     0.3735  -3.114 0.001846 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 13
## Log-likelihood: -501.1 on 8 Df
```

```
AIC(model4)
```

```
## [1] 1018.137
```

Model4 is a ZIP model that uses the same predictors to predict the likelihood true zeros (if a group are fishers or not). Camper is no longer a significant predictor in the Poisson part of the model so it may be useful to remove and it see the effect.

```
model5 <- zeroinfl(count~persons+child|persons+child+camper, offset=log(LOS), data=fish)
summary(model5)
```

```
##
## Call:
## zeroinfl(formula = count ~ persons + child | persons + child + camper,
##     data = fish, offset = log(LOS))
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.07146 -0.74051 -0.36662 -0.00788 10.62119
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.12105    0.15428  -7.266 3.70e-13 ***
## persons      0.45264    0.04285  10.563  < 2e-16 ***
## child       -0.56372    0.09156  -6.157 7.44e-10 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8085     0.5219   3.465  0.00053 ***
## persons      -0.9759     0.2129  -4.584 4.56e-06 ***
## child         2.0829     0.3466   6.009 1.87e-09 ***
## camper       -1.1738     0.3694  -3.178  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 13
## Log-likelihood: -501.1 on 7 Df
```

```
AIC(model5)
```

```
## [1] 1016.171
```

Removing the insignificant camper variable in the Poisson part of the ZIP model improved (decreased) the AIC of the overall model.

```
vuong(model5, model3)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A  p-value
## Raw                    2.141718 model1 > model2 0.016108
## AIC-corrected          2.037533 model1 > model2 0.020798
## BIC-corrected          1.854091 model1 > model2 0.031863
```

The Vuong test indicates that the zero-inflated Poisson regression model explains the data better than the regular Poisson regression model in all 3 tests ($p<0.05$). The AIC of model5 (1016.171) is also much lower than the AIC of model3 (1133.513).

We can also discuss the Residuals vs Fitted plot to see the effect of the outlier we had seen in the EDA.

```
res.df <- data.frame(resid = residuals(model5),
                     fit = fitted(model5))
# Residuals vs Fitted plot
ggplot(res.df, aes(x=fit, y=resid))+
  geom_point()+
  labs(y="Residuals from ZIP Model", x="Fitted Values from ZIP Model", title="Figure 11")
```

Figure 11

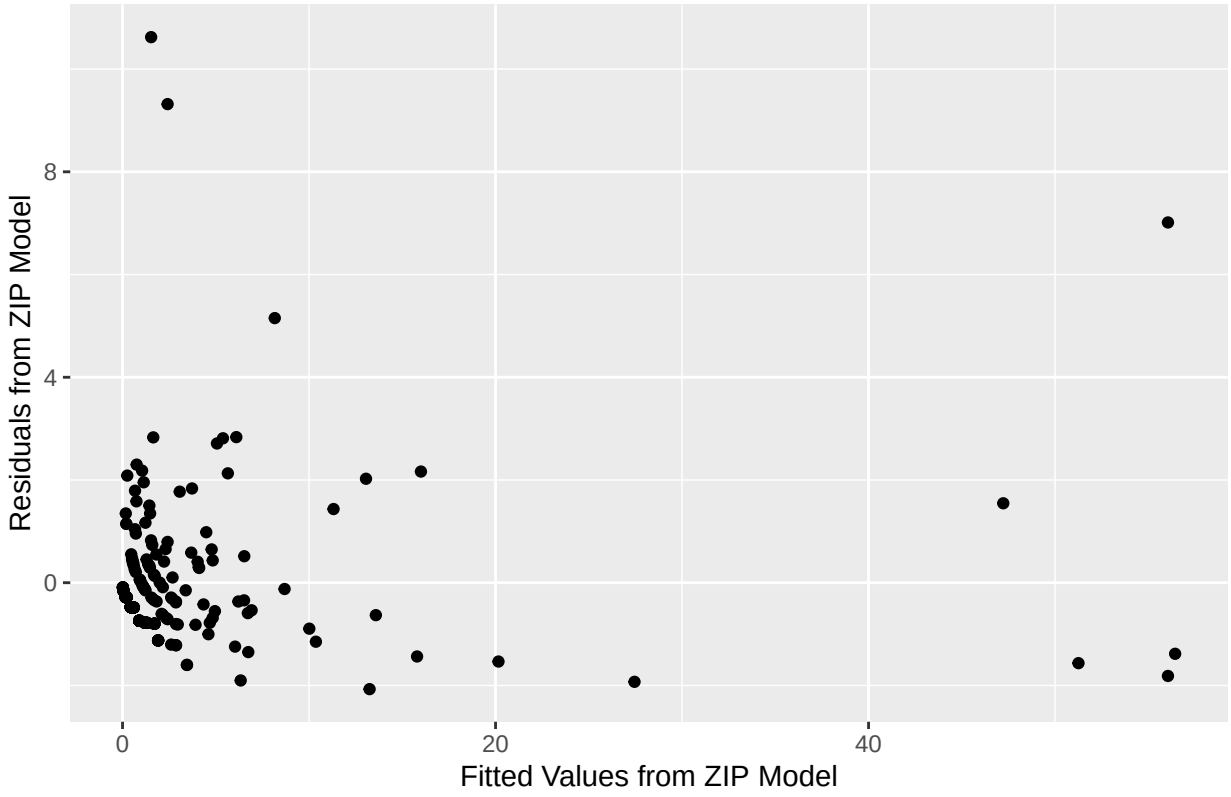


Figure 11 may be cause for some concern. There should be randomness and no pattern. There appears to be a cluster of points in the plot. There also appears to be a few points that may be relatively influential points

with high residuals. A solution for this will be discussed in the concerns of the conclusion.

Now we should determine if overdispersion is an issue.

```
model6 <- glm(count~persons+child+camper, offset=log(LOS), family=quasipoisson, data=fish)
summary(model6)
```

```
##
## Call:
## glm(formula = count ~ persons + child + camper, family = quasipoisson,
##     data = fish, offset = log(LOS))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2727  -1.3924  -0.9814  -0.0017  10.2987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.0387     0.3009  -6.776 9.05e-11 ***
## persons       0.6308     0.0814   7.749 2.42e-13 ***
## child        -1.1001     0.1683  -6.537 3.58e-10 ***
## camper        0.2782     0.1966   1.415    0.158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 4.609045)
##
##     Null deviance: 1432.11  on 249  degrees of freedom
## Residual deviance:  788.45  on 246  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Using the quasipoisson regression model, we see that the camper variable is no longer significant with the Wald test. This is the same result we saw using the ZIP model.

```
model7 <- glm(count~persons+child, offset=log(LOS), family=quasipoisson, data=fish)
summary(model7)
```

```
##
## Call:
## glm(formula = count ~ persons + child, family = quasipoisson,
##     data = fish, offset = log(LOS))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.014  -1.315  -1.034  -0.044   9.795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.89990    0.27587  -6.887 4.70e-11 ***
## persons      0.65653    0.07819   8.397 3.62e-15 ***
## child       -1.13910    0.16299  -6.989 2.57e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for quasipoisson family taken to be 4.413445)
##
##     Null deviance: 1432.11  on 249  degrees of freedom
## Residual deviance:  798.14  on 247  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
anova(model7, model6, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: count ~ persons + child
## Model 2: count ~ persons + child + camper
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1       247     798.14
## 2       246     788.45  1   9.6902 2.1024 0.1483
```

Removing the camper variable improved the model and the drop-in-deviance test (using the F distribution because of the quasi model) also concludes that the camper variable is insignificant. The dispersion parameter used is estimated to be 4.41 which is greater than 1. This will increase our standard errors for the predictor coefficient tests, but all the variables that we had previously determined were significant are still significant. This means the overdispersion was only affecting the camper variable which was already removed in our best ZIP model.

Conclusion

The model with the lowest AIC and best fit for the data is the zero-inflated Poisson regression model5.

```
exp(coef(model5))
```

```
## count_(Intercept)     count_persons       count_child  zero_(Intercept)
##         0.3259387         1.5724557         0.5690878         6.1011700
##      zero_persons        zero_child       zero_camper
##         0.3768669         8.0281155         0.3091849
```

Interpreting Coefficients in Context

Among fishers, the mean number of fish caught increases by 57.2% for every additional person holding other variables constant. Among fishers, the mean number of fish caught decreases by 43.1% for every additional child holding other variables constant. The intercepts in this model do not make much practical sense to interpret in this context as the variables will never allow the intercept to be used (persons > 0).

The odds that the group are non-fishers decreases by a factor of 0.377 for every additional person holding other variables constant. The odds that the group are non-fishers increases by a factor of 8.028 for every additional child holding other variables constant. The odds that the group that brings a camper are non-fishers is 0.309 times the odds of a group who did not bring a camper are non-fishers.

This model is the best fit of the data using our methods, although there still may be some cause for concerns. As described above, there is an issue with the Residuals vs Fitted plot. We discussed the effect of overdispersion and concluded that the variables selected in the final model are still significant. The outlier found in the EDA should be explored further before simply being removed from the model. Was this a true data point or was there an error in data entry in any way? Will this influence the research question present? More information would be necessary. It may also be interesting to discuss confounding variables or other reasons that could affect the number of fish caught (weather, if they brought a boat, fishing laws, etc.).