B.SC. STATISTICS (HONOURS) SEMESTER VI

STATISTICAL MODELLING OF COUNT DATA WITH OVER-DISPERSION AND ZERO-INFLATION PROBLEMS

NAME : SAMRAT HALDER
ROLL NO : 438
DEPARTMENT: STATISTICS
Paper CODE : HSTDS6043D
SUPERVISOR : PROF. MADHURA DAS GUPTA

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

STUDENT'S SIGNATURE: _____

DATE: 05.04.2023

Contents

Chapter 1

1. INTRODUCTION

1.1 Overview

Count data is encountered on daily basis and dealings. In econometrics the interest in count data models is a reflection of the general interest in modelling discrete aspects of individual economic behaviour. Count data models are specific types of discrete data regression. For example: Consumer demand – Number of products that a consumer buys on Amazon,

Health demand – The number of doctor visit

Quite often in statistics we come across the term "count data". Modelling count data is a common task in economics and social sciences. The classical Poisson regression model for count data is often of limited use in these disciplines because empirical datasets may typically exhibit an over-dispersion and/or excess number of zeros.

This issue can be addressed by fitting a negative binomial regression model instead of a Poisson regression model. However, although these models being part of the family of generalised linear models which capture over-dispersion rather well, they are, in many situations, not applicable for modelling excess zeros. To address this issue zero augmented models where a second model component captures only zero counts are used. Hurdle models combine a left truncated count component with a right censored hurdle component. Zero inflated models take somewhat a different approach; they are mixture models that combine a count component and a point mass at zero.

1.2 Motivating Example

In this study, we will analyse German health survey data for the year 1998 to model the relationship between number of visits to doctor(numvisit) and patient's health condition(badh) and patient's age. The dataset consists of 1127 observations on the following 3 variables

numvisit : number of visits to doctor during 1998

badh: 1= patient claims to be in bad health

0= not in bad health

age: age of patient—>20-60

1.3 Outline of the project

The main motivation behind this paper is to fit an appropriate model to the above described dataset taking numvisit as the response variable(y) and badh(x1), age(x2) as the predictors. The data is a zero inflated count dataset where occurrence of zeros is excess, there is a clear indication of presence of over-dispersion is the dataset. We perform a overdispersion test on the dataset to validate the above indication. For the Poisson model, since there are zero inflation and over-dispersion issues in the response, Hurdle Poisson model, Hurdle Negative Binomial model, zero-inflated- Poisson (ZIP) model and zero-inflated-negative-binomial (ZINB) model would be used to handle these issues. Model comparisons among the Poisson model hurdle model, ZIP and ZINB is conducted to select the best model to fit the data using the AIC criterion and the cross-validation criterion.

Chapter 2

2. GRAPH OF THE DATA SET

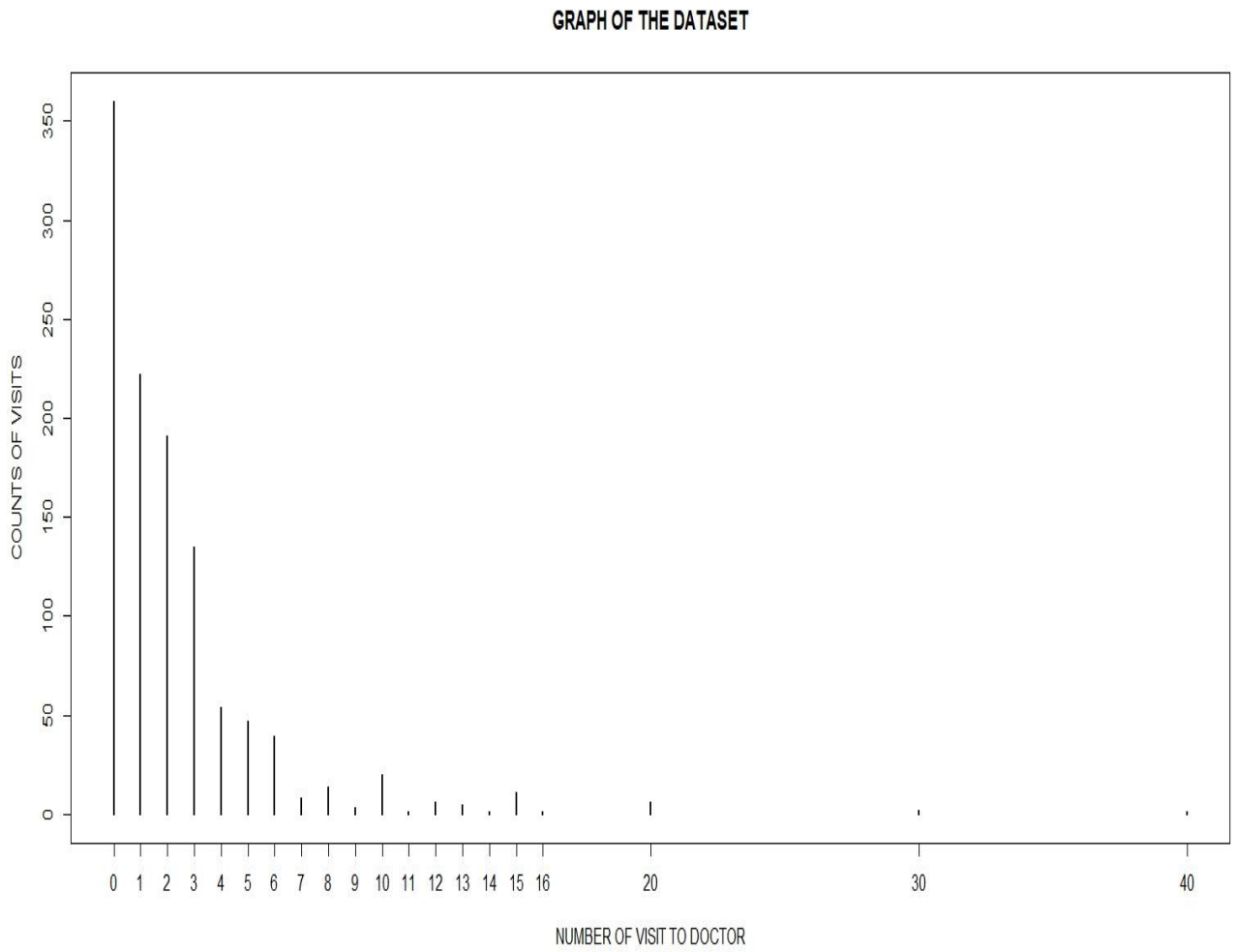

Fig1

Chapter 3

3. ORDINARY LEAST SQUARE REGRESSION

Firstly, we fit the model Y (No. of visits) on X1(Health condition) and X2 (Age of the patients) through Ordinary Least Square (OLS) method.

The regression model is $yi = xi\beta + \epsilon i$

This method requires certain significant assumptions. The assumptions are Normality of errors, constant variance, independence of errors, assumption of consistency, asymptotic normality etc. From fig1 we can say that this is a count data set. we suspect that the normality of errors assumption would not hold good. As, it is a count data set we can't apply OLS.

Chapter 4

4. Generalised Linear Models (GLM)

The GLM is an extension of general linear model. It allows the specification of models whose error term distribution is any one from the exponential family such as Poisson, Binomial, Negative Binomial etc., and not only normal. That is, GLM models generalise the error term distribution to the exponential family. For example, Logistic regression (where dependent variable is categorical), Poisson (where dependent variable is a count variable) etc. are GLMs. It generalizes linear regression by allowing the model parameters to be linearly related to the mean of response variable via a link function and by allowing the magnitude of

the variance of each measurement to be a function of its predicted value. For example, a Binomial regression can use a logit or a probit link function, a Poisson regression uses a log link function, and so on. The model still assumes that observations are independent.

A common simplest approach to model count data is Poisson Regression. Let us fit a Poisson model to our data regressing no. of visits on the other two predictors as mentioned before.

Chapter 5

5. Poisson Regression

The simplest distribution used for modelling count data is the Poisson distribution with probability density function

The model :

Let $y_1, y_2, y_3, \ldots y_n$ be iid sample, where Y is a random variable denoting no. of visits to doctor in the year 1998. Y is a count variable following Poisson distribution with E(Y) =μ.

$x_{i1}, x_{i2}, \ldots x_{im}$ be m linearly independent regressors that are thought to determine y_i , i=1(|)1127.

$\underline{\beta} = (\beta_0, \beta_1, \ldots \beta_m)$ are (m+1)-dimensional vector of parameters to be estimated on the basis of the data.

The Poisson regression model specifies that y_i given $\underline{x_i}$ is independently Poisson distributed with density:

$$f(y_i \mid \underline{x}_i) = \frac{e^{-\mu_i}\ \mu_i{}^{y_i}}{y_i!},\ y_i = 0,1,2,\ldots.$$

And mean parameter:

E [Y|$\underline{x}_i$]= μ_i, $\mu_i > 0$.

Poisson distribution is a special case of the generalized linear model (GLM) framework. The canonical link is the *log* function resulting in a log-linear relationship between the mean and the linear predictor. The variance in the Poisson model is identical to the mean. Poisson regression models allow researchers to examine the relationship between predictors and count outcome variables.

We model the conditional mean μ_i by a log link function,

The model: Log $\mu_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$, i= 1(1)n=1127.

i.e., $\mu_i = e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}$ ---------- (1)

Estimation of model parameters:

We estimate the parameters by maximum likelihood method.

The likelihood equation is:

$$L(\underline{\beta}) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\, \mu_i^{y_i}}{y_i!}$$

Taking logarithm on both sides,

$$\text{Ln } L(\underline{\beta}) = \sum_{i=1}^{n} (-\mu_i + y_i \ln \mu_i - \ln y_i!)$$

$$= -\sum_{i=1}^{n} e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}} + \sum_{i=1}^{n} y_i (\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}) - \sum_{i=1}^{n} y_i! \quad \text{------(2)} \quad \text{from (1)}$$

Differentiating w.r.t β_j and equating to 0, we obtain the Maximum Likelihood Estimate of $\beta_j's,\ j = 1(1)m = 3$

$$\frac{\partial \text{Ln } L(\underline{\beta})}{\partial \beta_j} = 0 \quad \Rightarrow \quad -\sum_{i=1}^{n} e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}\, x_{ij} + \sum_{i=1}^{n} y_i x_{ij} = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}) x_{ij} = 0 \text{ ----------(3)}$$

the equations ③ are not in closed form . We solve the equations numerically using **Newton-Raphson** method to obtain the maximum likelihood estimates of $\beta_j's$, where j = 1,2,3.

As a first attempt to capture the relationship between number of visit to doctor and patient's health condition and patient's age, we fit the basic Poisson regression model. The output is given in Table 1

Table1: Poisson model output

Predictors	Estimate	Std Error	Z value	P- value
intercept	0.447022	0.071428	6.258	3.89e-10
badh	1.108331	0.046169	24.006	< 2e-16
age	0.005822	0.001822	3.195	0.0014

Now, as we have observed from fig1 a large no. of zero count values in our data, we may be interested to check whether our model is able to make satisfactory prediction of those counts.

Chapter 6

6. PROBLEM OF EXCESS ZEROS

If the amount of observed zeros is larger than the amount of predicted zeros, the model is underfitting zeros which indicates a zero-inflation in the data.

We observed 360 zero counts but our model only predicts about 147 zero value. The zero counts are being severely under fitted. The excess zero values in the data need to be taken care of.

6.1 Score test for zero inflation

Here we run an independent test to ensure the presence of excess zeros in the dataset. The test procedure is performed by first calculating the rate estimate $\hat{\lambda} = \bar{x}$. Then we count the number of observed zeros, denoted by n_0 and the total no of observations n. We then calculate the probability ($\widetilde{p_0}$) of observing zero which is given by $\widetilde{p_0} = \exp(-\hat{\lambda})$. The test hypothesis is given by,

H_0: The dataset is not zero inflated against $H_1: H_0^c$

Therefore, the test statistic for the above testing procedure is given by the formula,

$$S_{zi} = \frac{(n_0 - n\widetilde{p_0})^2}{n\widetilde{p_0}(1-\widetilde{p_0}) - n\bar{x}\widetilde{p_0}^2} \overset{H_0}{\rightarrow} \square^2_{(1)}$$

We reject H_0 if the p-value of the test procedure is less than 0.05 and conclude that the dataset is zero inflated.

Here the p-value for the testing of presence of zero inflation in the dataset is coming out to be 2.22e-16 which is much less than the desired level of significance, 0.05. Hence we reject H_0 and conclude that the dataset is zero inflated i.e. in other words, the frequency of observing zeros is more than observing other non-negative values in the dataset.

Chapter 7

7. Zero Augmented Models

In this section we briefly outline two models based on a zero-augmented distribution, which incorporates an autoregressive binary choice component and thus captures the (potentially different) dynamics of both zero occurrences and of strictly positive realizations.

There are two methods (models) to handle excess zero counts:

1. Hurdle-at-zero model
2. Zero inflated model.

Both the two models deal with high occurrence of zeros in the observed data but have an important distinction in how they interpret and analyse zero counts.

In a zero-inflated (ZI) model (Lambert 1992), zero observations have two different origins: “structural” and “sampling”. In other hand, a hurdle model (Mullahy 1986) is a two part model which assumes all zero data are from one ‘structural’ source. The positive (i.e., non-zero) data have ‘sampling’ origin following a truncated count distribution.

More clearly, Hurdle model assumes two types of subjects:

1. Those who never experience the outcome.
2. Those who always experience the outcome at least once.

Zero-inflated models conceptualize subjects as:

1. Those who never experience the outcome.
2. Those who can experience the outcome but not always.

In our data, we can intuitionally consider the zero values to originate from only one source. An individual either do not visit to the doctor at all or visits a positive number of time during the given year 1998. The interpretation here would be that one process governs whether a patient visits to doctor in the given year at least once . If yes, another process separately governs how many visits are made. Thus, intuitionally, a hurdle model seems appropriate.

We, therefore, go on to fit a Hurdle-at-zero Poisson model to our data.

7.1 Hurdle at Zero Poisson Models

Hurdle Count Data Models in general, were introduced by Mullahy (1986). These models allow for a systematic difference in the statistical process governing individuals (obsevations) with zero counts and individuals with one or more counts.

As mentioned, this model helps to handle the excess zeros in the dataset which cannot be predicted by the usual Poisson model.

Again as mentioned before, Hurdle model is a two part model:

1. One generating the zeros: - A binary outcome of whether a count variable (here, no. of visits) has a zero or positive value, governed by a Binary probability model.
2. One generating the positive values: - If y_i>0 (i.e., no. of visits > 0), the "hurdle is crossed". The conditional distribution of the positive values is governed by a zero-truncated Poisson model (a count model).

We consider the same setup, variables and parameters, as described in the standard Poisson model above, for the truncated Poisson part here.

The probability function of the hurdle-at-zero model is given by:

$$P(y_i=k) = \begin{cases} \pi_i & ; k=0 \\ (1-\pi_i)\ \dfrac{e^{-\mu_i}\ {\mu_i}^{y_i}}{y_i!\ (1-e^{-\mu_i})} & ; k = 1,2,3,\dots \end{cases}$$

π_i is the probability of zero count value.

We write, $\pi_i = \frac{e^{\delta_0+\sum_{j=1}^{k}\delta_j z_{ij}}}{1+e^{\delta_0+\sum_{j=1}^{k}\delta_j z_{ij}}}$, i= 1 (l) n

i.e., $\ln\frac{\pi_i}{(1-\pi_i)} = \delta_0 + \sum_{j=1}^{k}\delta_j z_{ij}$ ----------(2)

Where, $\delta = (\delta_0, \delta_1, \ldots\ldots, \delta_k)$ are the (k+1) dimensional vector of parameters to be estimated on the basis of data.

The vector of covariates $\underline{x_i}$' and z_i' of the two parts of the model may contain exactly the same variables or may be different.

Also, $\mu_i = e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}$ [as in equation (1)]

Let us define an indicator variable:

$$I_i = \begin{cases} 1 & if\ y_i = 0 \\ 0 & if\ y_i = 1,2,.. \end{cases} \qquad \text{i= 1(l)n}$$

Estimation of parameters:

Given independent observations, the likelihood function of the hurdle model is given by:

L ($\underline{\beta}, \underline{\delta}$)= L1 ($\underline{\delta}$) + L2 ($\underline{\beta}$)

Where, L1 ($\underline{\delta}$) is the likelihood function for the Binary process

and L2($\underline{\beta}$) is the likelihood function of the truncated Poisson process.

The two mechanisms are assumed to be independent. The likelihood functions of both parts are maximised separately to obtain the MLE's of all the parameters of the hurdle model.

Now, L1 ($\underline{\delta}$)= $\prod_{i=1}^{n}\pi_i{}^{I_i}(1-\pi_i)^{(1-I_i)}$

Ln L1 ($\underline{\delta}$)= $\sum_{i=1}^{n} I_i \ln\pi_i + \sum_{i=1}^{n}(1-I_i)\ln(1-\pi_i)$

$$= \sum_{i=1}^{n} \ln(1-\pi_i) + \sum_{i=1}^{n} y_i \ln \frac{\pi_i}{(1-\pi_i)}$$

$$= -\sum_{i=1}^{n} \ln(1+e^{\delta_0+\sum_{j=1}^{k}\delta_j z_{ij}}) + \sum_{i=1}^{n} y_i(\delta_0+\sum_{j=1}^{k}\delta_j z_{ij}) \quad \{\text{using (2)}\}$$

Now differentiating w.r.t δ_j and equating to 0, we obtain the MLE of $\delta_j's$, j=1 (l) k.

$$\frac{\partial \text{Ln L1}(\delta)}{\partial \delta_j} = 0 \quad => -\sum_{i=1}^{n} \frac{e^{\delta_0+\sum_{j=1}^{k}\delta_j z_{ij}} z_{ij}}{1+e^{\delta_0+\sum_{j=1}^{k}\delta_j z_{ij}}} + \sum_{i=1}^{n} y_i z_{ij} = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - \pi_i) z_{ij} = 0$$

The above set of equations are not in a closed form. Newton-Raphson numerical method is opted to solve these and obtain the MLE of $\underline{\delta}$.

Again, the likelihood function for the truncated Poisson is given by:

$$\text{L2}(\underline{\beta}) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!(1-e^{-\mu_i})}$$

$$\text{Ln L2}(\underline{\beta}) = \sum_{i=1}^{n}[-\mu_i + y_i \ln \mu_i - \ln y_i! - \ln(1-e^{-\mu_i})]$$

$$= -\sum_{i=1}^{n} e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}} + \sum_{i=1}^{n} y_i(\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}) - \sum_{i=1}^{n} y_i!$$

$$-\sum_{i=1}^{n}(1-e^{-e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}}) \qquad \text{---------- [From (1)]}$$

Now, differentiating w.r.t β_j and equating to 0, we obtain the Maximum Likelihood Estimate (MLE) of $\beta_j's,\ j = 1(|)m.$

$$\frac{\partial \text{Ln L2}(\underline{\beta})}{\partial \beta_j} = 0 \quad => -\sum_{i=1}^{n} e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}} x_{ij} + \sum_{i=1}^{n} y_i x_{ij} -$$

$$\sum_{i=1}^{n} \frac{e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}} e^{-e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}} x_{ij}}{1-e^{-e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}}} = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}} - \frac{e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}} e^{-e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}}}{1-e^{-e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}}}\right) x_{ij} = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \mu_i - \frac{\mu_i\, e^{-\mu_i}}{1- e^{-\mu_i}} \right) x_{ij} = 0$$

These equations are also solved by Newton-Raphson method, and MLE of $\underline{\beta}$ is obtained.

In case of our data, we have n=1127. The explanatory variables for both the parts are considered to be same i.e., $\underline{x_i}$' and z_i' contain the same variables. Then, m=k=3.

The regression equations obtained are as follows:

Count model (truncated Poisson with log link):

$$\text{Ln } \mu_i = 0.679398 + 0.876388\, x_{i1} + 0.008999\, x_{i2}$$

Zero hurdle model (binomial with logit link):

$$ln \frac{\pi_i}{(1-\pi_i)} = 0.967776 + 1.262903\, x_{i1} - 0.008290\, x_{i2}$$

The estimates of model parameters are in the following table

Table2: Hurdle at Zero Poisson model output

Count model coefficients (truncated Poisson with log link)				
Predictors	**Estimate**	**Std. Errors**	**z value**	**p value**
Intercept	0.679398	0.077140	8.807	< 2e-16
badh	0.876388	0.047917	18.290	< 2e-16
age	0.008999	0.001946	4.624	3.76e-06
Zero – hurdle model coefficients (binomial with logit link)				
Predictors	**Estimate**	**Std. Errors**	**z value**	**p value**
Intercept	0.967776	0.232376	4.165	3.12e-05
badh	1.262903	0.288887	4.372	1.23e-05
age	-0.008290	0.006049	-1.371	0.171

Now, again comparing the no. of zero counts predicted by the model with that of the observed, we find the predicted no. of zeros to be 360, which happens to be the no. of zeros in the observed data. The Hurdle model will always predict the same number of zeros as observed. Thus our problem of zero is resolved.

Chapter 8

8. Models for dealing with over-dispersion

In Statistics, **overdispersion** is the presence of greater variability in a data set than would be expected based on a given Statistical model.

From the data we calculate mean= 2.35315 and variance= 11.98175.

We know, equi-dispersion property of the Poisson distribution. The equality of mean and variance, i.e mean=variance. This is a restrictive property and often fails to hold in practice. Here the variability of our observed data is much greater than what a Poisson model predicts . Hence, we suspect overdispersion is present in our data and Poisson regression do not seem good enough.

The Negative Binomial model is used with count data instead of the Poisson model if there is over dispersion in the data.

8.1 Score test for overdispersion

Poisson model is a special case of negative binomial model. The negative binomial regression model reduces to the Poisson regression model when the overdispersion parameter $\alpha \to 0$. To access the adequacy of the negative binomial model over the Poisson regression model, we can test the hypothesis:

$$H_o: \alpha = 0 \quad \text{against} \quad H_1: \alpha > 0$$

This is to test the significance of the value of the overdispersion parameter α. The presence of the overdispersion parameter α in the NB regression model is justified when the null hypothesis H_o: α=0 is rejected. In order to test the above mentioned hypothesis, a score test statistic (S_{od}) has been given below.

$$S_{od} = \frac{[\sum_{i=1}^{n}\{(y_i-\widehat{\mu_i})^2-y_i\}]^2}{2\sum_{i=1}^{n}\widehat{\mu_i}^2} \xrightarrow{H_o} \chi^2_{(1)}$$

where $\widehat{\mu_i}$ is the estimated value after fitting a poisson regression model. Under the hypothesis, the limiting distribution of the score statistic is chi-squared with one degree of freedom. We will reject H_o if the p-value of the test is less than 0.05, which is the level of significance for the above testing procedure.

The value of the test statistic S_{od} for testing the presence of overdispersion in the dataset is coming out to be 71.61476. And the value of $\chi^2_{0.05,(1)}$ is 3.814. So, we reject H_0 and conclude that α is significantly greater than zero. Thus we can say that there is a presence of overdispersion in the dataset.

8.2 Negative Binomial Regression

Another way of modelling over-dispersed count data is to assume a negative binomial (NB) distribution. Unlike Poisson model the Negative Binomial model has less restrictive property that the variance is not equal to the mean, the variance and mean are not equivalent, variance > mean.

The Negative Binomial model also estimates the overdispersion parameter α.

We have three cases,

When α=0 the mean and variance will be equal and the distribution boils down to Poisson.

When α > 0 variance exceeds the mean and the distribution allows for overdispersion.

Finally, $\alpha < 0$ mean exceeds the variance and the distribution allows for under dispersion.

The model :

Just as in Poison regression, let $y_1, y_2, y_3, \ldots y_n$ be iid sample, where Y is a random variable denoting no. of visits to doctor in the year 1998. Y is a count variable following Negative Binomial distribution with E(Y) =μ.

The Negative Binomial regression model specifies that y_i given $\underline{x_i}$ is independently Negative Binomial distributed with density:

$$f(y_i \mid \underline{x_i}) = \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad , \quad y_i = 0,1,2,\ldots.$$

And mean parameter: $E[Y|\underline{x_i}] = \mu_i \ , \ \mu_i > 0.$

The conditional variance: $V[Y|\underline{x_i}] = (\mu_i + \alpha\mu_i^2) > \mu_i$

We model the conditional mean μ_i by a log link function to ensure non-negativity of the mean parameterμ_i.

The model: $\text{Log}\ \mu_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$, i= 1(i)n.

i.e., $\mu_i = e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}$

$$L = \sum_{i=1}^{n} \{\ln[\Gamma(y_i + \alpha^{-1})] - \ln[\Gamma(\alpha^{-1})] - \ln[\Gamma(y_i + 1)] - \alpha^{-1}\ln(1 + \alpha\mu_i) - y_i \ln(1 + \alpha\mu_i) + y_i \ln(\alpha) + y_i \ln(\mu_i)\}$$

Rearranging gives

$$L = \sum_{i=1}^{n} \{(\sum_{j=0}^{y_i - 1} \ln(j + \alpha^{-1})) - \ln[\Gamma(y_i + 1)] - (y_i + \alpha^{-1})\ln(1 + \alpha\mu_i) + y_i \ln(\alpha) + y_i \ln(\mu_i)\}$$

The parameters are to be estimated by maximum likelihood method.

Differentiating w.r.t β_j and α respectively and equating with 0

$\frac{\partial L}{\partial \beta_j} = 0$ ---------- (1)

$\frac{\partial L}{\partial \alpha} = 0$ ----------(2)

We solve the equations numerically using **Newton-Raphson** method to obtain the maximum likelihood estimates of $\beta_j's$, where j = 1,2,3. And α

Using maximum likelihood method estimate of α is 0.9975

Because we already showed the Poisson model is with over-dispersion, we use negative binomial model to fit the data here.

Table3: Negative Binomial Model output

Predictors	**Estimate**	**Std Error**	**Z value**	**P- value**
intercept	0.404116	0.130847	3.088	0.00201
badh	1.107342	0.111603	9.922	< 2e-16
age	0.006952	0.003397	2.047	0.04070

Since the data is overdispersed, Negative Binomial model may show better fit, as per theory. And we are going to fit Hurdle-at-Zero Negative Binomial model to our data as data has excess zeros.

8.3 HURDLE-AT-ZERO NEGATIVE BINOMIAL MODEL

The zero generating part of hurdle model is governed by a Binomial distribution and the positive counts are governed by a zero truncated Negative Binomial distribution.

The Hurdle-at-zero Negative Binomial model is given by:

$$f(y_i \mid \underline{x}_i) = \begin{cases} \pi_i & ; y_i = 0 \\ (1-\pi_{i)} \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \dfrac{\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}}{1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}} & ; y_i = 1,2,3,\ldots. \end{cases}$$

π_i is the probability of zero count value.

We write, $\pi_i = \dfrac{e^{\delta_0 + \sum_{j=1}^{k} \delta_j z_{ij}}}{1 + e^{\delta_0 + \sum_{j=1}^{k} \delta_j z_{ij}}}$, i= 1 (l) n.

Where, $\delta = (\delta_0, \delta_1, \ldots\ldots, \delta_k)$ are the (k+1) dimensional vector of parameters to be estimated on the basis of data.

Also, $\mu_i = e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}$

Estimation of parameters:

Given independent observations, the likelihood function of the hurdle model is given by:

$$L(\underline{\beta}, \underline{\delta}) = L1(\underline{\delta}) + L2(\underline{\beta}, \alpha)$$

Where, $L1(\underline{\delta})$ is the likelihood function for the Binary process (1st part of hurdle model)

and $L2(\underline{\beta}, \alpha)$ is the likelihood function of the truncated Negative Binomial process. (2nd part of Hurdle model)

The two mechanisms are assumed to be independent. The likelihood functions of both parts are maximised separately to obtain the MLE's of all the parameters of the hurdle model.

Now, L1 ($\underline{\delta}$)= $\prod_{i=1}^{n} \pi_i{}^{I_i}(1-\pi_i)^{(1-I_i)}$

Where, $I_i = \begin{cases} 1 & if\ y_i = 0 \\ 0 & if\ y_i = 1,2,.. \end{cases}$ i= 1(l)n

Then, Ln L1 ($\underline{\delta}$)= $\sum_{i=1}^{n} I_i \ln \pi_i + \sum_{i=1}^{n}(1-I_i)\ \ln(1-\pi_i)$

Putting value of π_i, differentiating Ln L ($\underline{\delta}$) w.r.t. δ_j, (j=1(|) k), and equating to 0, just in the same way as in Hurdle Poisson model, we obtain :

$$\frac{\partial \text{Ln L1}(\underline{\delta})}{\partial \delta_j} = 0 \quad => \ \sum_{i=1}^{n}(\ y_i - \pi_i\ )\ z_{ij}\ = 0$$

These set of equations are solved by Newton Raphson iterative method and MLE's of $\underline{\delta}$ are obtained.

Again, the likelihood function for the truncated Negative Binomial model is given by:

$$\text{L2}\ (\underline{\beta}, \alpha) = \prod_{i=1}^{n} \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \frac{(\frac{1}{1+\alpha\mu_i})^{1/\alpha}\ (\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i}}{1-(\frac{1}{1+\alpha\mu_i})^{1/\alpha}}$$

Then, Ln L2 ($\underline{\beta}, \alpha$) = $\sum_{i=1}^{n}[\ \sum_{j=0}^{y_i-1}\{\ln(j+\frac{1}{\alpha})\} - \ln y_i! - (y_i + \frac{1}{\alpha})\ln(1+\alpha\mu_i) +$

$$y_i \ln(\alpha\mu_i) - \ln(1 - (\frac{1}{1+\alpha\mu_i})^{1/\alpha}\ )\ ]$$

To maximise the log likelihood function, we take derivative of the function w.r.t. α and $\beta_j's$[j=1(|) m], and equate it to 0.

$$\frac{\partial \text{Ln L2}\ (\underline{\beta},\alpha)}{\partial \alpha} = 0 \quad => \sum_{i=1}^{n}[\ \sum_{j=0}^{y_i-1}\{\frac{1}{(j+\frac{1}{\alpha})}\ (-\frac{1}{\alpha^2})\} + \frac{1}{\alpha^2}\ln(1+\alpha\mu_i)\ - (y_i + \frac{1}{\alpha})\frac{\mu_i}{1+\alpha\mu_i} +$$

$$\frac{y_i}{\alpha} - \frac{\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\mu_i}{\alpha(1+\alpha\mu_i)}\frac{\ln(1+\alpha\mu_i)}{\alpha^2}\right)}{1-(\frac{1}{1+\alpha\mu_i})^{1/\alpha}}\ ] = 0$$

$$\Rightarrow \sum_{i=1}^{n}[\sum_{j=0}^{y_i-1}\{\frac{j}{1+j\alpha}\}+\frac{1}{\alpha^2}\ln(1+\alpha\mu_i)\ -\left(y_i+\frac{1}{\alpha}\right)\frac{\mu_i}{1+\alpha\mu_i}-$$

$$\frac{\alpha\mu_i-(1+\alpha\mu_i)\ln(1+\alpha\mu_i)}{\alpha^2(1+\alpha\mu_i)\{(1+\alpha\mu_i)^{\frac{1}{\alpha}}-1\}}\ ]=0 \qquad \text{----------(i)}$$

Again, $\frac{\partial \text{Ln L2}\ (\underline{\beta},\alpha)}{\partial\beta_j}=0 \Rightarrow \sum_{i=1}^{n}\left\{-\left(y_i+\frac{1}{\alpha}\right)\frac{\alpha}{1+\alpha\mu_i}\frac{\partial\mu_i}{\partial\beta_j}+\frac{y_i}{\mu_i}\frac{\partial\mu_i}{\partial\beta_j}-\frac{\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}+1}}{1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}}\frac{\partial\mu_i}{\partial\beta_j}\right\}=0$

$$\Rightarrow \sum_{i=1}^{n}\left\{\frac{y_i}{\mu_i}\frac{\partial\mu_i}{\partial\beta_j}-\frac{1+\alpha y_i}{1+\alpha\mu_i}\frac{\partial\mu_i}{\partial\beta_j}-\frac{\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}+1}}{1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}}\frac{\partial\mu_i}{\partial\beta_j}\right\}=0$$

Now, $\mu_i = e^{\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}}$. So, $\frac{\partial\mu_i}{\partial\beta_j}=\mu_i x_{ij}$.

$$\Rightarrow \sum_{i=1}^{n}\left\{\frac{y_i-\mu_i}{1+\alpha\mu_i}-\frac{\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}+1}}{1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}}\mu_i\right\}x_{ij}=0 \qquad \text{----------(ii)}$$

Again, the sets of equations (i) and (ii) are not in closed form. We solve them numerically using Newton-Raphson iterative method to obtain the MLE's of α and $\underline{\beta}$.

The regression equations obtained are as follows:

Count model (truncated Negative Binomial with log link):

$$\text{Ln}\ \mu_i = 0.137493 + 1.070967\ x_{i1} + 0.013670\ x_{i2}$$

Zero hurdle model (Binomial with logit link):

$$ln\frac{\pi_i}{(1-\pi_i)} = 0.967776 + 1.262903\ x_{i1} - 0.008290\ x_{i2}$$

And, $\hat{\alpha} = 0.9587$

We also find the predicted no. of zeros to be 360, which happens to be the no. of zeros in the observed data.

Therefore, this model is able to handle excess zeros and overdispersion, and so we may conclude logically that this model satisfactorily regresses the data, better than the previous models.

Table4: Hurdle-at-zero Negative Binomial output

Count model coefficients (truncated negbin with log link)				
Predictors	**Estimate**	**Std. Errors**	**z value**	**p value**
Intercept	0.137493	0.170393	0.807	0.419712
badh	1.070967	0.124662	8.591	< 2e-16
age	0.013670	0.004075	3.354	0.000796
Log(alpha)	-0.042212	0.158032	-0.267	0.789383
Zero – hurdle model coefficients (binomial with logit link)				
Predictors	**Estimate**	**Std. Errors**	**z value**	**p value**
Intercept	0.967776	0.232376	4.165	3.12e-05
badh	1.262903	0.288887	4.372	1.23e-05
age	-0.008290	0.006049	-1.371	0.171

Chapter 9

9. Estimation of regression parameters

We have fitted the variable count(y) on two predictors namely, badh(x1), age(x2). There are total n = 1127 observations in our hand. We have not included any intercept in the regression model. The table given below gives the rough estimate of the parameters obtained by the method of maximum likelihood using Newton-Raphson numerical analysis procedure.

Table5: Estimation of Regression Parameters

MODEL	PREDICTORS	NOTATIONS	ESTIMATED VALUES
Hurdle-at-Zero Poisson	badh	(β_1,δ_1)	(0.876388,1.262903)
	age	(β_2,δ_2)	(0.008999,-0.008290)
Hurdle-at-Zero Negative Binomial	badh	(β_1,δ_1)	(1.070967, 1.262903)
	age	(β_2,δ_2)	(0.013670, -0.008290)

We now consider a statistical model comparing measure Akaike Information Criterion (AIC) to test for the goodness of fit of all the models, compare them and reach to a theoretically justified decision.

Chapter 10

10. Models selection / Tests to compare Models

In this section, we will briefly introduce the tests carried out to compare the goodness of fit of the two fitted models.

10.1 Akaike Information Criterion(AIC)

The Akaike information criterion (AIC), named after Hirotugu Akaike, is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

AIC is founded on information theory. When a statistical model is used to represent the process that generated the data, the representation will almost never be exact; so some information will be lost by using the model to represent the process. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

Suppose we have a statistical model of some data. Let k be the number of estimated parameters in the model. Let L be the maximum value of the likelihood function for the model. Then the AIC value of the model is given by,

$$\text{AIC} = 2k - 2\ln\text{L}$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood

function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, because increasing the number of parameters in the model almost always improves the goodness of the fit.

10.2 Goodness of fit

From the above dispersion test it can surely be said that there is overdispersion in the dataset. So the idea of fitting a Poisson model to the above dataset completely rules out since in the Poisson distribution the assumption of equi dispersion is followed. With this we may say that fitting a negative binomial model shall be better but since the dataset is zero inflated, simple negative binomial model captures overdispersion well but it fails to capture the zero inflation in the dataset.

From the above graph, we can clearly observe the property of zero inflation in the dataset since zeros are occurring at a higher rate than the other observations.

The above statements can also be verified from the Table6 given below

MODELS	LOG LIKELIHOOD VALUES	AIC VALUES
POISSON MODEL	-2817.776	5641.552
NEG BINOMIAL MODEL	-2233.6425	4475.285
HURDLE-AT-ZERO POISSON MODEL	-2549	5110
HURDLE-AT-ZERO NB MODEL	-2229	4472

Now, as discussed earlier, the preferred model for fitting is the one with the minimum AIC value. Thus from the above table, we can conclude that HURDLE-AT-ZERO POISSON MODEL is the best fitted model for the above dataset.

Chapter 11

11. Wald Z test for Significance of the coefficient for the Hurdle Negative Binomial Model :

We know for the above dataset Hurdle-at-zero Negative Binomial model has the lowest AIC value and thus fits better than the other models. Now, we want to test whether the seven predictors that we have used in the above model are significant or not . Wald Z test is a parametric test which is a way to find out if explanatory variables in a model are significant.

Test procedure:

Let θ denote a parameter in the model. The maximum likelihood estimate of θ on the basis of the sample be denoted by $\hat{\theta}$.

Then, $\hat{\theta} \sim N\ (\theta, V(\hat{\theta}))$, where $V(\hat{\theta})$ is the variance of $\hat{\theta}$.

To test: - $H_0 : \theta = 0\ against\ H_1 : \theta \neq 0$

An appropriate test statistic is given by: T = $\frac{\hat{\theta}-0}{se(\hat{\theta})} \overset{a}{\sim} N(0,1)$ under H_0 , where, se $(\hat{\theta})$ is the standard error of $\hat{\theta}$. A reasonable estimate of the se $(\hat{\theta})$ can be given by $\frac{1}{\sqrt{I_n(\hat{\theta})}}$, where I_n is the Fisher information of the parameter θ evaluated at $\hat{\theta}$.

We reject H_0 against H_1 at α level of significance iff $|T_{obs}| > \tau_{\alpha/2}$, where, $\tau_{\alpha/2}$ is the upper $\frac{\alpha}{2}$% point of N (0, 1) distribution and T_{obs} is an observed value of T.

In our test we consider, $\alpha = 0.05$

The table7 given below helps us to decide whether we reject or accept H_0 and make conclusion about the significance of the predictors in the fitted regression model.

Count model coefficients (negbin with log link)						
Coefficients	**Estimate**	**Std. Errors**	**T-STAT**	**Critical value $\tau_{0.025}$**	**p value**	**Decision**
intercept	0.137493	0.170393	0.807	1.9599	0.419712	Accept
badh	1.070967	0.124662	8.591		< 2e-16	Reject
age	0.013670	0.004075	3.354		0.000796	Reject
Log(α)	-0.042212	0.158032	-0.267		0.789383	Accept
Zero-inflation model coefficients (binomial with logit link)						
Predictors	**Estimate**	**Std. Errors**	**T-STAT**	**Critical value $\tau_{0.025}$**	**p value**	**Decision**
intercept	0.967776	0.232376	4.165	1.9599	3.12e-05	Reject
badh	1.262903	0.288887	4.372		1.23e-05	Reject
age	-0.008290	0.006049	-1.371		0.171	Accept

Thus we conclude that in the Hurdle Negative Binomial model coefficients (neg bin with log link) H_0 is rejected for **badh** and **age**. For the Zero inflation model coefficients (binomial with logit link) H_0 is rejected for all the predictors except the variable **age.**

Chapter 12

12. Conclusion

This project consists of eleven chapters which discusses about modelling of count data. Count data is encountered on daily basis and dealings. A lot of data is in the form of counts. Further, with count data, the number 0 often appears as a value of response variable

In this project, the Poisson model, Hurdle at zero Poisson model, Negative Binomial model, Hurdle at zero Negative Binomial model have been applied to model the number of visits of the patients to the doctors. Number of visit to doctor deals with over dispersion and zero-inflation problems.

We used the AIC criterion to select the best model for modelling the count data. The result shows that Hurdle at zero Negative Binomial have better performance than Poisson model and Hurdle at zero Poisson for the count data.

In the future, we may use some variable selection techniques to build the best model including only significant predictors. For example, we can use stepwise selection procedure to eliminate all the insignificant predictors from the current models. We can also try the subset selection procedures to select significant predictors in all the models or use the lasso (least absolute shrinkage and selection operator)-based techniques for variable selection in Poisson model and ZIP models. With the variable selection procedures, not only we can find all the statistical significant predictors to the response variable, we may also remove possible collinearity problem among the predictors.

Chapter 13

13. Appendix

```
rm(list=ls())
library(msme)
library(MASS)
library(lattice)
library(sandwich)
library(COUNT)
data(badhealth)
d=badhealth;d
attach(d)

#graph
numvisit
table(numvisit)
plot(table(numvisit),main="GRAPH OF THE DATASET",xlab="NUMBER OF VISIT TO
DOCTOR",ylab="COUNTS OF VISITS")

#Poisson Regression
modelp=glm(numvisit~badh+age,family=poisson(link="log"));modelp
summary(modelp)

#predict number of zeros for Poisson
sum(numvisit==0)
preds=predict(modelp,type="response");preds
prop=dpois(0,preds);prop
round(sum(prop))
```

```
#Testing for zero inflation
library(vcdExtra)
zero.test(numvisit)

#hurdle at zero poisson model
library(pscl)
hudp=hurdle(numvisit~badh+age,family=poisson(link="log"));hudp
hurdle(numvisit~badh+age)
summary(hudp)

#Over dispersion
mean=apply(d,2,mean);mean
var=apply(d,2,var);var
library(Rfast2)
overdispreg.test(numvisit,matrix(c(badh,age),ncol=2))

#Negative Binomial Regression
modelnb=glm.nb(numvisit~badh+age,data=badhealth);modelnb
summary(modelnb)

#Predict number of zeros for Negative Binomial
pred=predict(modelnb,type="response");pred #estimated means
esttheta=summary(modelnb)$theta;esttheta
prob=esttheta/(esttheta+pred)
prop0=dnbinom(0,esttheta,prob);prop0
round(sum(prop0))
```

```
#hurdle at zero Negative Binomial model
hudnb=hurdle(numvisit~badh+age,data=badhealth,dist="negbin");hudnb
summary(hudnb)

#AIC for poisson
extractAIC(modelp)
#AIC for negative binomial
extractAIC(modelnb)
#AIC for hurdle Poisson model
extractAIC(hudp)
#AIC for hurdle Negative Binomial model
extractAIC(hudnb)
```

Chapter 14

14. ACKNOWLEDGEMENT

First, there are no words to adequately acknowledge the wonderful grace that my Redeemer has given me. There are many individuals who have come together to make this project a reality. I greatly appreciate the inspiration; support and guidance of all those people who have been instrumental for making this project a success.

I express my deepest thanks to my guide Prof. Madhura Das Gupta , Department of Statistics, St. Xavier's College (Autonomous) who guided me faithfully through this entire project. I have learned so much from her, both in the subject and otherwise. Without her advice, support and guidance, it find difficult to complete this work.

Lastly, I would like to extend my gratitude to St. Xavier's College (Autonomous) for the opportunity to present a dissertation project paper on a topic of my choice. I would also like to thank them for helping me to develop a research mindset in me.

Chapter15

15. Bibliography

1. Generalized Linear Models, Second Edition, P. McCullagh, Department of Statistics, University of Chicago and J.A. NELDER, Department of Mathematics, Imperial College of Science and Technology, London

2. An Introduction to Categorical Data Analysis, Second Edition, ALAN AGRESTI, Department of Statistics University of Florida, Gainesville, Florida.

3. SCORE TESTS FOR ZERO-INFLATION AND OVER-DISPERSION IN GENERALIZED LINEAR MODELS, Dianliang Deng and Sudhir R. Paul University of Regina and University of Windsor

4. Hurdle model - Wikipedia en.wikipedia.org (Visit Date: April2, 2023)

5. Count data - https://en.wikipedia.org/ (Visit Date: April2, 2023)

6. Overdispersion - Wikipedia en.wikipedia.org (Visit Date : April1, 2023)

7. Zero-inflated model - Wikipediaen.wikipedia.org (Visit Date : April1, 2023)

8. https://www.youtube.com/watch?v=T6cv8n9xBGQ (Visit Date: March31, 2023)