PROPERTIES AND APPLICATIONS OF THE ZERO-INFLATED REGRESSION MODEL

Tanmay Gayen Roll No. : DSTC-24/25-018

Preetam Biswas Roll No.:DSTC-24/25-014

INDIAN STATISTICAL INSTITUTE(Chennai)

PGDSMA COURSE

INTRODUCTION:

In this project , we explore statistical models for count data, focusing on challenges like overdispersion and excess zeros, which are common in real world datasets. Traditional poisson models often fail under such conditions. To address this, we examine flexible alternatives such as the zero-inflated poisson (ZIP), zero-inflated negative binomial (ZINB), and especially the zero-inflated generalized poisson (ZIGP) regression model. This model effectively handles both overdispersion and zero inflation, making it suitable for a wide range of applications.

Problem statement :

Challenges in modeling count data with excess zeros and dispersion issues

Equidispersion assumption limitation: The poisson regression model assumes equal mean and variance (equidispersion) , which is often violated in real-world count data, leading to overdispersion (variance > mean).

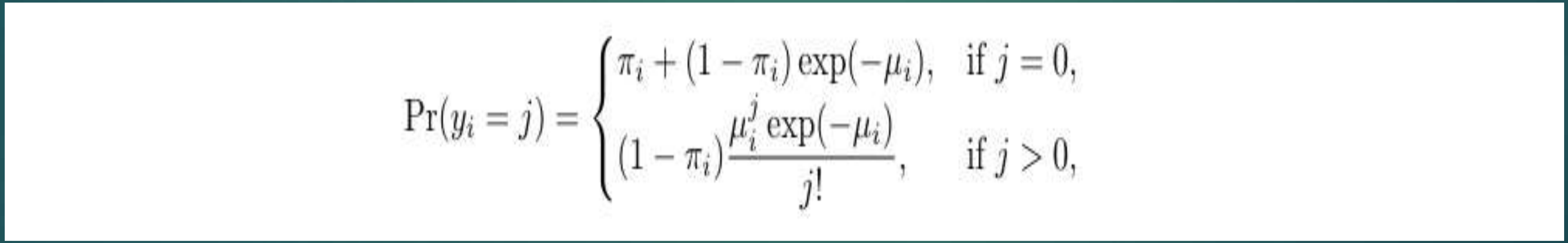
Excess zero inflation: Many datasets, especially in health and insurance, exhibit more zero counts than expected under standard poisson or negative binomial models, requiring specialized zero-inflated models.

The Zero-inflated Poisson Distribution:

The ZIP model is used for count data with an unusually high number of zeros. It assumes that zeros can come from two sources:

A **structural source** (always zero), with probability π,and **poisson process**, with probability (1−π), which can generate both zeros and positive counts.

This model combines a **binary (logit)** component to model excess zeros and a **poisson** component to handle the count data. It helps improve model fit when standard poisson regression underestimates the frequency of zeros.



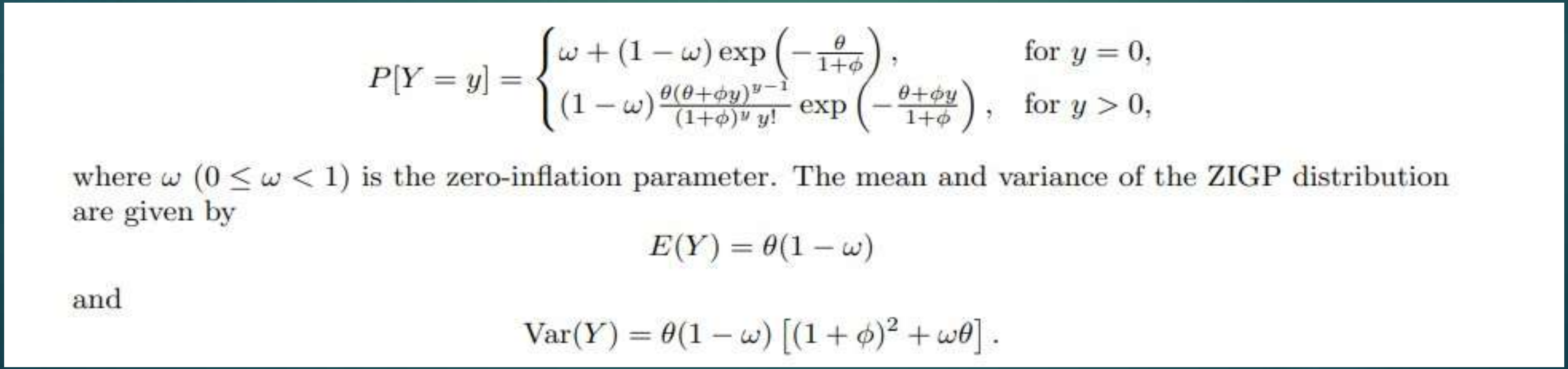
$$\Pr(y_i = j) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\mu_i), & \text{if } j = 0, \\ (1-\pi_i)\dfrac{\mu_i^j \exp(-\mu_i)}{j!}, & \text{if } j > 0, \end{cases}$$

The Zero-inflated generalized Poisson Distribution:

The ZIGP model is designed for count data with both **overdispersion** and **excess zeros**. It combines a **bernoulli distribution at zero** (to account for structural zeros) with a **generalized poisson distribution** (to model the count data). The model includes:

- **Θ** as the mean parameter,
- **Φ** as the dispersion parameter, and
- **w** as the zero-inflation parameter.

It offers greater flexibility than zip or zinb models and is estimated using **maximum likelihood methods**, making it ideal for complex real-world data



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

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

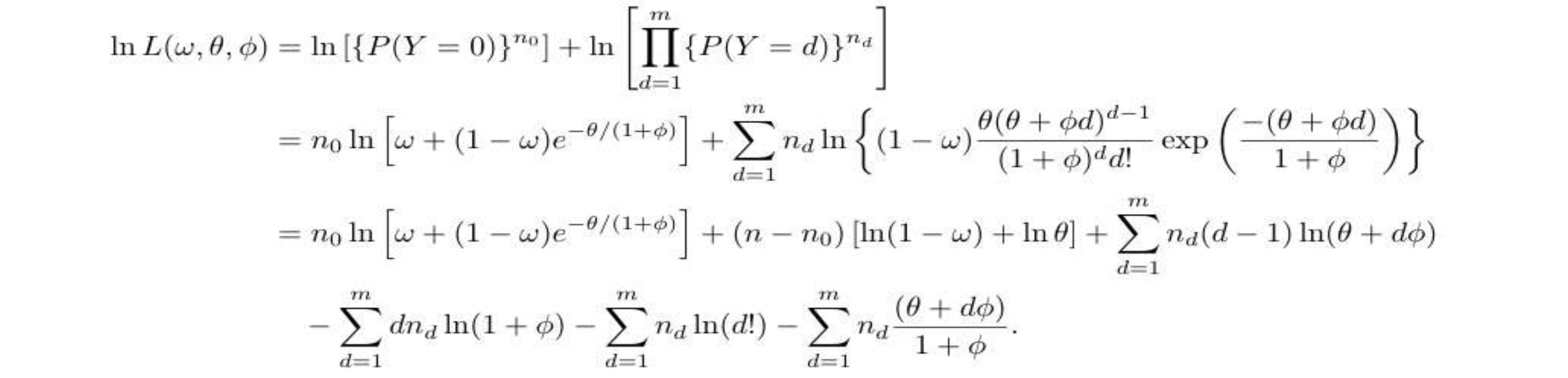
$$E(Y) = \theta(1-\omega)$$

and

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

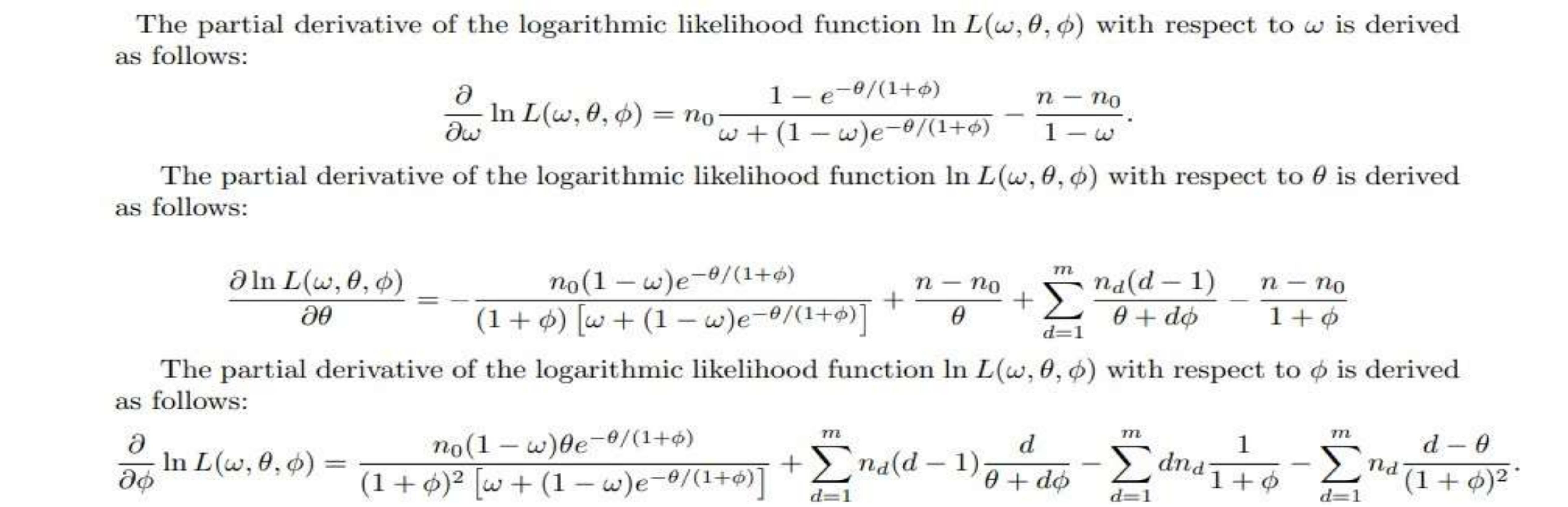
Log-likelihood function of ZIGP Distribution:

Then the log-likelihood function for the parameters ω,θ,andφ 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 log-likelihood function:



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

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

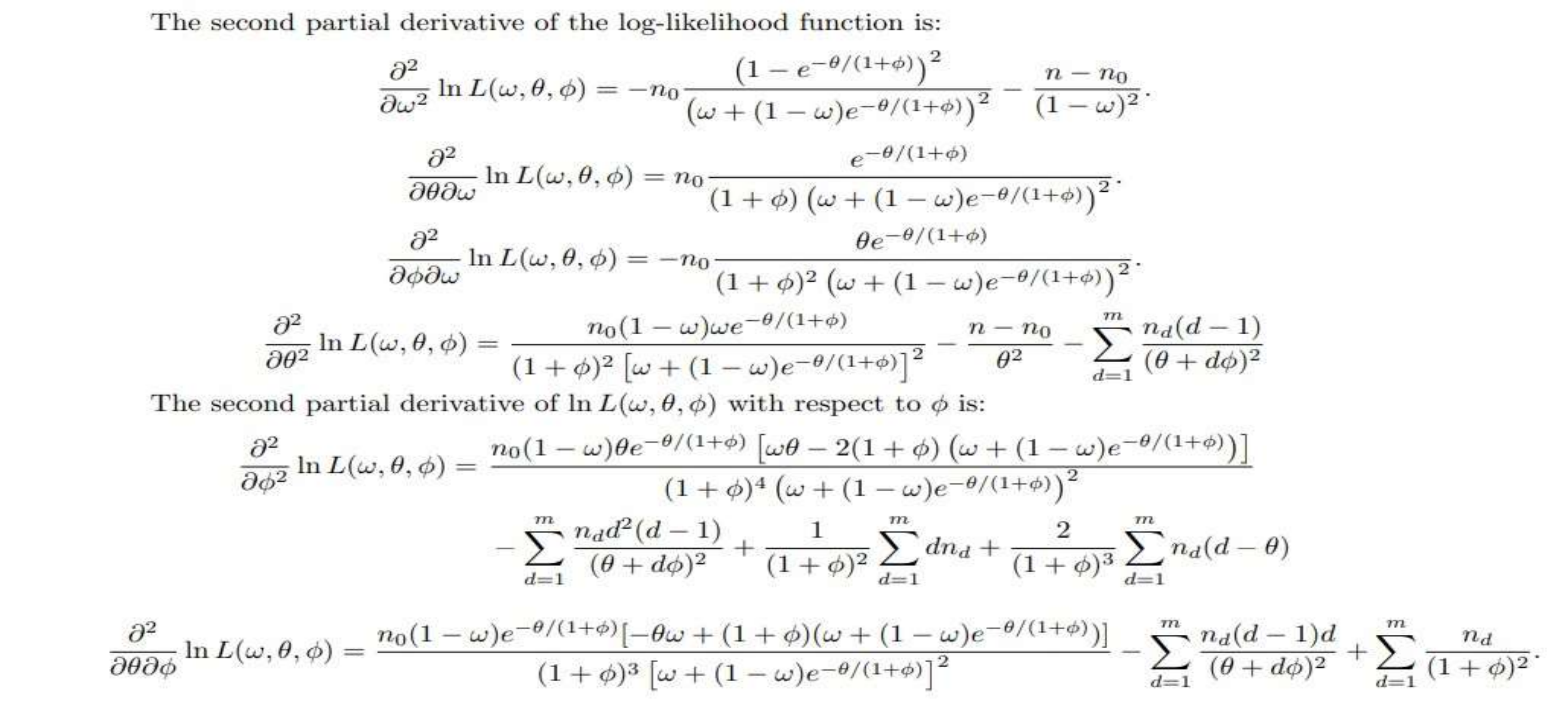
The partial derivative of the logarithmic likelihood function $\ln L(\omega, \theta, \phi)$ with respect to θ is derived as follows:

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

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:



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:

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

$$-\sum_{d=1}^{m}\frac{n_d d^2(d-1)}{(\theta+d\phi)^2} + \frac{1}{(1+\phi)^2}\sum_{d=1}^{m}dn_d + \frac{2}{(1+\phi)^3}\sum_{d=1}^{m}n_d(d-\theta)$$

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

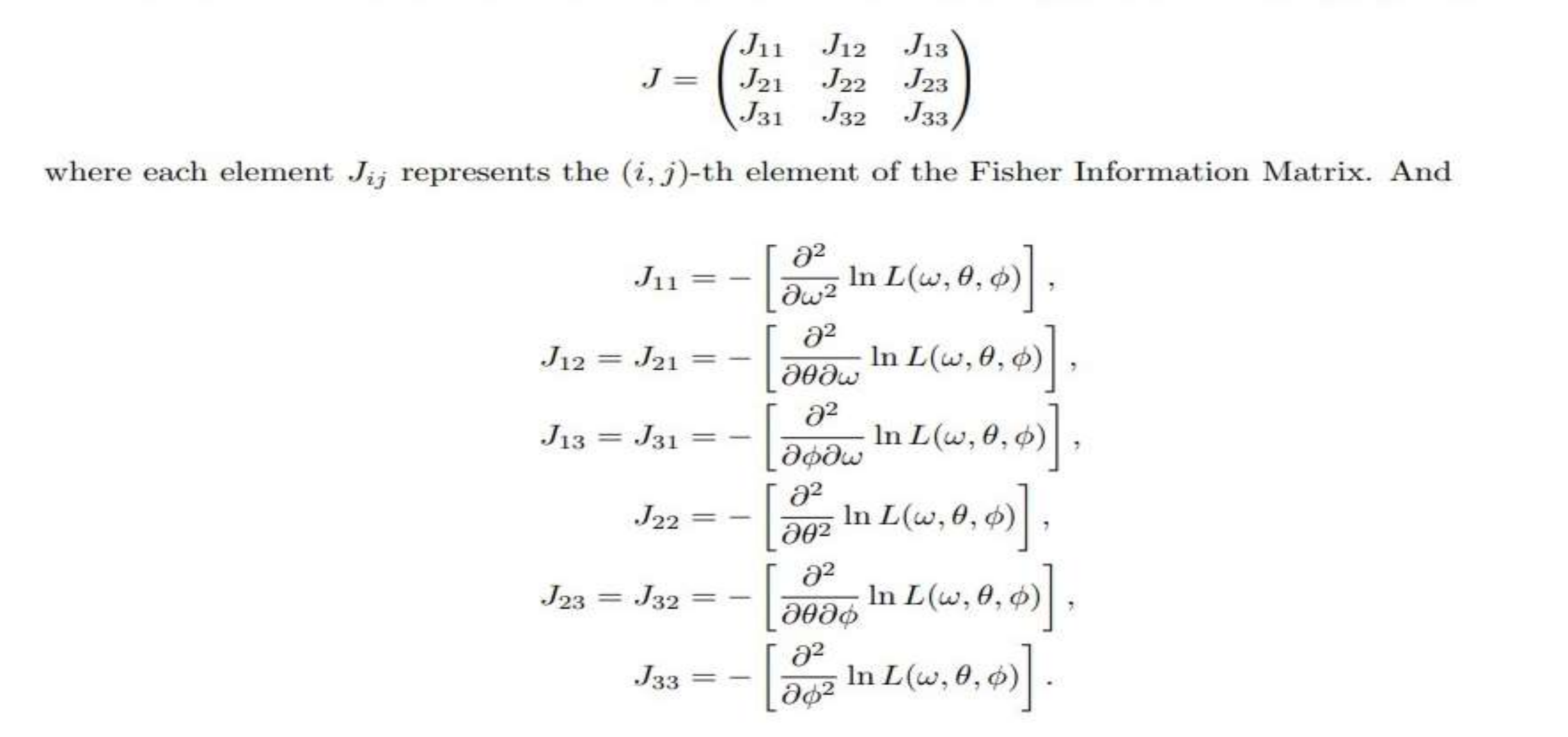
Some Properties of MLEs of ZIGP Parameters:

- There are no closed-form expressions for the standard errors of the MLEs $\hat{\omega}, \hat{\theta}, \hat{\phi}$.However, approximate standard errors can be obtained by computing the Fisher Information Matrix based on the sample data.
- Suppose yi (for i = 1,2,...,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:

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

Fisher information matrix:

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



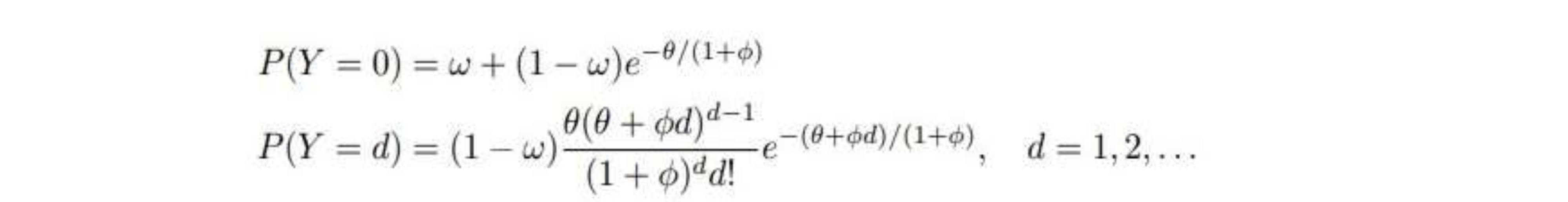
$$J = \begin{pmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{pmatrix}$$

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

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

Model specification:

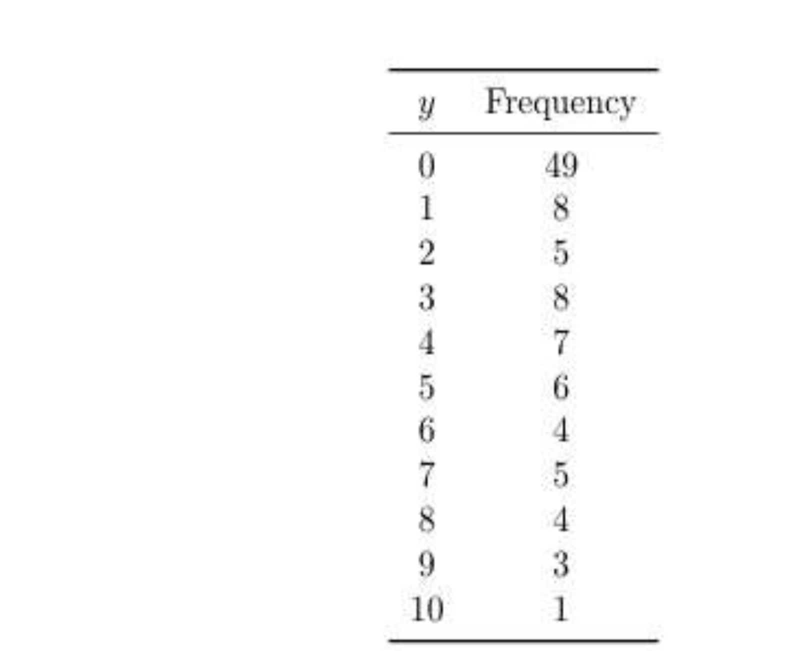
- The zero-inflated count model combines a point mass at zero with a count distribution:



$$P(Y=0)=\omega+(1-\omega)e^{-\theta/(1+\phi)}$$

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

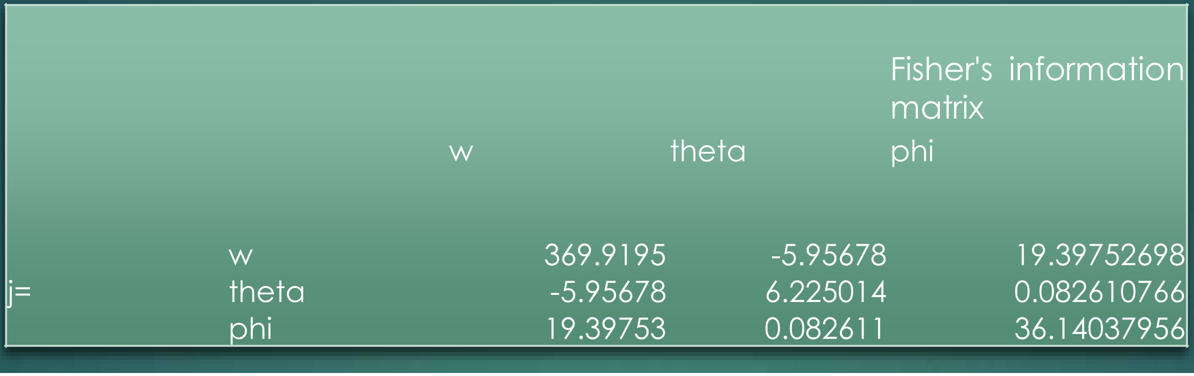
Data summary:



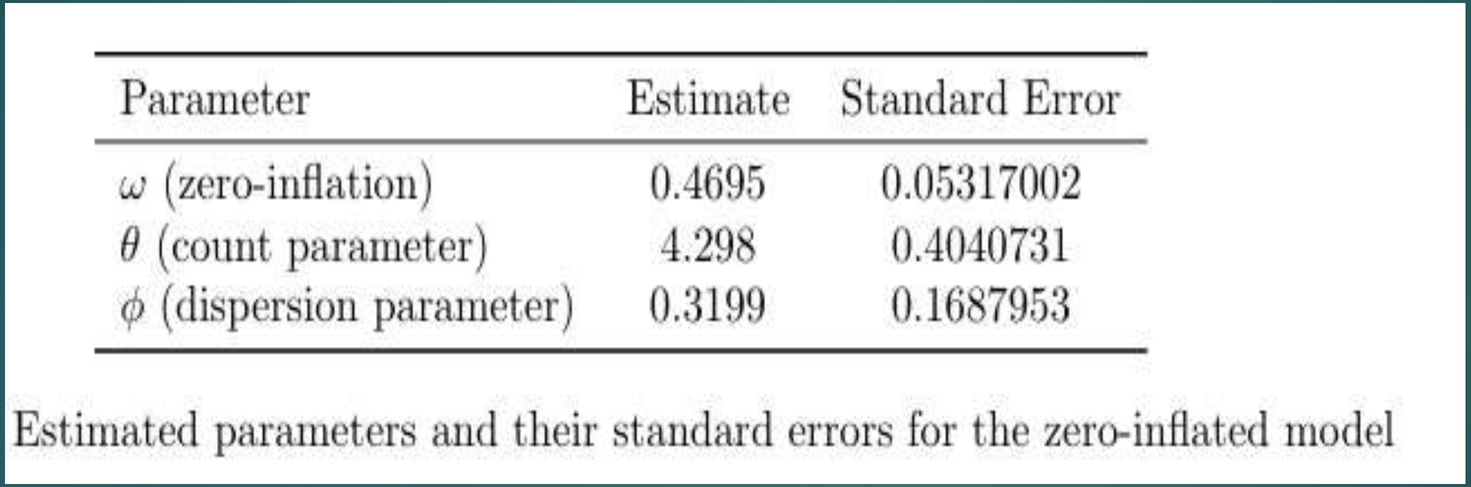
y	Frequency
0	49
1	8
2	5
3	8
4	7
5	6
6	4
7	5
8	4
9	3
10	1

Observed frequency distribution of counts

Results:



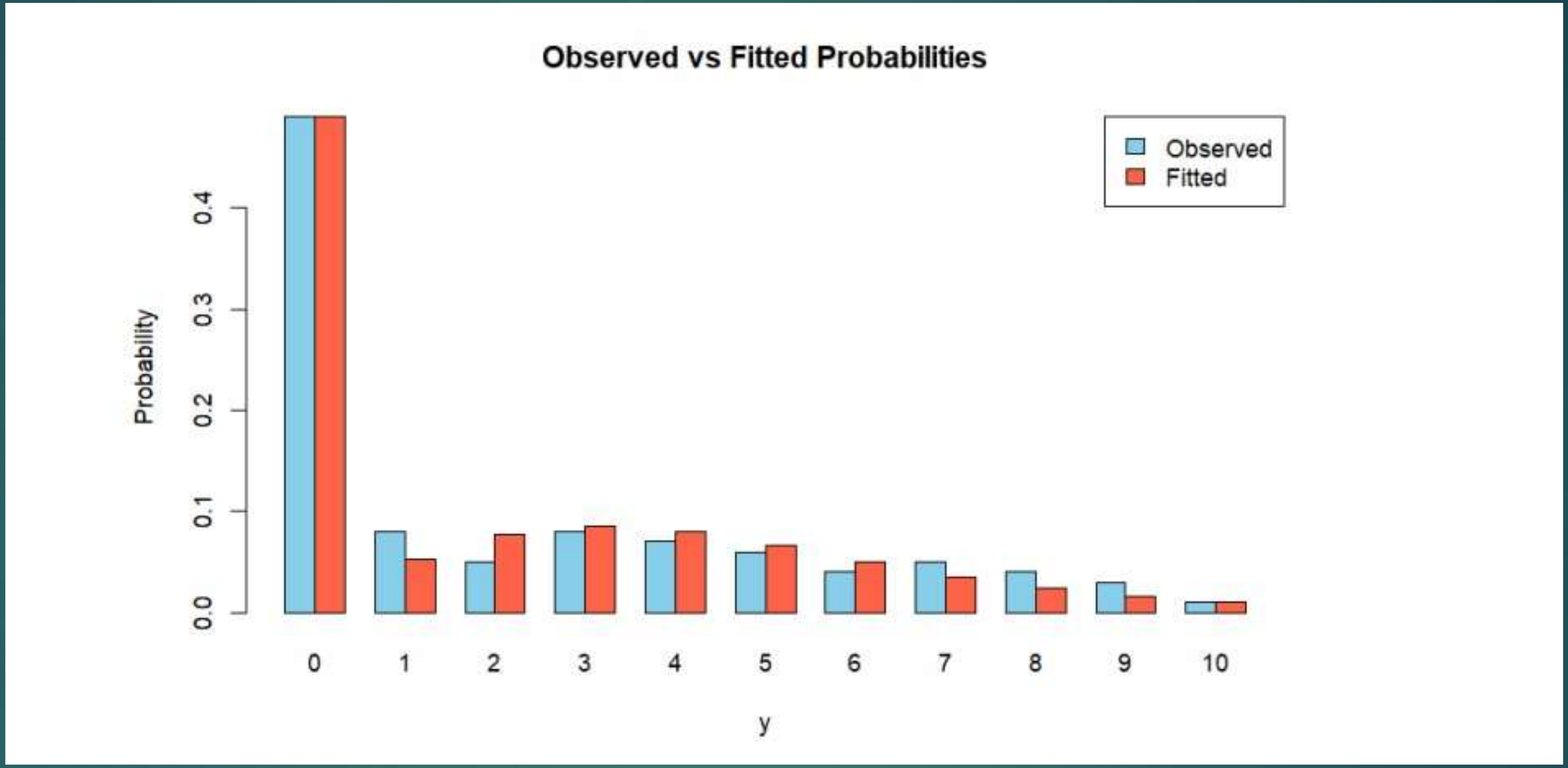
				Fisher's information matrix
		w	theta	phi
	w	369.9195	-5.95678	19.39752698
j=	theta	-5.95678	6.225014	0.082610766
	phi	19.39753	0.082611	36.14037956



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

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

Goodness-of-Fit:



The fitted model captures zero inflation well.

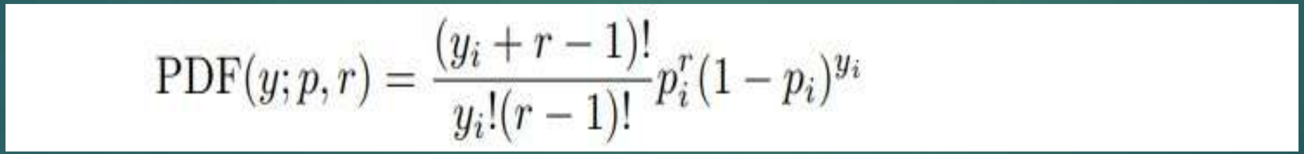
Zero inflated negative binomial:

Zero-Inflated Negative Binomial (ZINB) Model

The ZINB model is used for count data with **overdispersion** and **excess zeros**. It assumes that zeros come from two sources:

1. **Structural zeros** (e.g., no attempt occurred), modeled using a **logistic (binary) component**, and
2. **Sampling zeros and positive counts**, modeled using a **Negative Binomial distribution**.

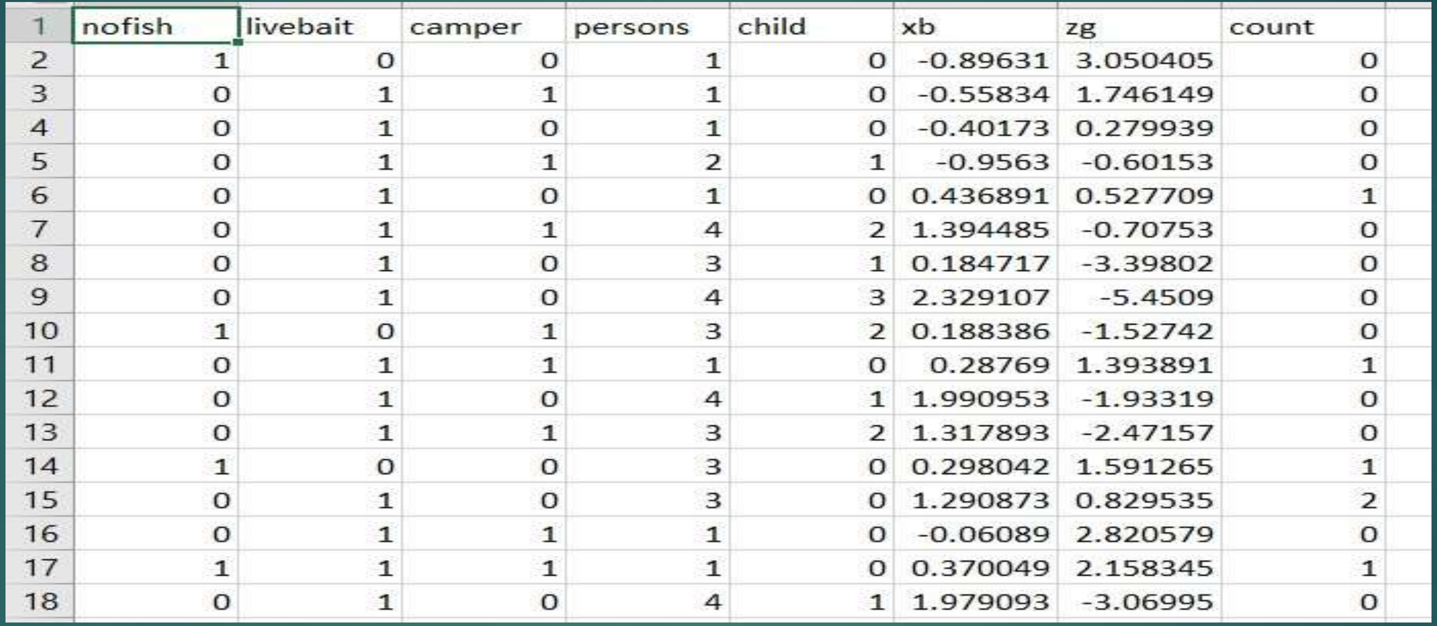
- This model effectively separates true zero events from count-generating processes, providing better estimates for highly variable and zero-heavy data.
- PDF of ZINB



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

- Where y _i is the observed count (number of successes), p_i is the probability of success, and r is the number of successes before the process stops.

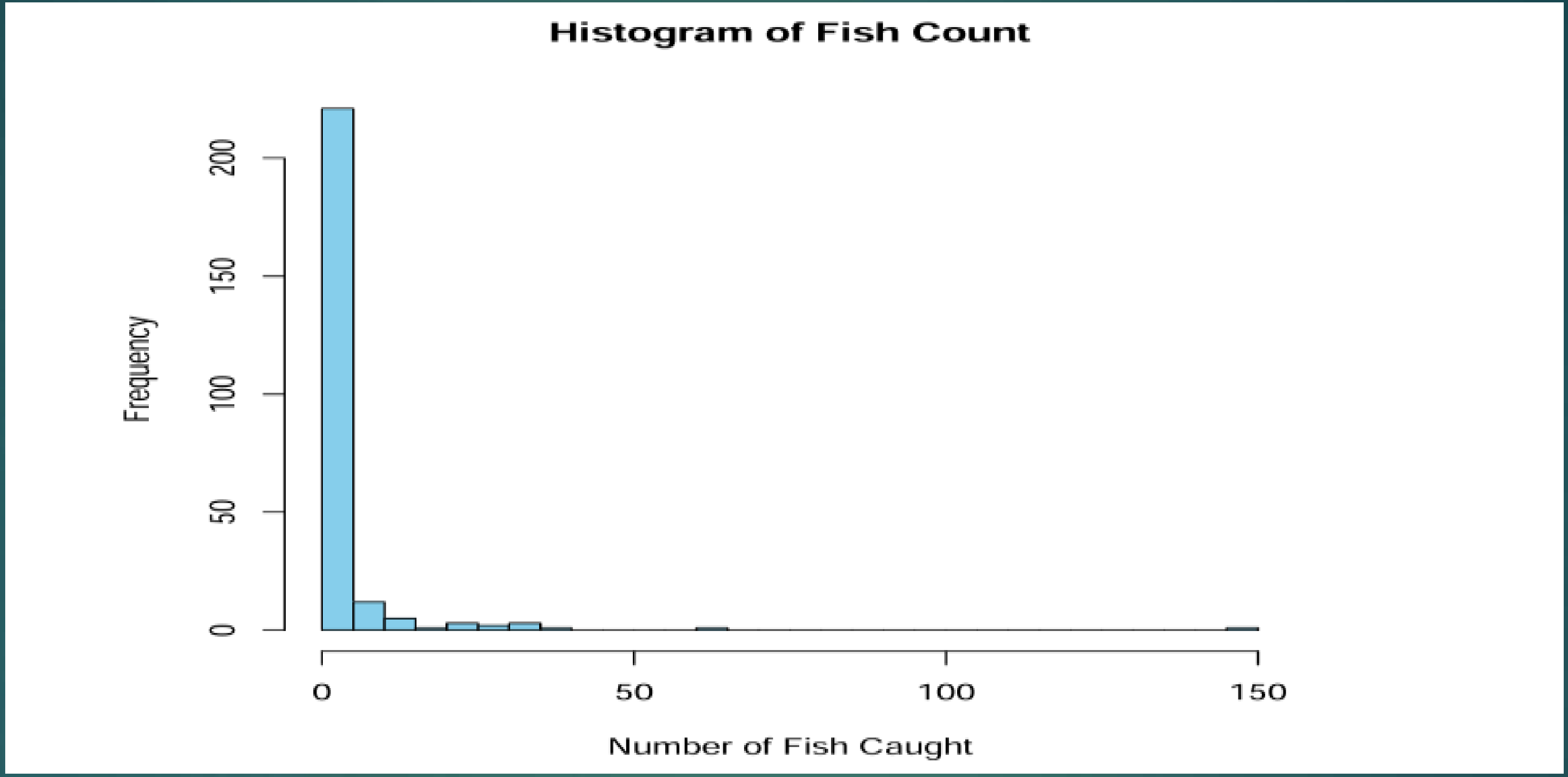
Modeling Excess Zeros with Zero-Inflated Negative Binomial Regression: (Data set)



1	nofish	livebait	camper	persons	child	xb	zg	count
2	1	0	0	1	0	-0.89631	3.050405	0
3	0	1	1	1	0	-0.55834	1.746149	0
4	0	1	0	1	0	-0.40173	0.279939	0
5	0	1	1	2	1	-0.9563	-0.60153	0
6	0	1	0	1	0	0.436891	0.527709	1
7	0	1	1	4	2	1.394485	-0.70753	0
8	0	1	0	3	1	0.184717	-3.39802	0
9	0	1	0	4	3	2.329107	-5.4509	0
10	1	0	1	3	2	0.188386	-1.52742	0
11	0	1	1	1	0	0.28769	1.393891	1
12	0	1	0	4	1	1.990953	-1.93319	0
13	0	1	1	3	2	1.317893	-2.47157	0
14	1	0	0	3	0	0.298042	1.591265	1
15	0	1	0	3	0	1.290873	0.829535	2
16	0	1	1	1	0	-0.06089	2.820579	0
17	1	1	1	1	0	0.370049	2.158345	1
18	0	1	0	4	1	1.979093	-3.06995	0

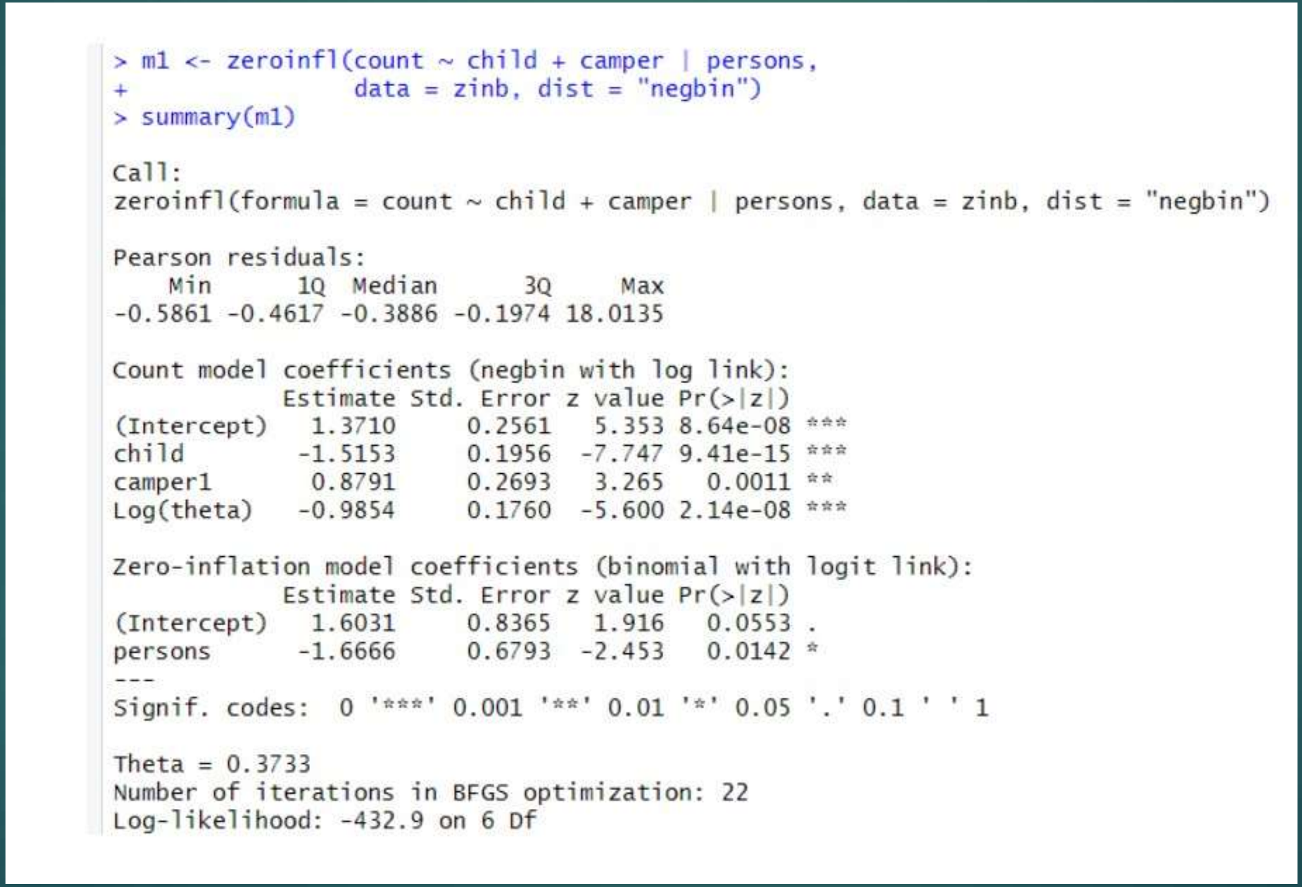
- We analyze data from 250 groups that visited a park, each reporting the number of fish caught (count), the number of children in the group (child), the total number of people in the group (persons), and whether they brought a camper (camper, a binary variable). The primary objective is to model the number of fish caught and understand the factors contributing to excess zeros—cases where no fish were caught.

Histogram of Data set:



- We see that from the histogram the ZINB regression model well fitted.

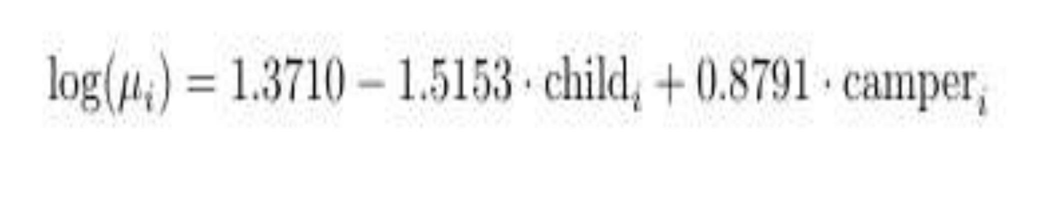
Analysis of Data set:



```
> 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
```

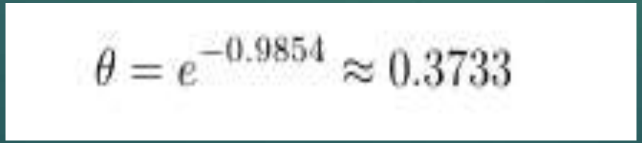
1. Count Model (Negative Binomial Regression with log link):

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



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

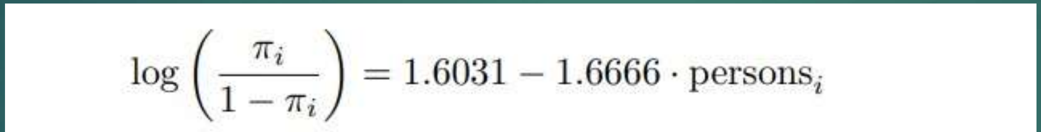
where: µ i is the expected count for observation i (given it’s from the count component), child is a covariate indicating the number of children, camper is a binary variable (1 if the person is a camper, 0 otherwise). The dispersion parameter θ for the negative binomial distribution is:



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

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:

- • πi is the probability of an excess zero for observation i,
- • persons is the number of persons in the group/household

Combined Interpretation

- If the number of children increases, the expected count decreases significantly (negative correlation coefficient).
- The camper status has a positive impact on the expected number of events.
- 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.

Conclusion:

This project investigates statistical models suitable for count data exhibiting overdispersion and excess zeros, specifically focusing on Zero-Inflated Generalized Poisson (ZIGP), Zero-Inflated Poisson (ZIP), and Zero-Inflated Negative Binomial (ZINB) regression models. Traditional Poisson regression often fails in such contexts, leading to biased inferences.

The ZIGP model provides flexibility to handle both overdispersed and underdispersed data, while the ZINB model is particularly effective when the Negative Binomial distribution fits the count process better. The models incorporate a zero-inflation component to account for excess zeros, and maximum likelihood estimation (MLE) is used for parameter estimation, with the Fisher information matrix aiding in standard error computation.

A simulation study underscored the importance of correctly specifying the underlying distribution. An empirical application using fish catch data demonstrated the practical utility of zero-inflated models, with ZINB effectively identifying predictors for both the count and zero-inflation processes . The findings emphasize that ignoring overdispersion and zero inflation can lead to misleading conclusions. ZIGP and ZINB provide robust alternatives. Future work may involve Bayesian extensions or more complex models tailored to specific domains.

Thank You