Negative Binomial Fishing Data Analysis

Minjee Kim

2025-02-02

1. Introduction

The dataset consists of visitor records at a state park, where visitors reported the number of fish they caught. However, we do not have direct information on whether each visitor fished, introducing potential structural zeros in the data.

The dataset includes the following variables:

livebait: A binary indicator for whether the visitor used live bait while fishing. camper: A binary indicator for whether the visitor camped at the park. persons: The total number of people in the visitor's group (count variable from 1 to 4). child: The number of children in the visitor's group (count variable from 0 to 3). This variable seems to be nested in "persons" variable. count: The number of fish caught by the visitor (highly skewed towards 0, with a long tail distribution up to 149 fish).

```
fishing_data <- read.csv("fishing.csv")
```

Explanatory Variables

The explanatory variables we will use to predict the number of fish caught are the following:

```
##              Variable Category Count Percentage (%)
## 1      Live Bait Used        0    34           13.6
## 2      Live Bait Used        1   216           86.4
## 3    Camped Overnight        0   103           41.2
## 4    Camped Overnight        1   147           58.8
## 5          Group Size        1    57           22.8
## 6          Group Size        2    70           28.0
## 7          Group Size        3    57           22.8
## 8          Group Size        4    66           26.4
## 9  Number of Children        0   132           52.8
## 10 Number of Children        1    75           30.0
## 11 Number of Children        2    33           13.2
## 12 Number of Children        3    10            4.0
```

```
## 'data.frame':    250 obs. of  5 variables:
##  $ livebait: Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 1 2 ...
##  $ camper  : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 1 2 2 ...
##  $ persons : int  1 1 1 2 1 4 3 4 3 1 ...
##  $ child   : int  0 0 0 1 0 2 1 3 2 0 ...
##  $ count   : int  0 0 0 0 1 0 0 0 0 1 ...
```

Response Variable

The variable we are interested in predicting is the count of fish caught by each group of visitors.

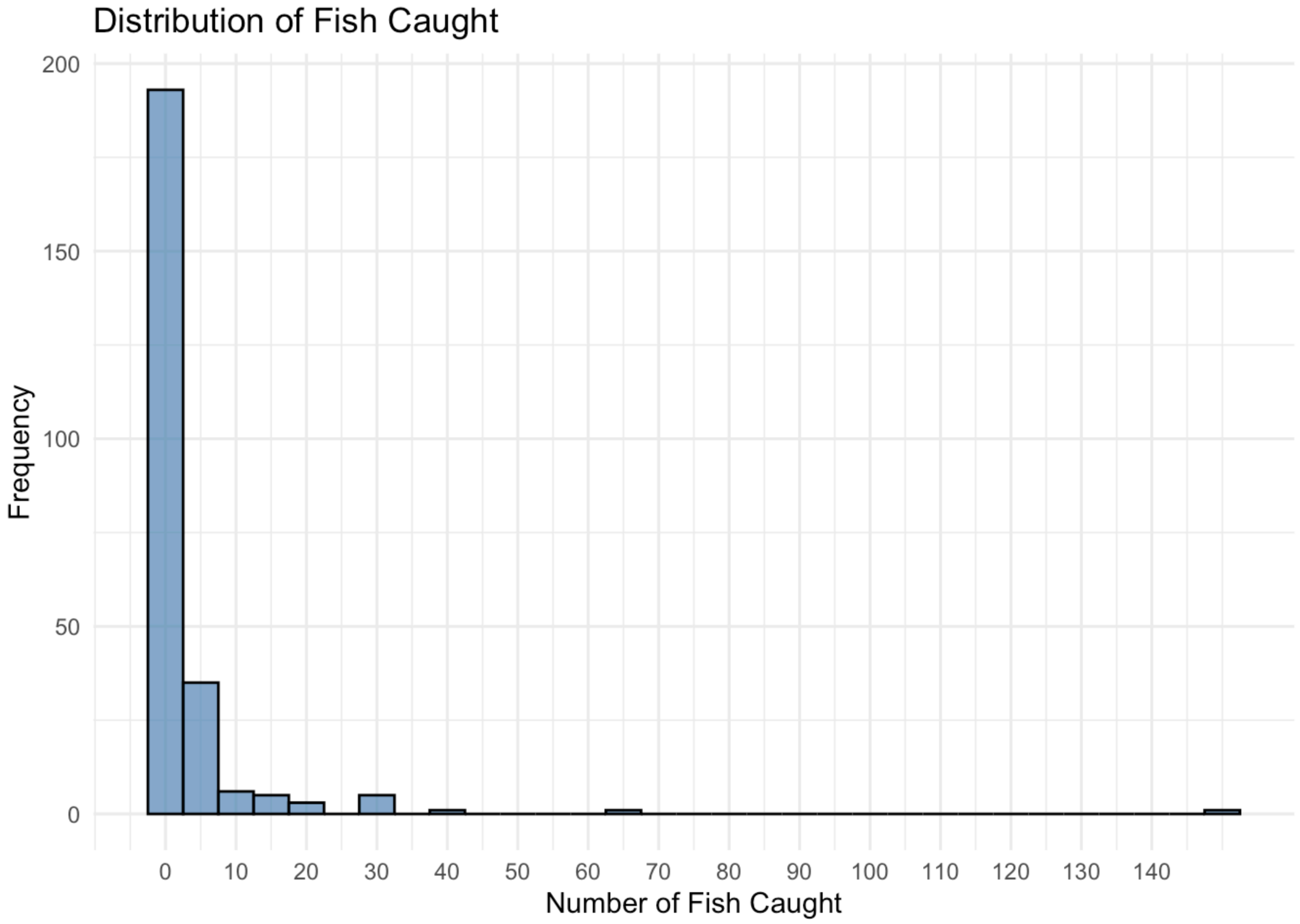
```
mean(fishing_data$count)
```

```
## [1] 3.296
```

```
var(fishing_data$count)
```

```
## [1] 135.3739
```

The mean of the response variable is 3.3 and the variance is much higher at 135.4. This suggests a large over-dispersion.



With a graph, it is much easier to see the majority of samples caught zero fish. This does not necessarily imply that all visitors did not go fishing - since the true rate of visitors who went into the state park to fish is unknown.

2. Preliminary Exploratory Data Analysis

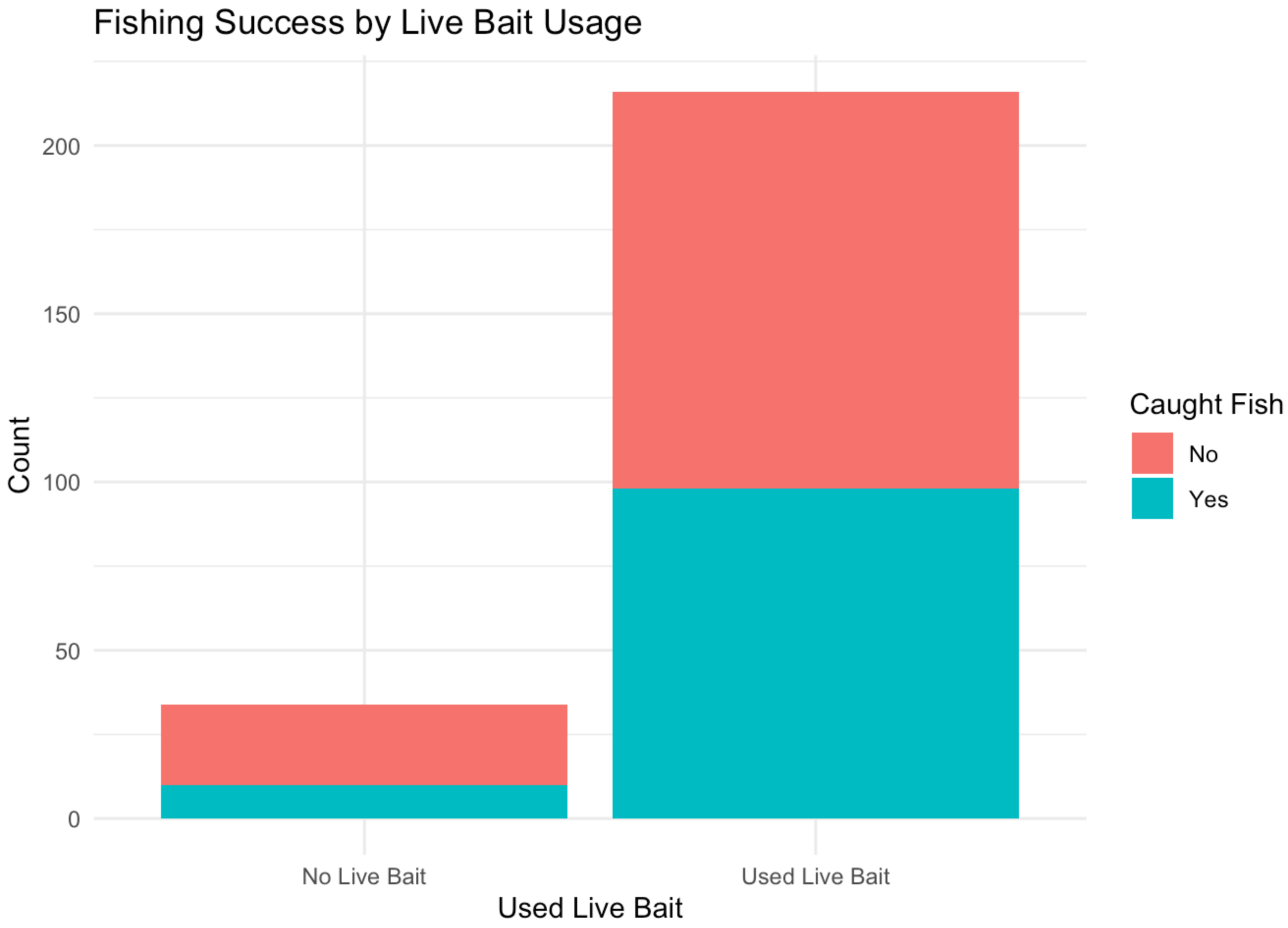
Let's examine some possible challenges in the data we should be aware of before building the model.

2.1 Zero Overinflation

We do not have information of whether each visitor group went to the park to fish or not. This implies that the large number of “no fish caught” could either be from fishermen not having any luck or from visitors that simply did not visit the park to go fishing. With this analysis, we should account for this latent factor into our model.

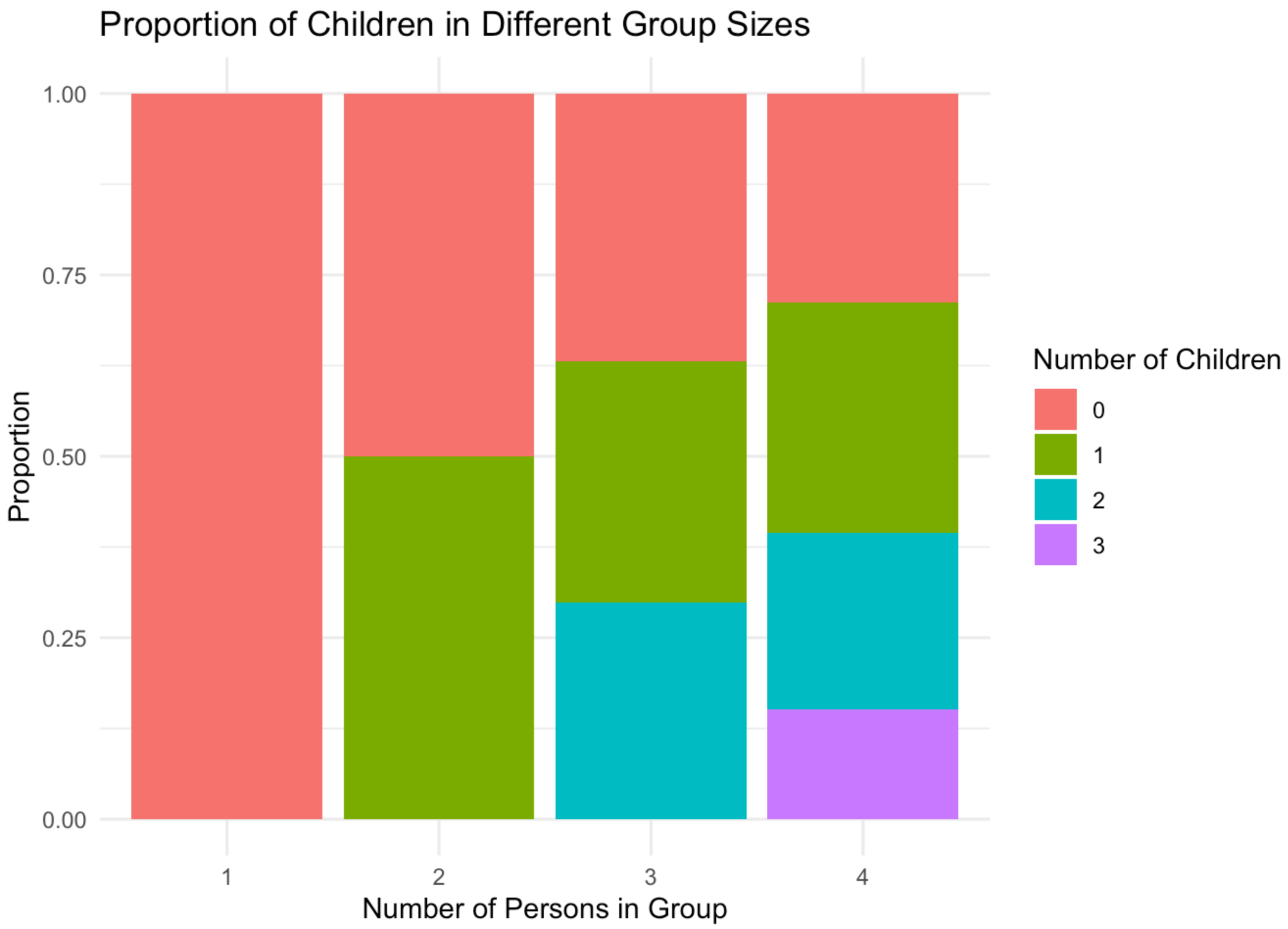
2.2 Constraint with Livebait use

On the other hand, if a group of visitors indicated their use of livebait, it is a definitive evidence that they did indeed go fishing. So, even if the group did not catch any fish, we can borrow information from the “livebait use” to decide whether the group went fishing or not. Possibly, this suggests a high correlation between the livebait use variable and the response (number of fish caught).



2.3 Group Size and Number of Children

We should keep in mind that the size of groups (“persons” variable) accounts for both the number of adults and children. This implies that the number of children (“child” variable) is nested in the persons variable. Not only that, there are no children visitors without adults present, meaning a very high correlation between the two, even if the “child” is removed from “persons”, the effects are still present.



3. Model Selection

Things to consider in the model: - over-inflated count of zero: we do not have any information on whether each observation of groups went to the park to fish - “child” count is nested in the “persons” count - constraint on “livebait”: if livebait is used it is reasonable to assume that the group was indeed fishing

3.1 Negative Binomial

We are interested in predicting how many fish will be caught by fishermen at a state park. Since the response of interest is count variable and the data has a large dispersion, we could consider the **Poisson**, **Negative Binomial**, and **Zero-Inflated Negative Binomial**. The dispersion is very high, so it is not reasonable to work with Poisson.

The Negative Binomial model extends the Poisson model by allowing for over-dispersion:

$$Var(Y) = E(Y) + \theta E(Y)^2$$

where θ is an overdispersion parameter.

Advantages:

- Handles count data with variance greater than the mean
- More flexible than Poisson

Limitation:

- Does not explicitly separate people who fished from those who did not.

$$\log(E(\text{count})) = \beta_0 + \beta_1 \times \text{livebait} + \beta_2 \times \text{camper} + \beta_3 \times \text{persons} + \beta_4 \times \text{child}$$

,

where

$$\text{Var}(\text{count}) = E(\text{count}) + \theta E(\text{count})^2, \quad \theta$$

```
## Likelihood ratio test
##
## Model 1: count ~ livebait + camper + persons * child
## Model 2: count ~ livebait + camper + persons + child
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -397.18
## 2   6 -398.58 -1 2.8069    0.09386 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A challenge we were concerned with was that persons and child variable might have a significant interaction. Based on the likelihood ratio test, the interaction is not significant on the prediction of counts.

Negative Binomial Regression Results

	Estimate	Std..Error	p.value
(Intercept)	-2.976	0.485	0.00000
livebaitUsed Live Bait	1.538	0.403	0.00014
camper1	0.518	0.231	0.02475
persons	1.062	0.111	0.00000
child	-1.801	0.184	0.00000

Based on the Negative Binomial Model, we can conclude:

1. Live bait use increases the predicted count of caught fish by exp(1.53) = 4.62 times.
2. Whether they camped or not has a positive affect on the number of fish caught.
3. Bigger groups tend to catch more fish.
4. The more children a group is with, the less fish they tend to catch.
5. The overdispersion parameter is significant - Poisson would not have been a good fit.

3.2 Zero-Inflated Negative Binomial

We have an **inflated** count of zeros due to two sources of “catching no fish”. Zero-inflated negative binomial is well-suited to handle this problem, because it assumes two processes:

1. A binary process that determines whether a group fished
2. A count process that models the number of fish caught given that they did go fishing.

$$P(Y_i = 0) = \pi_i + (1 - \pi_i) P_{NB}(0|X_i)$$

$$P(Y_i = y) = (1 - \pi_i) P_{NB}(y|X_i), \quad y > 0$$

- π_i is the probability of a **structural zero** (i.e., they did not fish).
- $P_{NB}(y|X_i)$ is the **negative binomial distribution** (accounts for overdispersion).
- X_i includes predictors like **livebait**, **camper**, **persons**, and **child**.

In other words,

$$E(\text{number of fish caught} = k) = P(\text{did not go fishing}) * 0 + P(\text{went fishing}) * E(y = k|\text{gone fishing})$$

For model selection, stepwise algorithm was used via AIC to select the best ZINB model.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in value[[3L]](cond): system is computationally singular: reciprocal
## condition number = 2.40545e-18FALSE
```

```
##
## Call:
## zeroinfl(formula = count ~ livebait + persons + child, data = fishing_data,
##     dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -0.76680 -0.57065 -0.29901  0.03957  9.70805
##
## Count model coefficients (negbin with log link):
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -2.7267     0.5120  -5.326 1.00e-07 ***
## livebaitUsed Live Bait   1.6857     0.4297   3.923 8.76e-05 ***
## persons                  1.0325     0.1204   8.574  < 2e-16 ***
## child                   -1.1829     0.2669  -4.432 9.35e-06 ***
## Log(theta)              -0.4103     0.2143  -1.914   0.0556 .
##
## Zero-inflation model coefficients (binomial with logit link):
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.13302    2.70513  -0.789 0.430399
## livebaitUsed Live Bait  0.01575    1.93915   0.008 0.993521
## persons                -0.52511    0.53486  -0.982 0.326206
## child                   2.75585    0.80208   3.436 0.000591 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.6635
## Number of iterations in BFGS optimization: 31
## Log-likelihood: -394.8 on 9 Df
```

The stepwise algorithm chose a zero-inflation model wtih live bait, persons, and child, but only child variable is signficant. Furthermore, comopared with a ZINB model with only child in the zero model part, the AIC and BIC of the simplified model are lower. Take this zinb_model to be our model of choice.

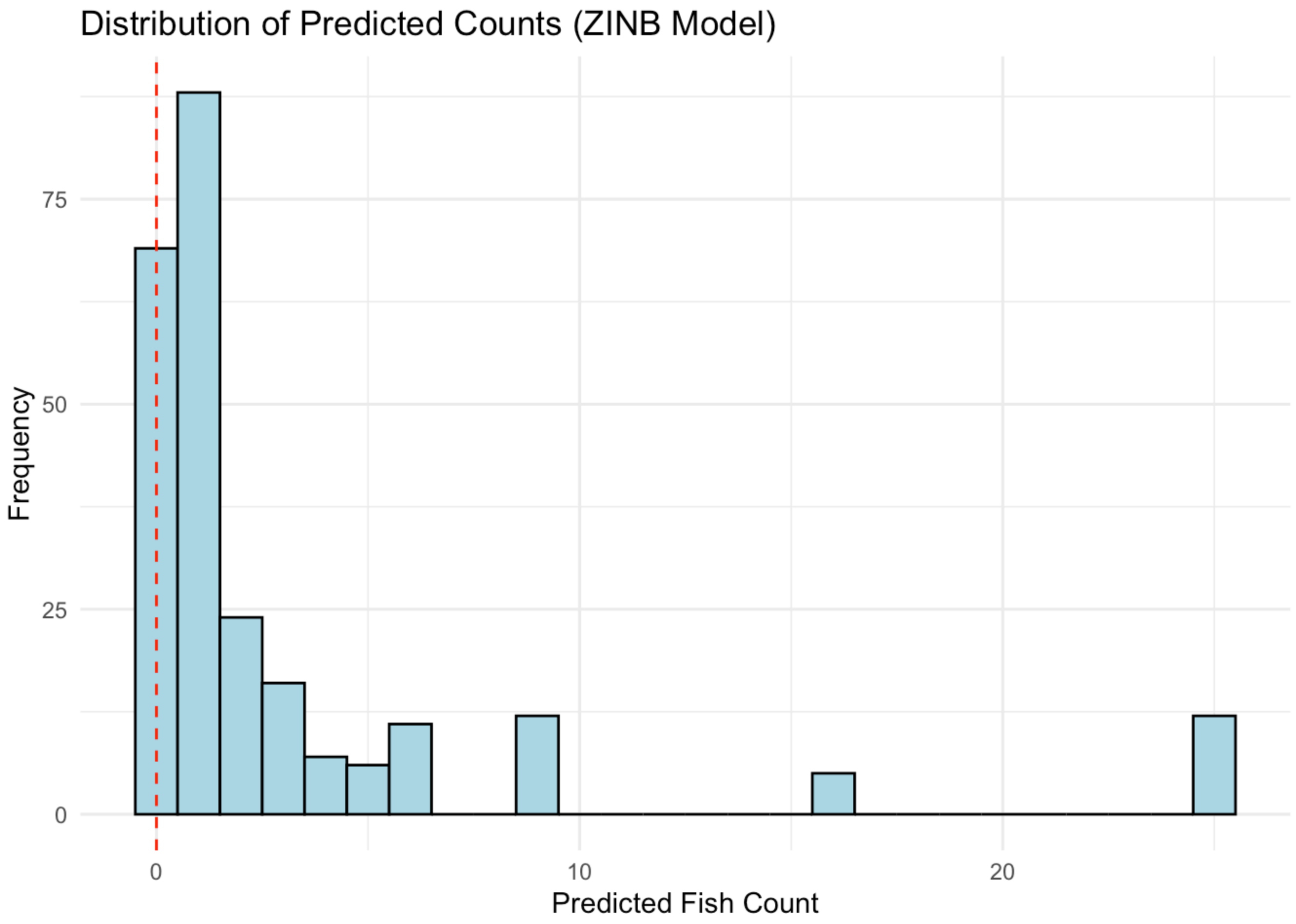
```
zinb_model <- zeroinfl(count ~ livebait + persons + child + camper |
                            child,
                            data = fishing_data,
                            dist = "negbin")
AIC(zinb_model, zinb_model_final)
```

```
##                  df      AIC
## zinb_model        8 802.5904
## zinb_model_final  9 807.6000
```

```
BIC(zinb_model, zinb_model_final)
```

```
##                  df      BIC
## zinb_model        8 830.7621
## zinb_model_final  9 839.2932
```

Based on both AIC and BIC, we can proceed by using the ZINB_model.



This plot shows the distribution of predicted fish count based on the ZINB model.

3.3 Interpretation

ZINB Count Model Coefficients

	Estimate	Std..Error	p.value
(Intercept)	-2.975	0.467	0.00000
livebaitUsed Live Bait	1.523	0.397	0.00013
persons	1.052	0.108	0.00000
child	-1.241	0.263	0.00000
camper1	0.463	0.234	0.04807
Log(theta)	-0.454	0.189	0.01618

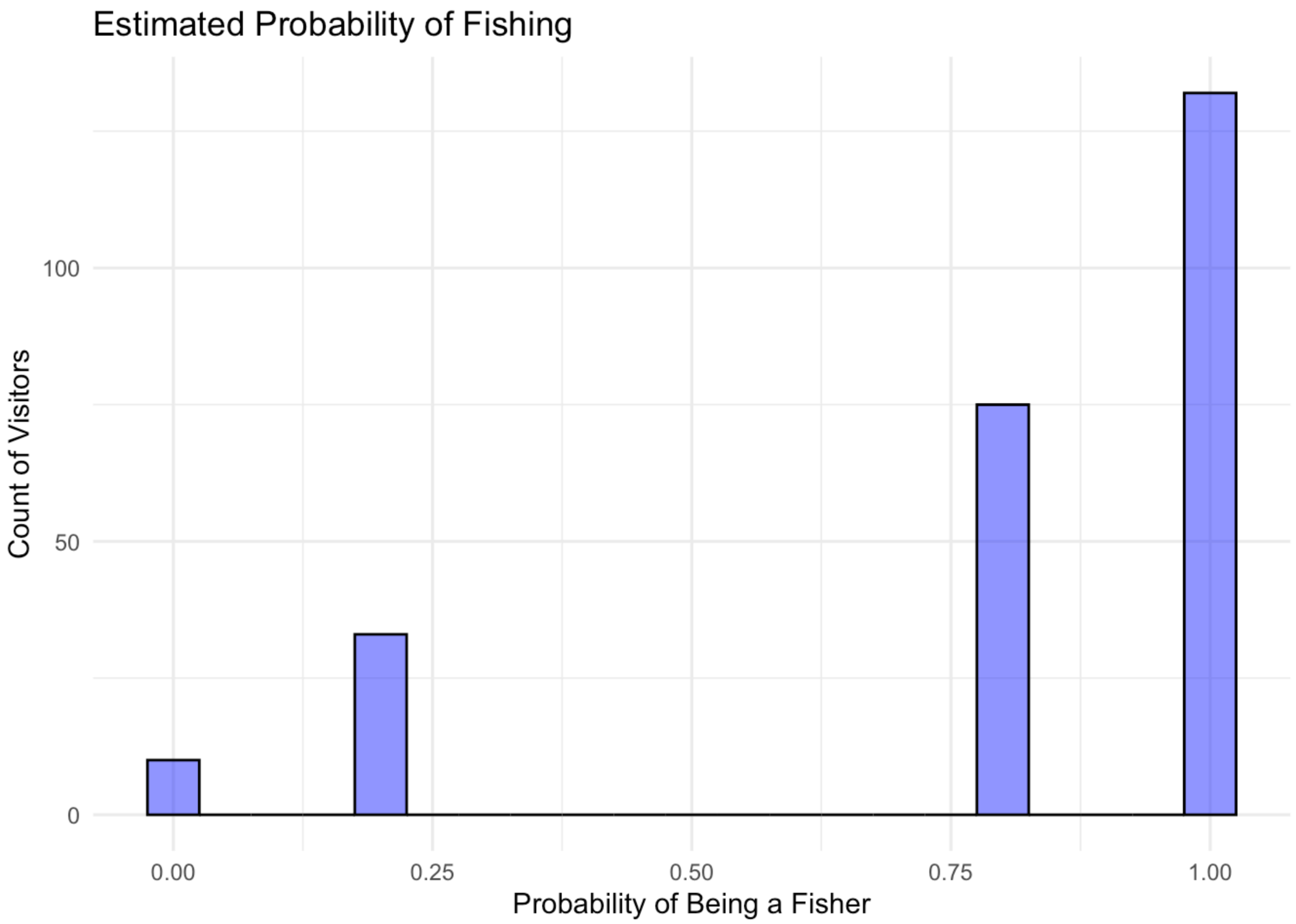
ZINB Zero-Inflation Model Coefficients

	Estimate	Std..Error	p.value
(Intercept)	-4.259	1.398	0.00231
child	2.839	0.808	0.00044

Based on the zero inflated negative binomial model, we can draw some similar interpretation as the negative binomial model -

1. Use of livebait, larger size of the group, and camping at the park all had a positive impact on the number of fish caught.
2. Larger number of children brought by the group had a negative affect on the number of fish caught.
3. Theta, the overdispersion parameter was still significant, indicating an overdispersion in the count regression part of the model.
4. The model selection chose only the “child” variable to be effective in the logistic regression to select out the people that did not go to the park for fishing (structural zeroes in the response).

```
## [1] 0.7884872
```



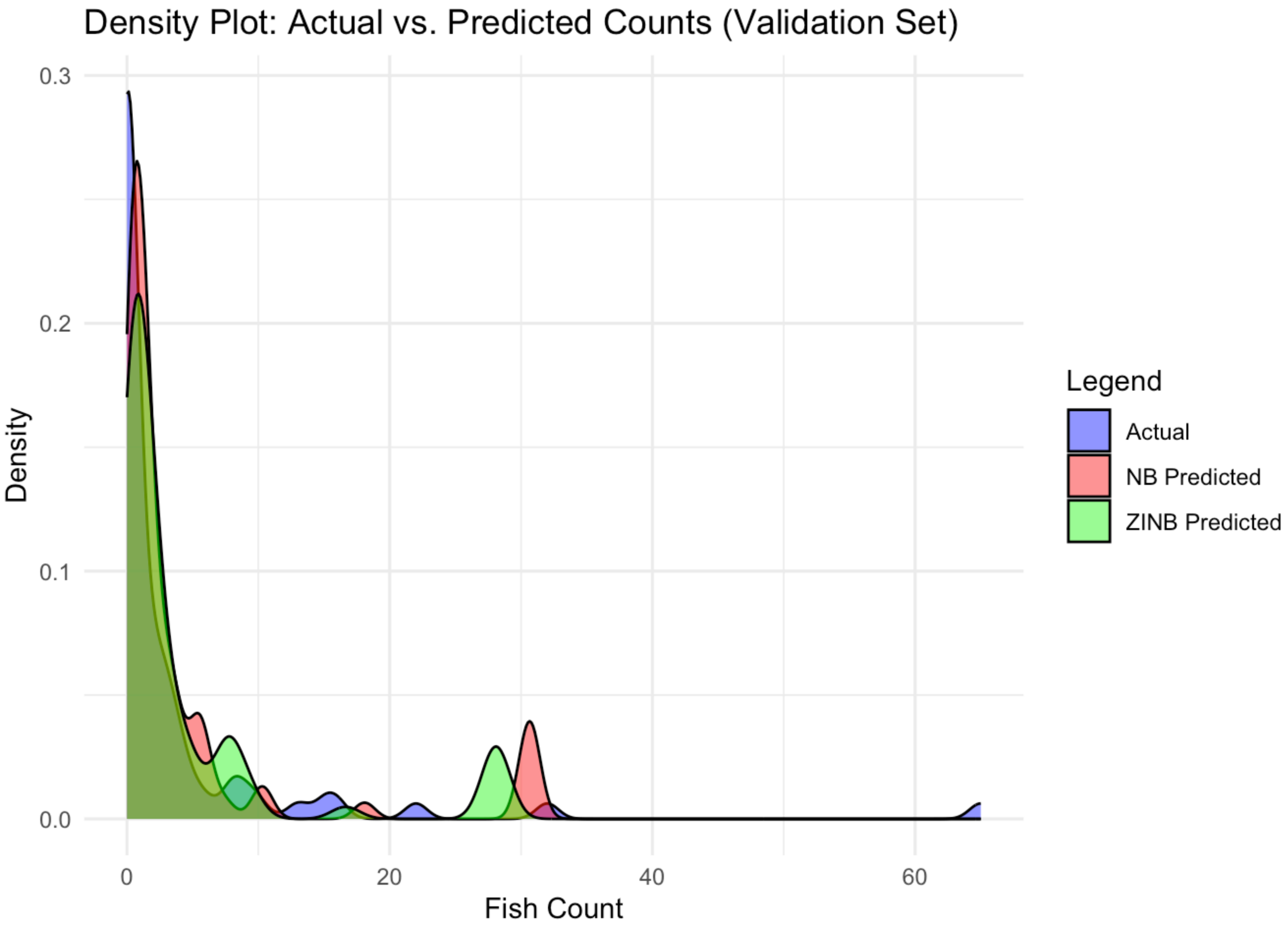
The ZINB model provides some estimates of probability of whether a group went to the park to go fishing or not. Based on the prediction results, the model believes that around 76% of the groups did indeed visit the state park for fishing purposes. In fact, the percentage of groups that the model was certain that they did not visit for fishing was very small (17.2%).

3.4 ZINB vs. NB Performance Comparison

```
##   Model     RMSE      MAE
## 1    NB 6.605428 3.104319
## 2  ZINB 6.467877 3.112134
```

Based on the 70-30 training split prediction, it is clear that the negative binomial model performed slightly better and at least comparable to the zero inflated negative binomial model. This makes sense based on our understanding that the ZINB model identified “not fishing” groups to be quite few. The additional binary process of the ZINB model that sorts out the “not fishing” groups was essentially adding more complexity without contributing to the results, since the binary process itself was redundant.

Very disappointing results.



4. Conclusion

Now, we can draw some conclusions about the data based on our models.

4.1 Model Choice

First, ZINB model is redundant. While the Negative Binomial is well justified by the signficant overdispersion variable, Zero inflated model is unnecessary for this data, since the majority of the zeroes in the response seem to be reasonable - fishermen really did not catch any fish.

The state park must be KNOWN for being a fishing spot!

4.2 Remarks

Revisiting the “livebait” variable - earlier, we claimed that the livebait variable serves as a constraint on knowing the true state of whether the group did indeed go fishing or not. Let’s revisit the livebait variable -

Table of Livebait vs. Caught Fish

	No	Yes
No Live Bait	24	10
Used Live Bait	118	98

There are only a small portion of the groups (34/ 250) who did not livebaits. This explains our ZINB predictions that most groups did indeed visit the park for fishing, since most groups actually used livebaits.

Now, another conclusion we drew, was that the number of children had a consisitently, significantly negative impact on the number of fish caught by each group. Let's look into the relationship between the number of children and fish caught.

Table of Child Count vs. Caught Fish

	No	Yes
0	56	76
1	46	29
2	30	3
3	10	0

All ten groups that brough three children did not catch any fish! This explains the unstable calculations (when you run the Negative binomial model, treating the number of children as categorical variables) on child = 3.

Among the ten groups that did bring three children with them, let's see whether any of the groups went "for fishing" based on the livebait use -

```
## 
##      No Live Bait Used Live Bait
##   No            2              8
```

Surprising! 8/10 groups that brought three children actually used livebaits. This implies that the eight groups came to the park with three children for the purpose of fishing, but absolutely none of them caught any fish!

Conclusion? Don't bring three children if you want to catch some fish!