Properties and Applications of the Zero-Inflated Generalized Poisson (ZIGP) Regression Model

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

Post Graduate Diploma in Statistical Methods and Analytics

Submitted by

Tanmay Gayen

ROLL NO:

REGISTRATION NO:

Preetam Biswas

ROLL NO:

REGISTRATION NO:

INDIAN STATISTICAL INSTITUTE(ISI),Chennai
(An Institution of National Importance)
Chennai Centre 110, Nelson Manickam Road Aminjikarai Chennai, 600 029.

DECLARATION

we, **Tanmay Gayen and Preetam Biswas**, hereby declare that this report entitled **"Properties and Applications of the Zero-Inflated Generalized Poisson (ZIGP) Regression Model"** submitted to INDIAN STATISTICAL INSTITUTE(ISI), chennai towards the partial requirement of **PGDSMA** course, is an original work carried out by me under the supervision of **Dr.Surajit Pal** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. We have sincerely tried to uphold academic ethics and honesty. Whenever a piece of external information or statement or result is used then, that has been duly acknowledged and cited.

Aminjikarai Chennai, 600 029 **Tanmay Gayen**
May, 2025 **Preetam Biswas**

ACKNOWLEDGEMENT

I thank everyone who helped me see this project through to completion. I would like to first express my profound gratitude and deep regard to Dr.Surajit Pal,Indian Statistical Institute, and sincerely wish to acknowledge his vision, guidance, valuable feedback, and constant support throughout the duration of this project. I am also thankful to all of my teachers and my friends for their steadfast encouragement and time.

I am lastly grateful to the Indian Statistical Institute for providing the necessary resources and facilities to complete this project to the best of my ability.

June,2024

Tanmay Gayen,Preetam Biswas
PGDSMA
Chennai Centre 110,
Nelson Manickam Road Aminjikarai Chennai, 600 029

ABSTRACT

Overdispersion is a common issue in Poisson modeling. The Generalized Poisson (GP) distribution offers flexibility by accommodating both overdispersed and underdispersed count data. In this study, we provide a brief overview of various overdispersed and zero-inflated regression models. To examine the consequences of fitting an incorrect model to data generated from a different distribution, we simulate data from a zero inflation generalized Poisson (ZIGP) distribution and fit a regression model to it. Here we used a maximum likelihood estimator (MLE) to estimate the ZIGP distribution parameter. Then find the Fisher information matrix and calculate the standard error of this parameter.

Contents

1 Introduction

Statistical models for count data have found applications in various fields, including insurance, dental epidemiology, healthcare facilities, risk classification, and medicine. The Poisson model is a commonly used method for analyzing such data. However, a major limitation of the Poisson model is the assumption that the mean and variance are equal, a property known as equidispersion. However, in practical situations, it is often observed that the variance of the sample data is greater than the mean. The phenomenon of excess variability is called overdispersion and has been widely studied in the literature Dean (1992).Failure to adequately account for existing overdispersion results in significant underestimation of standard errors and can lead to incorrect conclusions regarding regression parameters. As a result, various models and estimation techniques have been developed to manage the data overdispersed. Among these models, the two widely used forms of the negative binomial (NB) model, known as the regression models NB-1 and NB-2, are particularly notable are discussed by McCullagh and Nelder (1989) and Kamalja and Wagh (2018).
Another common distribution used to handle overdispersion or underdispersion is the *generalized Poisson distribution* (GPD). The classical GP regression model, often referred to as the *GP-1 model*, is a natural extension of the Poisson regression model. A restricted version of this model, known as the *GP-2 regression model*, was also introduced. Both GP-1 and GP-2 regression models extend the standard Poisson regression framework and are effective in addressing various forms of dispersion in count data. Thus, along with the negative binomial (NB) model, the generalized Poisson (GP) models are commonly employed to manage dispersion in the data.

2 The Zero-inflated Generalized Poisson Distribution

. Consul and Jain (1973)initially proposed a functional form of the generalized Poisson distribution (GP) to address both overdispersion and underdispersionn count data.Let the random variable Y represent the count data following a generalized Poisson (GP) distribution. The probability mass function is Y is defined a

$$p_Y(y) = P[Y = y] = \frac{\theta(\theta + \phi y)^{y-1}}{(1+\phi)^y \, y!} \exp\left(-\frac{\theta + \phi y}{1+\phi}\right), \quad y = 0, 1, 2, 3, \ldots$$

where the mean parameter is θ ($\theta > 0$) and the dispersion parameter is ϕ ($\phi > 0$). The mean and variance of the GP distribution are given by

$$E(Y) = \theta \quad \text{and} \quad V(Y) = \theta(1+\phi)^2.$$

The zero-inflated generalized Poisson (ZIGP) model combines a degenerate Bernoulli distribution at zero with a baseline generalized Poisson (GP) distribution. Various forms of the ZIGP distribution have been developed, and ZIGP regression models are widely used to analyze real-world zero-inflated count data.such as Shahsavari et al. (2023),Brooks et al. (2017),Zamani and Ismail (2014) etc.The probability mass function (PMF) of the ZIGP distribution is given by

$$P[Y = y] = \begin{cases} \omega + (1-\omega)\exp\left(-\frac{\theta}{1+\phi}\right), & \text{for } y = 0, \\ (1-\omega)\frac{\theta(\theta+\phi y)^{y-1}}{(1+\phi)^y \, y!}\exp\left(-\frac{\theta+\phi y}{1+\phi}\right), & \text{for } y > 0, \end{cases}$$

where ω $(0 \leq \omega < 1)$ is the zero-inflation parameter. The mean and variance of the ZIGP distribution are given by

$$E(Y) = \theta(1-\omega)$$

and

$$\mathrm{Var}(Y) = \theta(1-\omega)\left[(1+\phi)^2 + \omega\theta\right].$$

The parameters of the ZIGP distribution can be estimated using either the maximum likelihood method or the method of moments. Here we used the maximum likelihood method. A sufficiently large sample size (preferably $n \geq 300$) is recommended to ensure reliable estimates.

Let n_d denote the number of sample units with exactly d defects for $d = 0, 1, 2, \ldots, m$. Then, the log-likelihood function for the parameters ω, θ, and ϕ based on a sample of size n can be expressed as follows:

$$\begin{aligned}
\ln L(\omega,\theta,\phi) &= \ln\left[\{P(Y=0)\}^{n_0}\right] + \ln\left[\prod_{d=1}^{m}\{P(Y=d)\}^{n_d}\right] \\
&= n_0 \ln\left[\omega + (1-\omega)e^{-\theta/(1+\phi)}\right] + \sum_{d=1}^{m} n_d \ln\left\{(1-\omega)\frac{\theta(\theta+\phi d)^{d-1}}{(1+\phi)^d d!}\exp\left(\frac{-(\theta+\phi d)}{1+\phi}\right)\right\} \\
&= n_0 \ln\left[\omega + (1-\omega)e^{-\theta/(1+\phi)}\right] + (n-n_0)\left[\ln(1-\omega) + \ln\theta\right] + \sum_{d=1}^{m} n_d (d-1)\ln(\theta + d\phi) \\
&\quad - \sum_{d=1}^{m} d n_d \ln(1+\phi) - \sum_{d=1}^{m} n_d \ln(d!) - \sum_{d=1}^{m} n_d \frac{(\theta + d\phi)}{1+\phi}.
\end{aligned}$$

The partial derivative of the logarithmic likelihood function $\ln L(\omega,\theta,\phi)$ with respect to ω is derived as follows:

$$\frac{\partial}{\partial\omega}\ln L(\omega,\theta,\phi) = n_0 \frac{1 - e^{-\theta/(1+\phi)}}{\omega + (1-\omega)e^{-\theta/(1+\phi)}} - \frac{n-n_0}{1-\omega}.$$

The partial derivative of the logarithmic likelihood function $\ln L(\omega,\theta,\phi)$ with respect to θ is derived as follows:

$$\frac{\partial \ln L(\omega,\theta,\phi)}{\partial\theta} = -\frac{n_0(1-\omega)e^{-\theta/(1+\phi)}}{(1+\phi)\left[\omega + (1-\omega)e^{-\theta/(1+\phi)}\right]} + \frac{n-n_0}{\theta} + \sum_{d=1}^{m}\frac{n_d(d-1)}{\theta + d\phi} - \frac{n-n_0}{1+\phi}$$

The partial derivative of the logarithmic likelihood function $\ln L(\omega,\theta,\phi)$ with respect to ϕ is derived as follows:

$$\frac{\partial}{\partial\phi}\ln L(\omega,\theta,\phi) = \frac{n_0(1-\omega)\theta e^{-\theta/(1+\phi)}}{(1+\phi)^2\left[\omega + (1-\omega)e^{-\theta/(1+\phi)}\right]} + \sum_{d=1}^{m} n_d(d-1)\frac{d}{\theta + d\phi} - \sum_{d=1}^{m} d n_d \frac{1}{1+\phi} - \sum_{d=1}^{m} n_d \frac{d-\theta}{(1+\phi)^2}.$$

The second partial derivative of the log-likelihood function is:

$$\frac{\partial^2}{\partial\omega^2}\ln L(\omega,\theta,\phi) = -n_0 \frac{\left(1 - e^{-\theta/(1+\phi)}\right)^2}{\left(\omega + (1-\omega)e^{-\theta/(1+\phi)}\right)^2} - \frac{n-n_0}{(1-\omega)^2}.$$

$$\frac{\partial^2}{\partial\theta\partial\omega}\ln L(\omega,\theta,\phi) = n_0 \frac{e^{-\theta/(1+\phi)}}{(1+\phi)\left(\omega + (1-\omega)e^{-\theta/(1+\phi)}\right)^2}.$$

$$\frac{\partial^2}{\partial\phi\partial\omega}\ln L(\omega,\theta,\phi) = -n_0 \frac{\theta e^{-\theta/(1+\phi)}}{(1+\phi)^2\left(\omega + (1-\omega)e^{-\theta/(1+\phi)}\right)^2}.$$

$$\frac{\partial^2}{\partial\theta^2}\ln L(\omega,\theta,\phi) = \frac{n_0(1-\omega)\omega e^{-\theta/(1+\phi)}}{(1+\phi)^2\left[\omega+(1-\omega)e^{-\theta/(1+\phi)}\right]^2} - \frac{n-n_0}{\theta^2} - \sum_{d=1}^{m}\frac{n_d(d-1)}{(\theta+d\phi)^2}$$

The second partial derivative of $\ln L(\omega,\theta,\phi)$ with respect to ϕ is:

$$\frac{\partial^2}{\partial\phi^2}\ln L(\omega,\theta,\phi) = \frac{n_0(1-\omega)\theta e^{-\theta/(1+\phi)}\left[\omega\theta-2(1+\phi)\left(\omega+(1-\omega)e^{-\theta/(1+\phi)}\right)\right]}{(1+\phi)^4\left(\omega+(1-\omega)e^{-\theta/(1+\phi)}\right)^2}$$
$$-\sum_{d=1}^{m}\frac{n_d d^2(d-1)}{(\theta+d\phi)^2} + \frac{1}{(1+\phi)^2}\sum_{d=1}^{m} dn_d + \frac{2}{(1+\phi)^3}\sum_{d=1}^{m} n_d(d-\theta)$$

$$\frac{\partial^2}{\partial\theta\partial\phi}\ln L(\omega,\theta,\phi) = \frac{n_0(1-\omega)e^{-\theta/(1+\phi)}[-\theta\omega+(1+\phi)(\omega+(1-\omega)e^{-\theta/(1+\phi)})]}{(1+\phi)^3\left[\omega+(1-\omega)e^{-\theta/(1+\phi)}\right]^2} - \sum_{d=1}^{m}\frac{n_d(d-1)d}{(\theta+d\phi)^2} + \sum_{d=1}^{m}\frac{n_d}{(1+\phi)^2}.$$

The zero-inflated generalized Poisson (ZIGP) model effectively accommodates overdispersion and excess zeros commonly encountered in count data by integrating a degenerate distribution at zero with a generalized Poisson framework. Through maximum likelihood estimation, one can derive estimators for the model parameters ω, θ, and ϕ, with the associated log-likelihood and its partial derivatives providing essential tools for optimization and inference. The availability of closed-form expressions for the first and second-order derivatives of the log-likelihood function aids in numerical maximization techniques such as Newton-Raphson or Fisher scoring. The flexibility and analytical tractability of the ZIGP model make it a powerful alternative to traditional count models for analyzing zero-inflated and overdispersed data structures encountered in diverse fields such as epidemiology, insurance, and manufacturing.

Some Properties of MLEs of ZIGP Parameters

There are no closed-form expressions for the standard errors of the MLEs $\hat{\omega}$, $\hat{\theta}$, and $\hat{\phi}$. However, approximate standard errors can be obtained by computing the Fisher Information Matrix based on the sample data.

Suppose y_i (for $i = 1, 2, \ldots, n$) denotes the number of nonconformities observed in the i-th item of a sample of size n drawn from a ZIGP(ω,θ,ϕ) process. Let $(\hat{\omega},\hat{\theta},\hat{\phi})$ represent the maximum likelihood estimates of the unknown parameters (ω,θ,ϕ).

Then, the estimated probability of observing zero defects in an item from the ZIGP process, based on the fitted model, is given by:

$$\hat{P}_0 = P(y=0) = \hat{\omega} + (1-\hat{\omega})\exp\left(-\frac{\hat{\theta}}{1+\hat{\phi}}\right)$$

The Fisher Information Matrix J for the ZIGP model is a 3×3 symmetric matrix, expressed as:

$$J = \begin{pmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{pmatrix}$$

where each element J_{ij} represents the (i,j)-th element of the Fisher Information Matrix. And

$$
\begin{aligned}
J_{11} &= -\left[\frac{\partial^2}{\partial\omega^2}\ln L(\omega,\theta,\phi)\right],\\
J_{12} = J_{21} &= -\left[\frac{\partial^2}{\partial\theta\partial\omega}\ln L(\omega,\theta,\phi)\right],\\
J_{13} = J_{31} &= -\left[\frac{\partial^2}{\partial\phi\partial\omega}\ln L(\omega,\theta,\phi)\right],\\
J_{22} &= -\left[\frac{\partial^2}{\partial\theta^2}\ln L(\omega,\theta,\phi)\right],\\
J_{23} = J_{32} &= -\left[\frac{\partial^2}{\partial\theta\partial\phi}\ln L(\omega,\theta,\phi)\right],\\
J_{33} &= -\left[\frac{\partial^2}{\partial\phi^2}\ln L(\omega,\theta,\phi)\right].
\end{aligned}
$$

Once the Fisher Information Matrix is evaluated at the MLEs, its inverse provides the estimated asymptotic covariance matrix of the parameter estimates. The square roots of the diagonal elements of this inverse matrix yield approximate standard errors for $\hat{\omega}$, $\hat{\theta}$, and $\hat{\phi}$. These standard errors are crucial for constructing confidence intervals and conducting hypothesis tests. Although the derivation of the Fisher Information Matrix involves complex second-order derivatives, the analytical expressions obtained facilitate implementation in statistical software and enable reliable inference from ZIGP models applied to real-world zero-inflated count data.

3 Model Specification

The zero-inflated count model combines a point mass at zero with a count distribution:

$$
\begin{aligned}
&P(Y=0) = \omega + (1-\omega)e^{-\theta/(1+\phi)}\\
&P(Y=d) = (1-\omega)\frac{\theta(\theta+\phi d)^{d-1}}{(1+\phi)^d d!}e^{-(\theta+\phi d)/(1+\phi)}, \quad d = 1,2,\ldots
\end{aligned}
$$

4 Data Summary

y	Frequency
0	49
1	8
2	5
3	8
4	7
5	6
6	4
7	5
8	4
9	3
10	1

Table 1: Observed frequency distribution of counts

5 R Implementation

The maximum likelihood estimation was implemented in R:

```
y <- 0:10
freq <- c(49, 8, 5, 8, 7, 6, 4, 5, 4, 3, 1)
n <- sum(freq)
loglik <- function(params) {
  omega <- params[1]
  theta <- params[2]
  phi <- params[3]

if (omega <= 0 || omega >= 1 || theta <= 0 || phi <= 0) {
    return(Inf)
  }

  log_probs <- numeric(length(y))

  for (i in seq_along(y)) {
    yi <- y[i]
    if (yi == 0) {
      prob <- omega + (1 - omega) * exp(-theta / (1 + phi))
    } else {
      prob <- (1 - omega) * theta * (theta + phi * yi)^(yi - 1) / ((1 + phi)^yi
        ↪ * factorial(yi)) *
        exp(-(theta + phi * yi) / (1 + phi))
    }

    if (prob <= 0) {
      return(Inf)
    }

    log_probs[i] <- log(prob)
  }

  ll <- sum(freq * log_probs)

  return(-ll)
}

init_params <- c(0.5, 1, 0.5)

#Maximize log-likelihood (hessian = TRUE is added)
fit <- optim(
  par = init_params,
  fn = loglik,
  method = "L-BFGS-B",
  lower = c(0.0001, 0.0001, 0.0001),
  upper = c(0.9999, Inf, Inf),
  hessian = TRUE
)

# Step 5: Results
omega_hat <- fit$par[1]
theta_hat <- fit$par[2]
phi_hat <- fit$par[3]

cat("Estimated␣parameters:\n")
```

```
cat("omega␣=", omega_hat, "\n")
cat("theta␣=", theta_hat, "\n")
cat("phi␣=", phi_hat, "\n")
#######################################
# fitted probabilities
fitted_probs <- numeric(length(y))

for (i in seq_along(y)) {
  yi <- y[i]
  if (yi == 0) {
    prob <- omega_hat + (1 - omega_hat) * exp(-theta_hat / (1 + phi_hat))
  } else {
    prob <- (1 - omega_hat) * theta_hat * (theta_hat + phi_hat * yi)^(yi - 1) /
      ((1 + phi_hat)^yi * factorial(yi)) * exp(-(theta_hat + phi_hat * yi) / (1
         ↪  + phi_hat))
  }
  fitted_probs[i] <- prob
}
# Observed relative frequencies
observed_probs <- freq / sum(freq)
##Plot observed vs fitted
barplot(
  rbind(observed_probs, fitted_probs),
  beside = TRUE,
  names.arg = y,
  col = c("skyblue", "tomato"),
  legend.text = c("Observed", "Fitted"),
  args.legend = list(x = "topright"),
  xlab = "y",
  ylab = "Probability",
  main = "Observed␣vs␣Fitted␣Probabilities"
)
###################################
# Step 8: Standard errors
if (!is.null(fit$hessian)) {
  varcov <- tryCatch(
    solve(fit$hessian),
    error = function(e) {
      message("Warning:␣Hessian␣not␣invertible.")
      return(NULL)
    }
  )

  if (!is.null(varcov)) {
    se <- sqrt(diag(varcov))
    cat("\nStandard␣errors:\n")
    cat("SE(omega)␣=", se[1], "\n")
    cat("SE(theta)␣=", se[2], "\n")
    cat("SE(phi)␣=", se[3], "\n")
  }
}
```

6 Results

The maximum likelihood estimates are as follows:

Parameter	Estimate
ω (zero-inflation)	0.4695
θ (count parameter)	4.298
ϕ (dispersion parameter)	0.3199

Table 2: Parameter estimates

Parameter	Estimate	Standard Error
ω (zero-inflation)	0.4695	0.05317002
θ (count parameter)	4.298	0.4040731
ϕ (dispersion parameter)	0.3199	0.1687953

Table 3: Estimated parameters and their standard errors for the zero-inflated model

7 Goodness-of-Fit

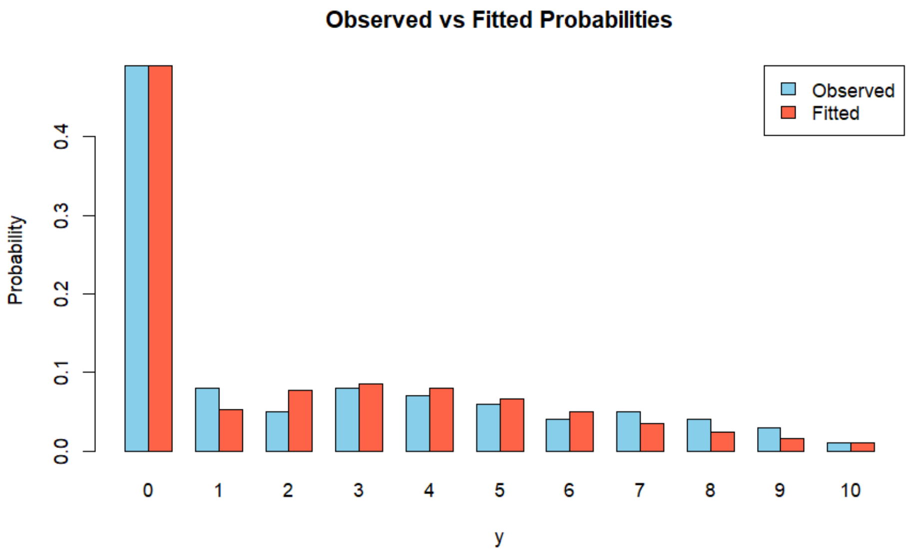


Figure 1: Comparison of observed and fitted probabilities

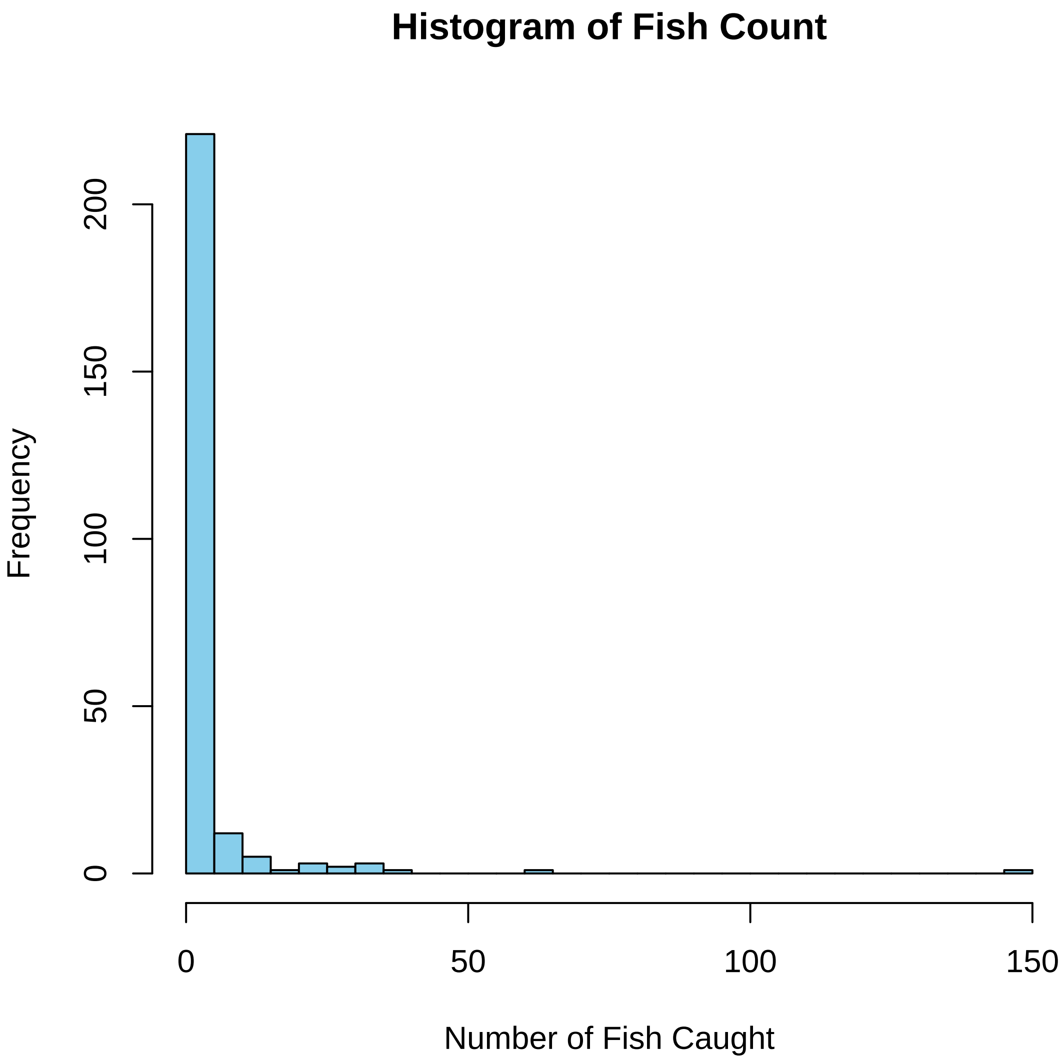
The fitted model captures zero inflation well, but may need adjustment for higher counts.

8 Modeling Excess Zeros with Zero-Inflated Negative Binomial Regression: Applications in R

Data Summary

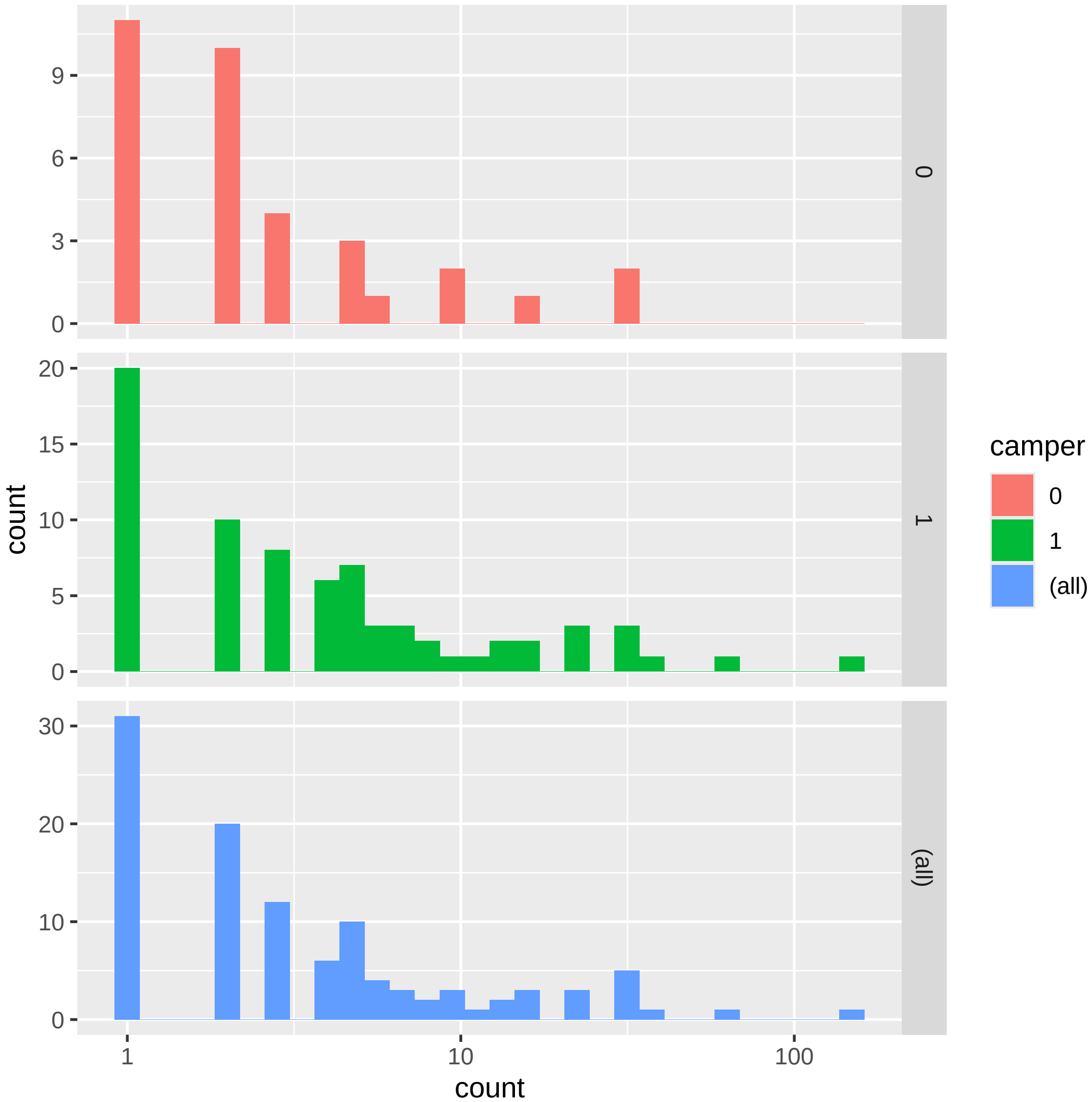
We analyze data from 250 groups that visited a park, each reporting the number of fish caught (`count`), the number of children in the group (`child`), the total number of people in the group (`persons`), and whether they brought a camper (`camper`, a binary variable). The primary objective is to model the number of fish caught and understand the factors contributing to excess zeros—cases where no fish were caught. Specifically, we aim to identify how group composition and camping status influence the likelihood of zero catches and the overall catch count.

```
require(ggplot2)
require(pscl)
require(MASS)
require(boot)
zinb <- read.csv("fish data.csv")
head(zinb)

##   nofish livebait camper persons child         xb         zg count
## 1      1        0      0       1     0 -0.8963146  3.0504048     0
## 2      0        1      1       1     0 -0.5583450  1.7461489     0
## 3      0        1      0       1     0 -0.4017310  0.2799389     0
## 4      0        1      1       2     1 -0.9562981 -0.6015257     0
## 5      0        1      0       1     0  0.4368910  0.5277091     1
## 6      0        1      1       4     2  1.3944855 -0.7075348     0

hist(zinb$count,
     breaks = 30,
     col = "skyblue",
     main = "Histogram of Fish Count",
     xlab = "Number of Fish Caught",
     ylab = "Frequency")
```




```r
table(zinb$count == 0)

##
## FALSE  TRUE
##   108   142

zinb <- within(zinb, {
  nofish <- factor(nofish)
  livebait <- factor(livebait)
  camper <- factor(camper)
})

summary(zinb)

##  nofish  livebait camper     persons          child              xb
##  0:176   0: 34    0:103   Min.   :1.000   Min.   :0.000   Min.   :-3.275050
```

```
## 1: 74   1:216    1:147   1st Qu.:2.000   1st Qu.:0.000   1st Qu.: 0.008267
##                          Median :2.000   Median :0.000   Median : 0.954550
##                          Mean   :2.528   Mean   :0.684   Mean   : 0.973796
##                          3rd Qu.:4.000   3rd Qu.:1.000   3rd Qu.: 1.963855
##                          Max.   :4.000   Max.   :3.000   Max.   : 5.352674
##        zg              count
## Min.   :-5.6259   Min.   :  0.000
## 1st Qu.:-1.2527   1st Qu.:  0.000
## Median : 0.6051   Median :  0.000
## Mean   : 0.2523   Mean   :  3.296
## 3rd Qu.: 1.9932   3rd Qu.:  2.000
## Max.   : 4.2632   Max.   :149.000

#############
## histogram with x axis in log10 scale
ggplot(zinb, aes(count, fill = camper)) +
  geom_histogram() +
  scale_x_log10() +
  facet_grid(camper ~ ., margins=TRUE, scales="free_y")
```



The graph supports the need for a zero-inflated model (e.g., zero-inflated Poisson or negative binomial) to account for the excess zeros and overdispersion. The camper variable appears to influence both the occurrence of zeros and the count distribution, warranting its inclusion in the model. Further modeling should assess its role in both the zero-inflation and count processes.

Zero-Inflated Negative Binomial Regression

The zero-inflated negative binomial (ZINB) regression model accounts for excess zeros by assuming that zero outcomes arise from two distinct processes. In the context of fishing, one process corresponds to groups that did not engage in fishing at all, for which the number of fish caught is necessarily zero. The second process involves groups that did go fishing, in which case the number of fish caught follows a count distribution—in this case, the negative binomial distribution.

The ZINB model combines a binary component and a count component. The binary component, typically modeled using logistic regression, estimates the probability that a zero count is structural (i.e., due to not fishing). The count component models the number of fish caught using a negative binomial distribution for groups that participated in fishing. The overall expected count is derived by combining these two components, effectively distinguishing between 'true' zero counts (from non-fishers) and 'sampling' zeros (from fishers who caught none).

$$E(n_{\text{fish caught}} = k) = P(\text{not gone fishing}) \times 0 + P(\text{gone fishing}) \times E(y = k \mid \text{gone fishing})$$

$$\text{PDF}(y; p, r) = \frac{(y_i + r - 1)!}{y_i!(r-1)!} p_i^r (1 - p_i)^{y_i}$$

where y_i is the count of successes, p_i is the probability of success, and r is the number of successes before the process stops. The likelihood function $L(\mu; y, \alpha)$ is given by:

$$L(\mu; y, \alpha) = \prod_{i=1}^{n} \exp\left(y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right)\right)$$

where y_i are the observed values, μ_i is the mean parameter, and α is a parameter that influences the distribution's variance.

Log-Likelihood Function

The log-likelihood function $\mathcal{L}(\mu; y, \alpha)$ is given by:

$$\mathcal{L}(\mu; y, \alpha) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right)\right]$$

where y_i are the observed values, μ_i is the mean parameter and α is a parameter that influences the variance of the distribution. which can be expressed in terms of our model by replacing μ_i with $\exp(x_i'\beta)$. Turning to the zero-inflated negative binomial model, the expression of the likelihood function depends on whether the observed value is a zero or greater than zero. From the logistic model of $Y_i > 1$ versus $Y_i = 0$:

$$p = \frac{1}{1 + e^{-x_i'\beta}}, \quad 1 - p = \frac{1}{1 + e^{x_i'\beta}}$$

$$\mathcal{L} = \begin{cases} \sum_{i=1}^{n} \left[\ln(p_i) + (1-p_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right] & \text{if } y_i = 0 \\ \sum_{i=1}^{n} \left[\ln(p_i) + \ln\Gamma\left(\frac{1}{\alpha} + y_i\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) + \left(\frac{1}{\alpha}\right)\ln\left(\frac{1}{1+\alpha\mu_i}\right) + y_i \ln\left(1 - \frac{1}{1+\alpha\mu_i}\right)\right] & \text{if } y_i > 0 \end{cases}$$

Finally, note that R does not estimate α but θ, the inverse of α.

Now let us build up our model. We are going to use the variables child and camper to model the count in the part of negative binomial model and the variable persons in the logit part of the model. We used pscl to perform a zero-inflated negative binomial regression. We begin by estimating the model with the variables of interest.

```
#m1 <- zeroinfl(count ~ child + camper | persons, data = zinb, dist = "negbin")
#summary(m1)
```

```
> m1 <- zeroinfl(count ~ child + camper | persons,
+                data = zinb, dist = "negbin")
> summary(m1)

Call:
zeroinfl(formula = count ~ child + camper | persons, data = zinb, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-0.5861 -0.4617 -0.3886 -0.1974 18.0135 

Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   1.3710     0.2561   5.353 8.64e-08 ***
child        -1.5153     0.1956  -7.747 9.41e-15 ***
camper1       0.8791     0.2693   3.265   0.0011 ** 
Log(theta)   -0.9854     0.1760  -5.600 2.14e-08 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)   1.6031     0.8365   1.916   0.0553 .
persons      -1.6666     0.6793  -2.453   0.0142 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Theta = 0.3733 
Number of iterations in BFGS optimization: 22 
Log-likelihood: -432.9 on 6 Df
```

The fitted model is a **Zero-Inflated Negative Binomial (ZINB)** regression model, which consists of two parts:

1. Count Model (Negative Binomial Regression with log link)

This part models the count of the response variable (e.g., number of events) when the outcome is not from the excess zeros:

$$\log(\mu_i) = 1.3710 - 1.5153 \cdot \text{child}_i + 0.8791 \cdot \text{camper}_i$$

where:

- μ_i is the expected count for observation i (given it's from the count component),
- `child` is a covariate indicating the number of children,
- `camper` is a binary variable (1 if the person is a camper, 0 otherwise).

The dispersion parameter θ for the negative binomial distribution is:

$$\theta = e^{-0.9854} \approx 0.3733$$

2. Zero-Inflation Model (Logistic Regression with logit link)

This part models the probability that an observation is from the structural zero group (i.e., always zero due to an excess-zero process):

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = 1.6031 - 1.6666 \cdot \text{persons}_i$$

where:

- π_i is the probability of an excess zero for observation i,
- `persons` is the number of persons in the group/household.

Combined Interpretation

- If `child` increases, the expected count decreases significantly (negative correlation coefficient).
- Being a `camper` increases the expected count (positive effect).
- More `persons` in a household reduce the probability of being in the zero-inflation group.

The model handles overdispersion via the negative binomial distribution and accounts for excess zeros via a logistic component.

9 Conclusion

This project explored the challenges of modeling count data with overdispersion and excess zeros, focusing on the Zero-Inflated Generalized Poisson (ZIGP),Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) regression models. Through theoretical analysis and empirical implementation, we demonstrate how these models effectively address the limitations of traditional Poisson regression when dealing with real-world data characterized by overdispersion and zero inflation.

From our analysis,the ZIGP distribution provides a flexible framework for handling both overdispersed and underdispersed count data, while the zero-inflation component accounts for excess zeros. The ZINB model offers an alternative approach when the negative binomial distribution is more appropriate for the count process.Maximum likelihood estimation proved effective for estimating ZIGP parameters, with the Fisher information matrix providing reliable standard errors. Our simulation study demonstrated the importance of correctly specifying the distribution when modeling count data.

The analysis of fish catch data illustrated the practical utility of zero-inflated models. The ZINB regression successfully identified significant predictors for both the count process (number of fish caught) and the zero-inflation process (probability of not fishing).The fitted models showed reasonable agreement with observed data, though some discrepancies at higher counts suggest potential areas for model refinement.

The project highlights that ignoring overdispersion and zero inflation can lead to biased estimates and incorrect inferences. The ZIGP and ZINB models provide robust alternatives that better capture the complexities of real-world count data. Future work could explore Bayesian approaches to these models or investigate more complex zero-inflated distributions for specific application domains.

In conclusion, when analyzing count data with excess zeros, researchers should carefully consider the underlying data generation process and select an appropriate model that accounts for both the counting mechanism and the zero-inflation process. The methods presented in this project offer valuable tools for such analyses across various scientific disciplines.

References

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). Modeling zero-inflated count data with glmmtmb. *BioRxiv*, page 132753.

Consul, P. C. and Jain, G. C. (1973). A generalization of the poisson distribution. *Technometrics*, 15(4):791–799.

Dean, C. B. (1992). Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457.

Kamalja, K. K. and Wagh, Y. S. (2018). Estimation in zero-inflated generalized poisson distribution. *Journal of Data Science*, 16(1):183–206.

McCullagh, P. and Nelder, J. A. (1989). Binary data. In *Generalized linear models*, pages 98–148. Springer.

Shahsavari, S., Moghimbeigi, A., Kalhor, R., Jafari, A. M., Bagherpour-Kalo, M., Yaseri, M., and Hosseini, M. (2023). Zero-inflated count regression models in solving challenges posed by outlier-prone data; an application to length of hospital stay. *Archives of Academic Emergency Medicine*, 12(1):e13.

Zamani, H. and Ismail, N. (2014). Functional form for the zero-inflated generalized poisson regression model. *Communications in Statistics-Theory and Methods*, 43(3):515–529.