**St. Xavier’s College, Kolkata**

*Semester VI Dissertation Paper*

**Name** ***Anirban Ray***

**Roll No.** ***3-14-14-0446***

**Registration No. *A01-1112-0854-14***

**Supervisor’s Name** ***Professor Debjit Sengupta***

**Title *Analysis of Fishing Data – Application of Count Regression***

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

**Introduction**

Based on a real-life dataset, finding its underlying characteristics and fitting a suitable regression model to perform prediction on the dependent variable based on some independent variables is one of the main interests in the theory of statistics. A standard (and unarguably, the most popular) method for this purpose is the *Method of Least Squares*, which works best if the data under study has a normal distribution. If the data is normal, we can directly use this method to fit a prediction formula. Otherwise, if the observed raw data are continuous (which is true in most of the practical situations), we can assume that it can be approximated by a normal distribution more or less satisfactorily by virtue of the *Central Limit Theorem*, in case the sample size is large. But if it is not so, we transform the data in such a way that it becomes normal (using *Box Cox Transformation*, *Johnson Transformation*, etc.), and then perform the ordinary least squares method. But if the data is of discrete type, i.e. if the data is based on counts, this method fails. In that case, we must look for some other approach of fitting a model.

**Generalized Linear Models**

One of the methods for setting up a regression model for the count data is the method of generalized linear models. This refers to a broad class of conventional linear regression models for a response variable given some continuous and/or categorical explanatory variables. These models generally have three components:

1. *Random Component* – This refers to the probability distribution of the response variable, say . This is analogous to the error term in classical linear regression model.
2. *Systematic Component* – This specifies the explanatory variables, say , of the model in form of a linear predictor, e.g. , where are the unknown parameters of the model.
3. *Link Function* – This specifies the link between the random and the systematic components. It says how the expected value of the response relates to the linear predictor of the explanatory variables, for example .

This method can be applied as long as the different realizations of the response variable, say are independent observations from an exponential family of distribution. Then based on the model, the parameters are estimated by the method of maximum likelihood rather than ordinary least squares. Thus, the estimation process relies on the large sample approximations.

**Possible Choices of Models for Count Regression**

Now, there are a lot of possible models for a count data, for all choices of link function and the distribution. Generally, it is customary to fit a Poisson model to the data with a logarithmic link function. But this model has very strict assumption of equality of mean and variance, which may not always hold in real life. In those cases, the first alternative approach that comes to our mind is the negative binomial model, which can successfully explain higher variance in the dataset compared to the mean. Again, if the dataset is such that there are too many zeroes due to the fact that some are generated by a binary process (called structural zeroes) and the rest are generated by a count process (called true zeroes), we can fit a zero-inflated Poisson or negative binomial model to find out the estimate of the success probability of the binary process and fit a model to the count process simultaneously. Again, it may so happen that the zeroes of the dataset may occur due to a process completely independent of another process which generates the nonzero counts. In that case, we can fit hurdle models.

**Data Description**

The dataset for our study is collected from [*https://stats.idre.ucla.edu/stat/data/fish.csv*](https://stats.idre.ucla.edu/stat/data/fish.csv). Here, we have data on groups of visitors who went to a state park, where fishing is allowed. They were questioned how many people and children were there in the group, how many fishes were caught and whether they brought a camper. We have data on the following variables:

1. persons () - number of persons in that group
2. child () - number of children in the group
3. camper () - whether the group brought a camper to the park or not
4. count () - number of fishes caught by a group

We also define two indicator variables and respectively denoting whether a group brought camper or not.

**Descriptive Approaches**

First, we plot the column diagram of the data. We obtain the figure as follows.

We also compute the basic descriptive measures of central tendency, dispersion, skewness and kurtosis. The values come out as given below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measures** |  |  |  |  |
| **Values** | 3.296 | 11.635 | 8.931 | 100.405 |

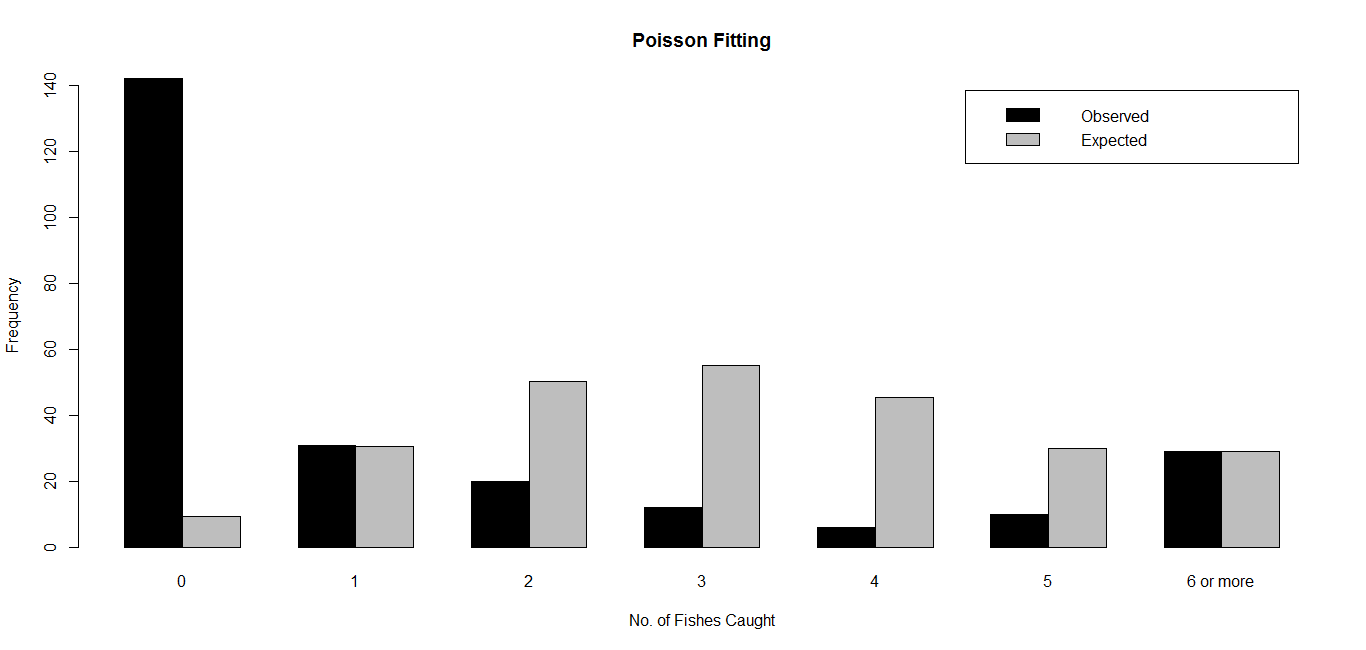
**Data Study**

From the column diagram and the descriptive measures, we observe that there are too many zeroes in our dataset compared to other values. Also, there is huge variation in the data, where its average is quite small. We also note that the data is highly positively skewed and leptokurtic. All of these may be due to the existence of too much zeroes (almost 60%). On the other hand, some groups have caught a very high number of fishes, giving rise to a small number of very large observations. This leads to huge variation in the data compared to its average, thereby deviating quite significantly from one of the basic assumptions of the Poisson model. So, we suspect that over-dispersion may be present in the dataset and if so, negative binomial models are expected to give us good results. Alternatively, we can also consider the fact that some visitors did not fish at all, and since we do not have data on whether a person fished or not, those observations are recorded as zeroes. Along with these zeroes, there are also other zeroes as some groups failed to catch any fish, just due to bad luck. This indicates that fitting a zero-inflated model to this data may also be a good idea, in case the standard models fail to explain the variation justifiably.

**Objective**

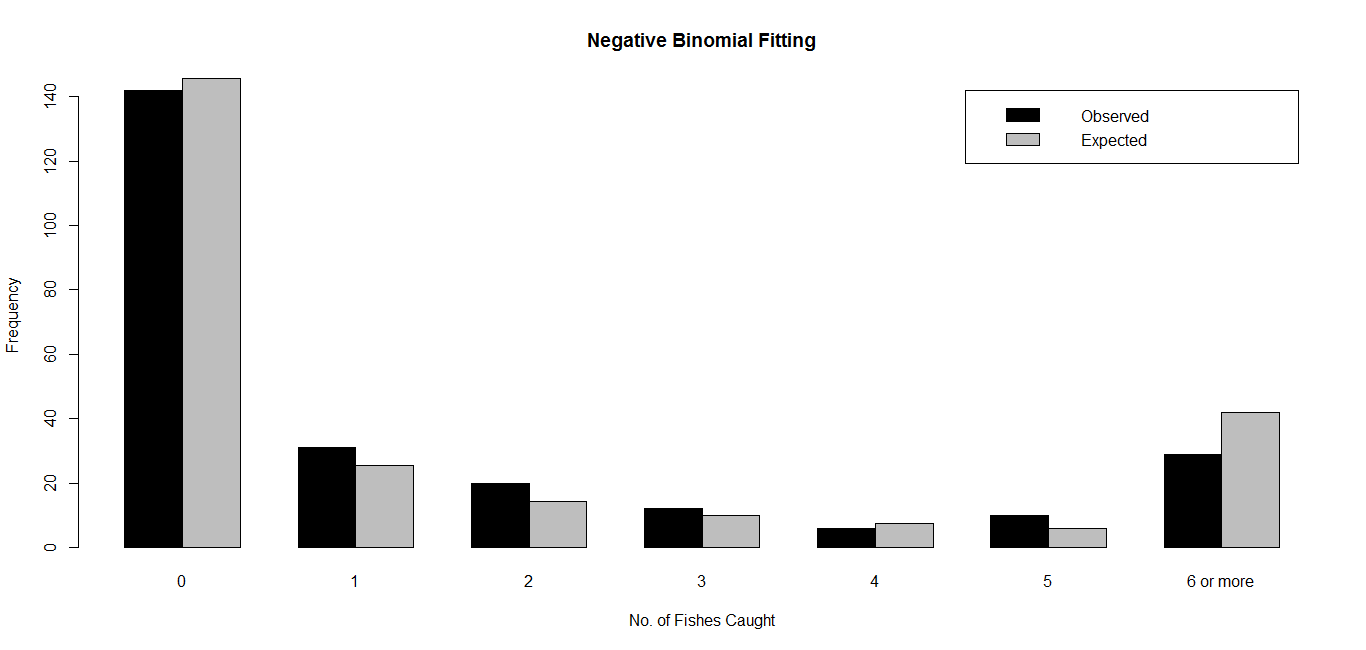
Our objective is to predict the number of fishes caught by a group. If a group of persons and children comes to the park with or without a camper, we wish to guess beforehand whether they will fish or not and if they do, how many fish they will catch. Hence, first of all we would like to estimate the probability that a group of visitors will fish, if possible. We then try to fit a count regression model of on , and , as evidently the variable of interest is a discrete random variable, taking only the non-negative integer values.

**Large Variation - Failure of Poisson Model**

To verify our very first suspicion that the Poisson model will not be appropriate for this dataset, we proceed to fit a Poisson model and test how well it is by Goodness of Fit test. The maximum likelihood estimator of the parameter of a distribution is given by, . Thus, we fit a model and obtain the value of the Pearson’s Chi-Square statistic as 2001.086, a very large value. Thus, we conclude that the fit is very poor. We plot the observed and expected frequencies together and the diagram clearly shows a huge discrepancy between those two.

**Accounting for Over-dispersion – Negative Binomial Model**

Next, we try to fit a negative binomial model with parameters and , as theoretically its variance () is more than its mean (). The maximum likelihood estimator of is the sample mean and that of is obtained by solving the following likelihood equation by numerical methods,

Hence, we fit a model to our fishing data, and the value of Pearson Chi-Square statistic, being just 11.318, leads us to the conclusion that this fit is far better than the previous one and this can be a satisfactory model as a regression model. The following figure also convey us the same indication.

**Negative Binomial Regression**

We now proceed to fit the regression, as negative binomial seems to explain the data quite well. For fitting the model, we use the logarithmic link function, i.e. our model is,

, where is the disturbance term of the model. Here, we have excluded any intercept term, as a group of size zero cannot possibly catch any fish. We obtain the fitted model as,

**Significance of Predictor Variables in Negative Binomial Regression**

Now, we wish to test whether the explanatory variables, namely, the number of persons in a group, the number of children in a group and whether the group brought a camper really affect the number of fishes caught or not. For that, we test the hypotheses against , and an appropriate test statistic is given by , which can be assumed to follow a distribution under , . We reject the null hypothesis at level of significance if and only if , where is the observed value of and is the upper point of a standard normal distribution.

Let us take . Then the critical point becomes . Based on the data, we prepare the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Estimate** | **Z value** | **Decision** |
|  | 1.061 | 9.273 | Reject |
|  | -1.781 | -9.623 | Reject |
|  | -1.625 | -4.918 | Reject |
|  | -1.004 | -3.362 | Reject |

Thus, we can conclude that at 5% level of significance, effects of all the predictors on the response is significant.

**Probability of Extra Zeroes – Zero-inflated Poisson Model**

Again, it is also logical to take into account the fact that some visitors did not fish in the park, i.e. one could wish to explain the excess zeroes present in the data which are the result of non-fishing. Then we may opt for a zero-inflated Poisson model. Here, we assume that a particular group of visitor will not fish with probability , and if that group actually fish, the number of fishes caught will be governed by a Poisson process with parameter . We also assume that these two processes, viz. the Bernoulli process (whether fishes or not) and the Poisson process (how many fishes) are independent. This assumption leads us to a zero-inflated Poisson distribution with parameters and .

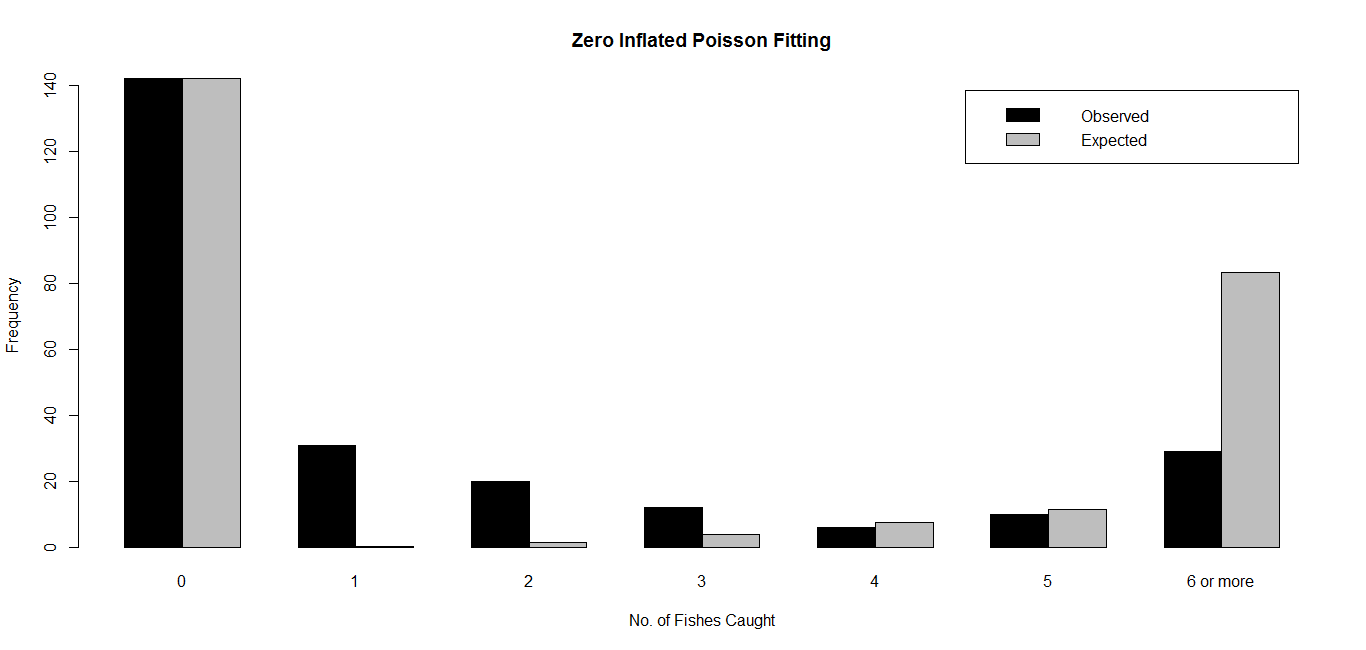
The p. m. f. of a distribution is given by,

, where and . In order to estimate the parameters, we can use either the maximum likelihood estimators, or the method of moment estimators. By the bootstrapping procedure of drawing repeated samples of size 1000 from a hypothetical population 500 times, we have verified that maximum likelihood estimators give us better estimates.

**Zero-Inflated Poisson Fitting**

In order to fit a zero-inflated Poisson distribution using the method of maximum likelihood, the maximum likelihood estimators of and can be found by numerically solving the following two maximum likelihood equations,

, where is the observed frequency of zero in the sample. We obtain the maximum likelihood estimates of and as 7.628 and 0.568 respectively. Hence, we can say that on an average, almost 57% of the visitors do not fish in the state park. Equivalently, one can say that a group of visitors will be fishing in the park with probability 0.43. Also, we can say that the groups, who fish at the park, on an average catch approximately 8fishes. But it should be noted that this model does not give us a good fit, which can be checked from the following diagram of observed and expected frequencies.

 **Zero-Inflated Poisson Regression**

Though the fit seems to be poor, if we fit a regression model with logarithmic link function (i.e. ) for the Poisson process and logit link function (i.e. ) for the Bernoulli process, the fitted model is obtained as,

**Significance of Predictor Variables in Zero-Inflated Poisson Regression**

Now, we wish to test the significance of the coefficients of the model based on the sample estimates, as we have done for the negative binomial regression. The results obtained are presented in the following table. For each of the columns, the 1st sub column represents the model of the response and 2nd one represents the model of the logit.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Estimate** | | **Z value** | | **Decision** | |
|  | 0.829 | -0.923 | 18.862 | -4.632 | Reject | Reject |
|  | -1.137 | 1.905 | -12.224 | 5.840 | Reject | Reject |
|  | -0.798 | 1.664 | -4.673 | 3.227 | Reject | Reject |
|  | -0.074 | 0.830 | -0.471 | 1.922 | Accept | Accept |

Thus, we can conclude that at 5% level of significance, effects of all the predictors other than the presence of camper, i.e. on the response is significant. Thus, under this setup, bringing a camper to the park does not significantly affect whether the group of visitors would fish, or if they do fish, it does not affect the number of fishes caught significantly.

**Comparison of the two models**

To compare the negative binomial and the zero-inflated Poisson model, we can use *Akaike Information Criterion*. It is defined as, , where is the number of free parameters in the fitted model that needs to be estimated. Between two possible models for a given dataset, the one with lower value of AIC is considered to be the better one. Here, for the negative binomial model, AIC comes out as , where the zero-inflated Poisson model has as its AIC value. So, we can conclude that zero-inflated Poisson model is not better than the negative binomial model, rather it is worse. So, we should fit a negative binomial model to the available data for any inferential purpose. We can now proceed to see what conclusions can we make from the fitted negative binomial model based on the sample data.

**Interpretation of Coefficients in Negative Binomial Regression**

Since our model is quite different from that of the method of ordinary least squares, the model coefficients need to be interpreted accordingly, considering the logarithmic link function. For and , is the increment in the logarithm of the response for unit increase in , provided that the other predictors remain unchanged, . Thus, signifies that if we increase in by unit, the expected number of fishes caught will be increased by . Thus, based on the fitted model, presence of one additional person in a group will increase the expected number of fishes caught by almost , and if one more child is there, that will decrease it by more or less . Again, will signify the proportional increase in successful fishing when a group brings a camper to the park compared to group with identical compositions of its members who have not brought any camper. Hence, we can say that bringing camper will increase the number of fishes caught by approximately . Combining, we can conclude that a large group of visitors with very few children (or none at all) with a camper are expected to catch more fishes compared to other groups.

**Bibliography**

1. *UCLA* - Institute for Digital Research and Education (accessed on 07.01.2017)
2. *STAT 504* – Analysis of Discrete Data (accessed on 21.01.2017)
3. *Wikipedia* – Zero-inflated Model (accessed on 15.01.2017)