Comparison of Statistical Models in Modeling Over-Dispersed Count Data with Excess Zeros

Ndwiga Antony Macharia[1], Oscar Ngesa[2], Anthony Wanjoya[3], Damaris Felistus Mulwa[4]

[1, 2 ,3]*Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya.*

***Abstract*:-Generalised Linear Models such as Poisson and Negative Binomial models have been routinely used to model count data. But, these models assumptions are violated when the data exhibits over-dispersion and zero-inflation. Over-dispersion is as a result of excess zeros in the data. For modelling data with such characteristics several extensions of Negative Binomial and Poisson models have been proposed, such as zero-inflated and Hurdles models. Our study focus is on identifying the most statistically fit model(s) which can be adopted in presence of over-dispersion and excess zeros in the count data. We simulate data-sets at varying proportions of zeros and varying proportions of dispersion then fit the data to a Poisson, Negative Binomial, Zero-inflated Poisson, Zero-inflated Negative Binomial, Hurdles Poisson and Negative Binomial Hurdles. Model selection is based on AIC, log-likelihood, Vuong statistics and Box-plots. The results obtained, suggest that Negative Binomial Hurdles performed well in most scenarios compared to other models hence, the most statistically fit model for over-dispersed count data with excess zeros.**

Keywords: **Zero-inflated models, Hurdles models, Over-dispersion, Excess zeros, Simulation, Zero-inflation, Vuong test**

I. INTRODUCTION

Over-dispersed count data with zero inflation is becoming so common in recent researches. Such data is encountered in a wide range of places.[1], gave areas in which a scientist can expect such kind of data; medicine, recreational facilities, econometric data....agricultural data, or econometrics data. Especially when the event of interest is rare. For instance, the number of car accidents per day, reported cases of infectious diseases, number of absentees in a school, or number of crime cases. This kind of data may have a variety of characteristics that a scientist needs to take account of in the modelling phase. Normal distribution, cannot be adopted in modelling this kind of data, since, the disciplines of OLS (Ordinary Least-square) regression models are not complied. That is, normality, linearity and homoscedastcity.

Instead, Generalized Linear Models (GLM) may be adopted. [2], used the Poisson regression model to analyse count data.[3],[4], showed that negative binomial (NB) model can be used when data has over-dispersion caused by heterogeneity of data. Although, they cannot account for zero inflation in the data, since, for these models to be used the zero proportion must be necessarily linked to the distribution of positive counts. For instance, Poisson model assumes equi-dispersion; mean and variance should be equal, otherwise the model will be violated, [5]. In real data, equi-dispersion is not commonly reflected. In most cases, variance is greater than the mean. This condition of Poisson variation is termed as over-dispersion. The negative binomial distribution is used for the data which is over-dispersed. The NB distribution has natural parameter which has an effect on relaxing the mean-variance relationship, and itis assumed to follow a gamma distribution. In a case where variance is smaller compared to the mean, the data is under-dispersed. The variance of a Poisson model is, $v(\mu) = \mu$) which shows that the variance is a deterministic function of the mean. There are several instances that can lead to violation of the equi-dispersion assumption namely; 1) When the data is hierarchically structured, this can be encountered in longitudinal studies. 2) The occurrence of over-dispersion, that is, variance of the data is not equal to the mean, which is a necessary condition for a Poisson distribution. 3) When there is zero inflation relative to Poisson model assumption, the negative binomial regression model maybe adopted which may improve the fit of the data. Presence of excess zeros in the data. One explanation for this, it can be assumed the sample is from two different sub-populations; one where the outcome of interest is always zero and the other behaves like a Poisson distribution. Several extensions of the existing models have been proposed to model zero inflation in count data. [6], proposed the zero-inflated count models. Zero-inflation, shows that the count data set has excessive number of zeros. The word inflation, emphasizes that the probability mass has a higher peak at zero which exceeds the levels allowed under a standard parametric family in the discrete distribution. Zero-inflated models tend to assume count datacomes from two different populations. First, non-exposed group that "never" experiences the event, and the exposed group where the events are generated byuse of a standard model. There are two sources of zeros in such scenario, one is assumed to come from the exposed population and the other from the non-exposed population. Zeros from the exposed population are modelled using a Poisson or Negative Binomial,[7][8]. It is important to note that, Zero-inflated Poisson models can be adopted for the data with large number zeros counts but cannot adequately account for over-dispersed data. Thus there was a need for a model that would cater for both zero inflation and over-dispersion. [9], proposed an extension of NB model, zero-inflated negative binomial (ZINB) regression model as an alternative to ZIP model.

[10],proposed the hurdle models for analysis of count data with both excess zeros and over-dispersion. The model is

composed of two stages hence the name "two-part" model. The main assumption is all the zeros are sampling zeros. The hurdles model can be grouped into two, Poisson hurdle (PH) and negative binomial hurdle (NBH). The first part is the binary response which determine whether the response has zero or non-zero outcome. Binary response is modelled using a binary model for instance logit, probit or complementary log-log. The second part, analysis the positive counts truncated-at-zero, where Poisson, geometric and negative binomial are used, thus estimating two equations. Some simulation studies have been performed to compare the model performance for zero-inflated and over-dispersed counts. [11], compared parameter estimations between Poisson hurdle and zero inflated Poisson; [12], compared Poisson, PH and ZIP varying degree of zero-inflation; [13], compared zero inflated negative binomial with negative binomial hurdle. However, the comparisons available in these studies are limited hence a comprehensive comparison of one part models (Poisson and NB), zero-inflated and hurdles models for over-dispersed and zero-inflated count data is desired.

In this paper a comprehensive comparison of the following models; one part models (Poisson, Negative Binomial), mixed models (Zero inflated models) and two part models (Hurdles models). Through data simulation, different scenarios of zero inflation and over-dispersion were examined.

II. METHODOLOGY

Statistical Models

In this study, a simulation technique in R to generate the simulated data set that was used to compare the models. Simulation mainly focused on:

existence of structural zeros in the data. In the structural and zero-component there are two binary covariate. The covariates $X_1$and X_1 are defined as indicators of the exposed and unexposed group respectively and the chance of an outcome originating from the exposed group is given as 0.5. Data has been simulated under the following models;

2.1 One-part models

2.1.1 Poisson model

Poisson model assumes equi-dispersion(mean=variance) a property that is hard to attain in real life situations. If X follows a Poisson distribution, then the probability distribution function (pdf);

$$\boldsymbol{f(X_i)} = \frac{e^{-\mu}\mu^{x}}{x!}, \quad y = 0,1,2,.... \qquad (1)$$

With mean and variance as; $E\{x\}$ =Var{x}= μ (implying equi-dispersion).

The case model Poisson regression, assumes, $x_1, \dots x_n$ are a realization of independent random variables $X_1, \dots X_n$following the distribution:

$$Y_t \sim Pois(\lambda_i) \qquad (2)$$

Define a link function g relating x to a linear predictor can be expressed as;

$$\begin{aligned} \mathrm{g}(\mu_i) &= n_i \\ &= \beta_0 + \beta_1 X_i + \beta_2 X_{2i} \\ &= x'\beta \end{aligned}$$

With β as the vector to be estimated, n_i the linear predictor and μ_i as the mean of the distribution function. The link function of the Poisson model can be given as;

$$n_i = \log(\mu_i)$$

The Poisson model has only one parameter μ_i and can be estimated by use of the Newton-Raphson Algorithm.

2.1.2 Negative Binomial model

NB is an extension of Poisson model and can be used when data is over-dispersed. It can handle over-dispersion due to the additional parameter that account for variability in the data. Let assume that a random variable $w_i$has a mean μ_i where I is a positive integer. The parameter μ_i also depends the heterogeneity component ϵ (error term). The NB of w_i can be expressed as;

$$P(w_i) = \frac{\Gamma\Omega+\mathrm{w}}{\Gamma(\Omega)+\mathrm{w!}}\left\{\frac{\beta}{1+\beta}\right\}^{w_i}\left\{\frac{1}{+\beta}\right\}^{\Omega}, \qquad (3)$$
$$w_i = 1, 2, \dots, \mathrm{n}$$

We can expressed mean as; $E(w_i) = \Omega\beta$ and variance $Var(w_i) = \Omega\beta + \Omega\beta^2$. The parameters $\mu = \Omega\beta$ and $\mathrm{k} = \frac{1}{\Omega}$ are taken as the expected value and the dispersion parameter when building a NB model. Whereby, $E(w_i) = \mu$ and $Var(w_i) = \mu + \mathrm{k}\mu^2$ can be expressed as the log link function of NB mode as;

$$\mathrm{logit}(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

With p covariates and $\beta_0, \beta_1, \dots, \beta_p$ as the regression coefficient. The parameters of the model ϕ and β are estimated using the Fisher's scoring Algorithm.

2.2. Mixed models

2.2.1 Zero-Inflated models

For the case model for the ZIP and ZINB they assume that $y_1, \dots y_n$are a realization of independent random variables $Y_1, \dots Y_n$following the distribution:

$$Y_t \sim ZIP(p_i, \mu_i), \qquad i = 1, 2, \dots, n$$
$$Y_t \sim ZINB(p_i, \lambda_i, \Omega^{-1}), \qquad i = 1, 2, \dots, n$$

The probability mass function of ZIP can expressed as;

$$Y_i = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} & ,when\ y_{i=0} \\ (1-p_i)\dfrac{e^{-\lambda_i}{\lambda_i}^{x}}{x!}, & when\ y_{i>0} \end{cases}$$

Where p_i can for ZIP can be modelled using a log model as;

$$Log(pi) = \frac{p_i}{(1-p_i)} = \beta_0 + \beta_1 X_i + \beta_2 X_{2i} + \cdots + \beta_p X_p$$

$$\log(\lambda_0) = \gamma 0 + \gamma 1 X1i + \gamma 2 X2i$$

2.2.2 Zero inflated Negative Binomial

Probability mass function of ZINB can be expressed as;

$$Y_i = \begin{cases} p_i + (1-p_i)(1+\Omega\lambda_i)^{-\Omega^{-1}} & ;when\ y_{i=0} \\ (1-p_i)\dfrac{\Gamma(\Omega^{-1}+\mathrm{w})(\Omega\lambda_i)^{y_i}}{\Gamma(y_i+1)\Gamma(\Omega^{-1})(1+\Omega\lambda_i)^{y_i+\Omega^{-1}}} & ;when\ y_{i>0} \end{cases}$$

Wherep_i can be modelled using a logit model according to[6];

$$Logit(pi) = \frac{p_i}{(1-p_i)} = \beta_0 + \beta_1 X_i + \beta_2 X_{2i} + \cdots + \beta_p X_p$$

$$\mathrm{logit}(\lambda_0) = \gamma 0 + \gamma 1 X1i + \gamma 2 X2i$$

Where,
pias the zero inflation factor or the probability of structural zeros in the data.

λ_i Expected value which can be expressed as; $\lambda_i = \exp(\alpha' X_i)$where α, is a $(m+1)*1$vector of unknown zero-inflated coefficient to be estimated to be estimated associated with, X_i (known covariates) and number of covariates in the model.
Ω^{-1}, as the dispersion parameter.

X_{ji}, β_{ji}, as the binary covariates for the simulation conditions. Parameters $\gamma 0, \gamma 1$and $\gamma 2$are collectively referred to as γwhile $\beta_0$and $\beta_1$as; βn sample size. If $pi$is 1, the outcome of interest Yiwill be set to be zero, for $pi$is zero, Yi will be simulated from either a Poisson or NB distribution for ZIP and ZINB distribution respectively.

2.3 Two part models

For the NBH and PHmodel assumes that $y_1, \ldots y_n$are a realization of independent random variables $Y1, \ldots Yn$ following the distribution;

$$Yi \sim \mathrm{PH}(\pi i, \lambda_i),\ i = 1, \ldots, n$$

$$Yi \sim \mathrm{NBH}\ (\pi i, \lambda_i, \Omega i),\ i = 1, \ldots, n$$

The probability mass function of the PH can be expressed as;

$$Y_i = \begin{cases} \pi i & ,when\ y_{i=0} \\ (1-\pi i)\dfrac{e^{-\lambda_i}{\lambda_i}^{\lambda_i}}{(1-e^{-\lambda_i})y!}, & when\ y_{i>0} \end{cases}$$

Where πi can for ZIP can be modelled using a log model as;

$$Log(\pi i) = \frac{\pi_i}{(1-\pi_i)} = \beta_0 + \beta_1 X_i + \beta_2 X_{2i} + \cdots + \beta_p X_p$$

$$\log(\lambda_i) = \gamma 0 + \gamma 1 X1i + \gamma 2 X2i$$

Probability mass function of HNB can be expressed as;

$$Y_i = \begin{cases} \pi i & ;when\ y_{i=0} \\ (1-p_i)\dfrac{\Gamma(\Omega^{-1}+y_i)(1+\Omega\lambda_i)^{-\Omega^{-1}}(\Omega\lambda_i)^{y_i}}{\Gamma(y_i+1)\Gamma(\Omega^{-1})} & ;when\ y_{i>0} \end{cases}$$
$$1-(1+\Omega\lambda_i)^{-\Omega^{-1}}$$

Where πi can be modelled using a logit model as;

$$Logit(\pi i) = \frac{\pi_i}{(1-\pi_i)} = \beta_0 + \beta_1 X_i + \beta_2 X_{2i} + \cdots + \beta_p X_p$$

$$logit = \gamma 0 + \gamma 1 X1i + \gamma 2 X2i$$

With $\pi_i$denotes the chance of zero count in the model and $\mu_i$denotes the expected value of left-truncated at zero negative binomial model.

Where,
$\pi_i$as the zero inflation factor or the probability of structural zeros in the data.

λ_i Expected value which can be expressed as; $\lambda_i = \exp(\mathcal{g}' X_i)$ where $\mathcal{g}$, is a $(m+1)*1$vector of unknown parameters to be estimated associated with X_i (known covariates) and mnumber of covariates.

$X_{ji}\ and\ \beta_{ji}$, as the binary covariates for the simulation conditions.

Parameters $\gamma 0, \gamma 1$and $\gamma 2$are collectively referred to as γ while $\beta_0$and$\beta_1$as; βn sample size. If $pi$is 1, the outcome of interest Yiwill be set to be zero, for pi is zero, Yiwill be simulated from either a Poisson or NB distribution for ZIP and ZINB distribution respectively

2.4. Simulation design

In my simulation the performance of both one-part models, zero-inflated models and hurdles has been compared given the simulation conditions. The zero proportion/structural zeros varies as shown (0.1, 0.2, 0.3, 0.5, 0.7 and 0.9), for instance 0.2 denotes 20%. Dataset has been generated from: Zero-Inflated (Poisson and NB) and Hurdles (Poisson and NB).For the over-dispersion in the non-zero part, we set the dispersion parameter(Ω) with the following values 1,5,10,15,20,30 and 50 are used. The larger the value of Ω the less dispersed the variable. We used NB distribution to simulate the response variable with varying zero proportion. Then for each data set simulated above using the four models, we run all the models including Poisson and NB model and compare the model fit.

2.5 Generating Simulated datasets

In order to attain a realistic prediction model, the simulation has focused on a scenario where structural zeros are present in the data. Let Y be our response variable, and two types of covariates, $X_1$and X_2have been simulated in theprediction model. Assume the covariateX_1, is a binary variable taking the values0 and 1 whose$\Pr(x1 = 0) = Pr(X1 = 1) = 0.5$ (This implies that X_1 can be denoted as the indicator of the susceptible group and the probability of an individual coming from the this group is 50%) and let set $X_2$to follow a standard normaldistribution, with mean 0 and standard deviation of 1.The regression coefficient β_0, $\beta_1$and $\beta_2$for the two covariates are set to be 1, 0.3and 0.5 respectively so that our model can allow for a medium and large value . The above covariates $\beta_1$and $\beta_2$are seen as realistic choices for comparisonbetween different scenarios of prediction for the two covariates. For us to increase the accuracy of the results, a simulation size of2000 and each a sample size n=500, was generated. The decision to use those number of simulation and sample size is based on previous research according to[6],[11].

III. RESULTS

Table 1, reflects on the AIC results from the simulation scenario varying the level of zero-proportions and levels of over-dispersion in the non-zero part. The AIC values from the table, show that an increase in over-dispersion from 1 to50 keeping the zero proportion constant leads to improvement of the models efficiency since the values of the AIC are getting smaller. The Poisson model produced the largest values of AIC in all simulation conditions proving to be the worst fit for the data followed by the NB model.

Table 2, gives similar results to those of table one, on the log-likelihood ratiosof the 6 models at varying conditions of zero inflation and over-dispersion. The log-likelihood values reduced gradually as the level of over-dispersion increased keeping zero proportion constant as it evident in all scenarios. Table 3, 4 and 5 gives results on the Vuong test at varying levels of zero-inflation and at constant level of over-dispersion 10, 20 and 50. When the v-statistic >1.64,then model 1 is preferred and vice-versa, the p-value is significance at 0.05 level. When the V-statistic is less than 1.96 but the p-vale not the significant, the two models are assumed to be equal in preference. The tables gave varying preference of models as the levels of zero inflation and over-dispersion changed. TheZINB and NBH models were more superior to the other models at various level of dispersion as it can be seen in the Vuong tests results. Figures 1, 2, 3, 4 and 5 shows the box plots at 0.5 zero-proportion at various levels of over-dispersion. It is evident that Poisson model and NB models performed poorly compared to other four models. There are Box-plots at all levels of zero proportions and dispersion.

Figure 3.1 shows some of the frequency plots of our variable on varying levels at 0.5 zero inflation and different levels of over-dispersion. It is evident that there is gradual change from one scenario to another. There is evidence of large mass at zero inall the frequency plots. There are frequency plots at all levels of zero proportion and dispersion.

Table 1: AIC results varying levels of zero proportions and over-dispersion.

Zero-proportion	Dispersion levels	Poisson	NB	ZIP	ZINB	PH	NBH
W=0.1	1	1854.868	1603.067	1592.335	1505.749	1592.391	1505.422
	5	1644.402	1530.117	1422.833	1424.654	1423.094	1424.96
	10	1592.497	1486.751	1388.915	1390.182	1389.365	1390.684
	15	1541.379	1438.436	1334.51	1336.511	1333.628	1335.628
	20	1509.406	1378.059	1319.739	1313.513	1315.45	1314.54
	30	1404.546	1398.892	1299.981	1301.709	1300.573	1302.385
	50	1403.821	1370.149	1261.271	1263.272	1261.85	1263.851
w=0.2	1	1735.897	1558.389	1479.043	1462.716	1478.91	1461.172
	5	1509.747	1491.955	1443.811	1436.62	1436.692	1429.774
	10	1340.724	1351.705	1291.557	1291.848	1291.15	1291.441
	15	1285.964	1292.037	1270.764	1272.047	1270.66	1271.973
	20	1252.267	1219.02	1208.592	1210.538	1209.462	1211.363
	30	1288.748	1188.748	1158.722	1160.695	1160.32	1162.287
	50	1270.455	1189.921	1184.669	1186.67	1184.409	1186.411
W=0.3	1	1514.777	1314.146	1274.214	1215.931	1275.369	1216.718
	5	1304.229	1297.528	1232.601	1217.819	1233.004	1217.346
	10	1252.787	1196.064	1187.406	1186.217	1188.342	1189.11

	15	1181.757	1147.27	1133.663	1134.886	1139.418	1141.221
	20	1136.62	1156.82	1137.703	1136.278	1138.788	1137.295
	30	1160.647	1213.7	1124.216	1126.101	1126.774	1128.682
	50	1140.671	1037.154	1011.507	1013.508	1011.246	1013.246
W=0.5	1	1217.553	1111.0255	1099.203	1095.1257	1098.913	1094.8461
	5	1197.883	1059.837	1044.712	1045.915	1045.57	1046.795
	10	1145.647	1028.498	1001.112	1001.738	1005.421	1005.857
	15	948.4479	883.2292	873.8491	875.8429	874.7466	876.7379
	20	916.8813	872.0255	765.4092	762.6073	766.0089	763.2086
	30	909.3163	817.3055	756.1152	758.1052	761.113	758.1214
	50	927.5441	814.9079	718.5312	719.5813	717.1642	717.1843
w=0.7	1	974.9643	759.8688	785.1444	789.5019	805.4872	789.8835
	5	947.4684	790.6423	785.8947	785.2784	786.6508	786.1283
	10	862.3864	738.968	712.8576	714.858	716.2471	718.2474
	15	723.5035	634.3374	630.8808	632.6487	631.8775	633.6412
	20	782.5832	608.7846	604.0946	602.4004	605.0406	603.3566
	30	756.7285	617.7997	596.1011	598.0993	596.7915	598.7893
	50	735.189	602.9745	588.9496	587.9499	588.7727	587.773
w=0.9	1	550.0461	343.6635	337.7279	318.2852	330.731	318.3424
	5	476.972	343.131	340.5723	294.377	311.7136	312.7379
	10	435.8858	313.5291	278.3307	278.8959	312.9879	311.988
	15	367.7192	277.9089	275.8271	277.8272	275.3203	277.3205
	20	371.5652	227.7987	218.6254	220.4399	220.7392	218.548
	30	385.1982	240.3517	223.7344	224.3247	221.5789	222.7416
	50	289.0465	238.0542	200.1299	202.1303	200.2184	200.2185

Table 2: Log-likelihood results varying levels of zero proportions (zero-%) and over-dispersion(D).

Zero %	D	Poisson	NB	ZIP	ZINB	PH	NBH
W=0.1	1	-916.639	-785.584	-734.832	-710.253	-734.845	-710.075
	5	-838.712	-774.114	-721.373	-720.411	-721.09	-720.162
	10	-833.249	-769.376	-688.457	-678.091	-688.683	-678.342
	15	-797.69	-685.218	-661.255	-661.256	-660.814	-660.814
	20	-791.703	-685.03	-658.869	-655.256	-700.725	-655.27
	30	-789.273	-675.446	-653.991	-653.854	-694.287	-654.193
	50	-683.911	-671.075	-647.636	-644.636	-647.925	-644.926
w=0.2	1	-884.537	-831.987	-773.799	-730.426	-773.84	-730.618
	5	-751.874	-816.978	-711.31	-712.346	-707.887	-707.887
	10	-667.362	-746.853	-639.778	-638.924	-639.575	-638.72
	15	-679.982	-732.019	-629.382	-629.237	-629.33	-628.986
	20	-683.134	-655.511	-628.296	-628.269	-628.731	-628.682
	30	-681.374	-651.073	-633.361	-633.348	-634.16	-634.143
	50	-702.228	-645.961	-636.334	-636.335	-636.205	-636.205
W=0.3	1	-754.389	-653.073	-631.107	-600.965	-631.685	-601.359

	5	-649.115	-634.764	-610.301	-601.91	-610.502	-601.673
	10	-623.394	-594.032	-588.703	-588.109	-588.171	-587.555
	15	-587.878	-569.635	-560.832	-560.443	-563.709	-563.611
	20	-565.31	-524.41	-517.851	-516.139	-518.394	-516.647
	30	-527.323	-502.85	-491.108	-491.051	-492.387	-492.341
	50	-482.336	-464.577	-449.754	-449.754	-449.623	-449.623
w=0.5	1	-735.776	-651.513	-543.602	-540.563	-543.457	-540.423
	5	-595.942	-525.919	-516.356	-515.958	-516.785	-516.397
	10	-569.824	-510.249	-494.556	-493.869	-496.711	-495.928
	15	-521.224	-487.615	-480.925	-480.921	-481.373	-481.369
	20	-505.407	-432.028	-426.746	-424.336	-427.044	-424.643
	30	-511.657	-454.677	-437.076	-437.076	-439.557	-439.557
	50	-520.77	-473.49	-402.756	-402.757	-402.571	-402.571
w=0.7	1	-484.482	-475.934	-386.572	-387.751	-396.744	-387.942
	5	-470.734	-391.321	-386.947	-385.639	-387.325	-386.064
	10	-428.193	-365.484	-350.429	-350.429	-352.124	-352.124
	15	-358.752	-313.169	-309.44	-309.324	-309.939	-309.821
	20	-488.292	-400.392	-396.047	-394.2	-396.52	-394.678
	30	-475.364	-404.9	-392.051	-392.05	-392.396	-392.395
	50	414.3759	-342.487	-321.975	-321.975	-321.886	-321.887
w=0.9	1	-372.023	-267.832	-262.864	-252.143	-259.366	-252.171
	5	-335.486	-267.566	-264.286	-240.189	-249.857	-249.369
	10	-314.943	-252.765	-250.165	-232.448	-232.494	-232.494
	15	-280.86	-234.955	-231.914	-231.914	-231.66	-231.66
	20	-292.783	-209.899	-203.313	-203.22	-203.37	-203.274
	30	-239.599	-166.176	-160.867	-125.163	-124.79	-124.371
	50	-266.504	-190.027	-179.065	-179.065	-179.109	-179.109

Table 3 (a): Vuong tests results varying levels of zero proportions at 10 level of over-dispersion

Zero proportions	models	V-statistic	P-value	preference
W=0.3	ZIP vs. PH	0.1545689	0.43858	ZIP=PH
	PH vs. ZINB	0.5894296	0.027779	ZINB
	ZIP VS NBH	0.444017	0.025966	NBH
	ZINB VS NBH	0.2094353	0.041705	NBH
	ZIP VS ZINB	-0.7098647	0.023889	ZINB
W=0.5	ZIP VS PH	1.37026	0.85303	ZIP=PH
	ZINB VS PH	1.96052	0.058551	ZINB
	ZIP VS NBH	0.6488366	0.025822	NBH
	ZINB VS NBH	1.343741	0.089516	NBH
	ZIP VS ZINB	-0.588163	0.027821	ZINB
W=0.7	ZIP VS PH	1.613822	0.53283	ZIP=PH
	ZINB VS PH	1.961359	0.053307	ZINB
	ZIP VS NBH	1.613936	0.053271	NBH

	ZINB VS NBH	1.613713	0.053295	NBH
	ZIP VS ZINB	-0.05606	0.014547	ZINB
W=0.9	ZIP VS PH	0.120384	0.44123	ZIP=PH
	ZINB VS PH	1.9814695	0.004569	ZINB
	ZIP VS NBH	0.1216431	0.045156	NBH
	ZINB VS NBH	0.1206431	0.045151	NBH
	ZIP VS ZINB	-0.10324	0.043168	ZINB

Table 3: Vuong test results varying levels of zero proportions at 20 level of over-dispersion.

Zero proportions	models	V-statistic	P-value	preference
W=0.3	ZIP VS PH	0.52544	0.29964	ZIP=PH
	ZINB VS PH	1.99195	0.015435	ZINB
	ZIP VS NBH	-0.565	0.028604	NBH
	ZINB VS NBH	0.473359	0.031798	ZINB
	ZIP VS ZINB	-0.90392	0.18302	ZINB
W=0.5	ZIP VS PH	0.354179	0.3616	ZIP=PH
	ZINB VS PH	1.964318	0.013644	ZINB
	ZIP VS NBH	-0.87501	0.019078	NBH
	ZINB VS NBH	0.368767	0.035615	NBH
	ZIP VS ZINB	-1.05176	0.14645	ZINB
W=0.7	ZIP VS PH	0.774216	0.2194	ZIP=PH
	ZINB VS PH	2.168115	0.015462	ZINB
	ZIP VS NBH	-0.67349	0.025032	NBH
	ZINB VS NBH	0.759934	0.022365	NBH
	ZIP VS ZINB	-0.89104	0.18672	ZINB
W=0.9	ZIP VS PH	0.193314	0.42336	ZIP=PH
	ZINB VS PH	2.241494	0.004046	ZINB
	ZIP VS NBH	-0.07771	0.046903	NBH
	ZINB VS NBH	0.181721	0.04279	NBH
	ZIP VS ZINB	-0.19557	0.042247	ZINB

Table 4: Vuong tests results varying levels of zero proportions at 50 level of over-dispersion.

Zero proportions	models	V-statistic	P-value	preference
W=0.3	ZIP VS PH	-0.2909	0.38555	PH=ZIP
	PH VS ZINB	-0.2915	0.038534	ZINB
	ZIP VS NBH	-0.2908	0.03856	NBH
	ZINB VS NBH	-0.2913	0.038539	NBH
	ZIP VS ZINB	0.72611	0.023389	ZINB
W=0.5	ZIP VS PH	-0.2723	0.39268	PH=ZIP
	PH VS ZINB	-0.2731	0.039238	ZINB
	ZIP VS NBH	-0.2722	0.39275	NBH
	ZINB VS NBH	-0.2729	0.039245	NBH

	ZIP VS ZINB	0.18474	0.042672	ZINB
W=0.7	ZIP VS PH	-0.1252	0.45017	PH=ZIP
	PH VS ZINB	-0.1254	0.04501	ZINB
	ZIP VS NBH	-0.125	0.045026	NBH
	ZINB VS NBH	-0.1252	0.046502	NBH
	ZIP VS ZINB	0.91277	0.018068	ZINB
W=0.9	ZIP VS PH	-0.1376	0.44528	PH=ZIP
	PH VS ZINB	-0.137	0.044551	ZINB
	ZIP VS NBH	-0.1379	0.044518	NBH
	ZINB VS NBH	-0.1373	0.04454	NBH
	ZIP VS ZINB	0.51006	0.0305	ZINB

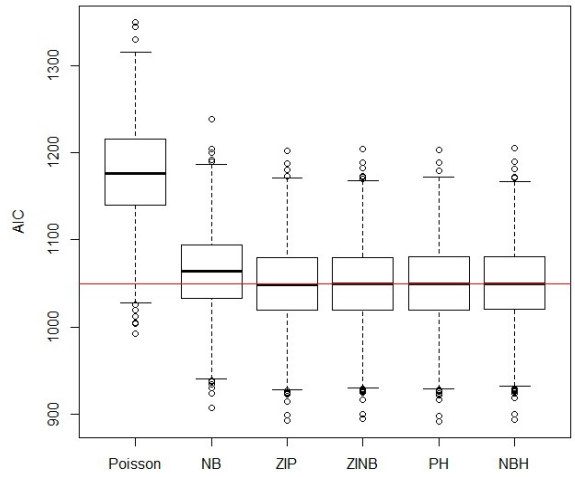

Figure 1: Box plot for the AIC from the six models at (W=0.5 and Ω=10)

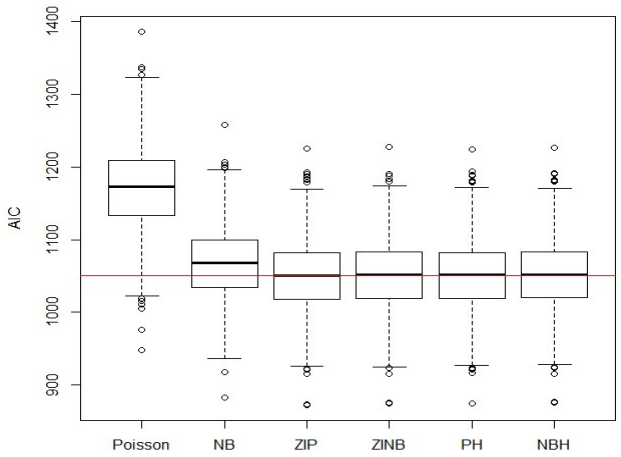

Figure 2: Box plot for the AIC from the six models at (W=0.5 and Ω=15)

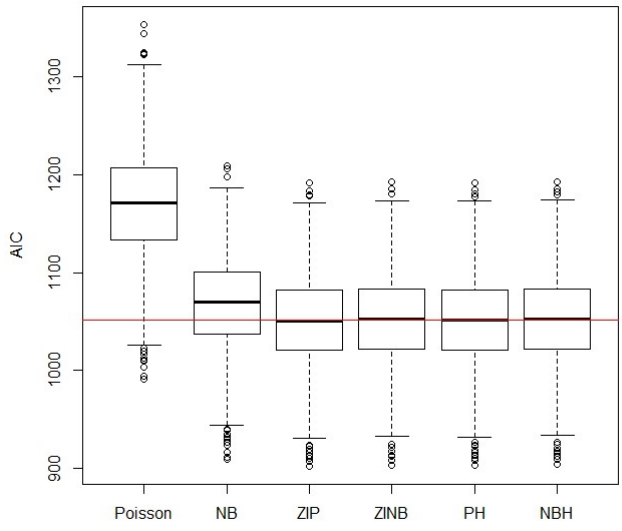

Figure 3: Box plot for the AIC from the six models at (W=0.5 and Ω=20)

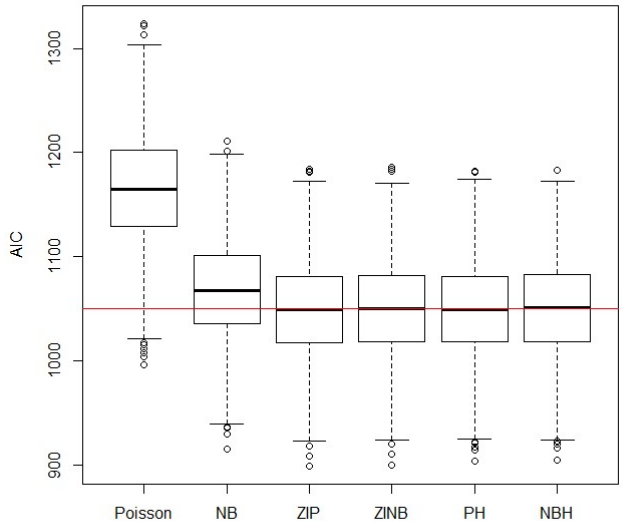

Figure 4: Box plot for the AIC from the six models at (W=0.5 and Ω=30)

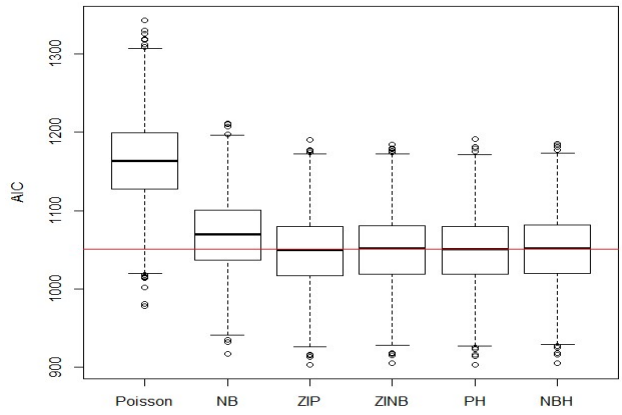

Figure 5: Box plot for the AIC from the six models at (W=0.5 and Ω=50)

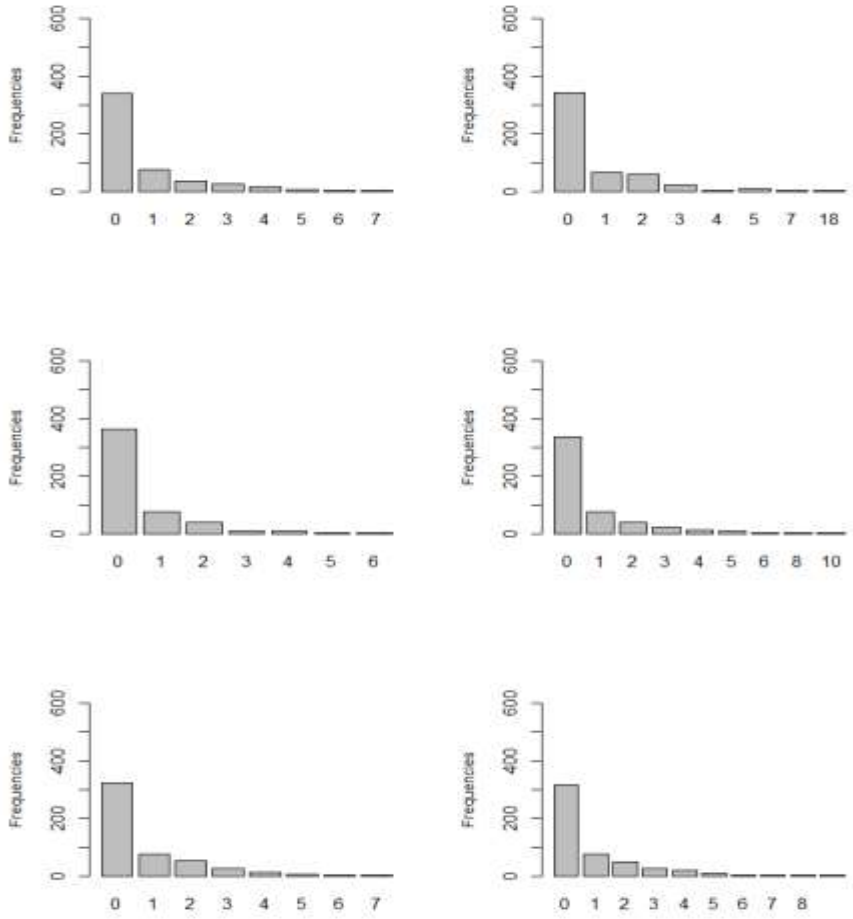

Figure 3.1: Frequency plot for the simulated response variable at w=0.5 and Ω(5,10,15,20,30 and 50 respectively from top right to left bottom)

IV. DISCUSSION

When a researcher, is modelling over-dispersed and zero-inflated count data, the ZI (ZINB and ZIP) and Hurdles models (PH and NBH) are preferred to the GLM models (Poisson and NB). The choice between ZI and Hurdles model should be based on the structure of the data. It is important to note that both ZINB and NBH models are more statistically fit when the data is over-dispersed and zero-inflated compared to ZIP and PH. This is evident in our result as they both had better performance in our Vuong results. From our results, the NBH model performed well compared to ZINB at all scenarios of level of dispersion hence the preferred choice. The ZIP and PH gave similar results in most scenarios. As per the results of the AIC, Log-likelihood ratio, Vuong test and the Box plots the NBH is the preferred choice for the highly dispersed count data with excess zeros.

V. CONCLUSION

Zero-inflated models have gained popularity in recent past due to their ability to handle over-dispersed and zero inflated count data. In most scenarios over dispersion is as a result of excess zeros in the count data. For a researcher, it is paramount to use the most statistically fit model in their analysis to achieve highly accurate results. This can be achieved by getting to explore and understand all the characteristics exhibited by their data. This will enable them to choose the most statistically fit model to adopt in analysis. As per the results from this study, the NBH model is the best model which can over-come both over-dispersion and zero-inflation in count data. The research is time consuming but on the brighter side, it provides the researcher with guidance and knowledge required when dealing with over-dispersion and excess zeros. Recommendation for further research would be, to explore a case where there is under-dispersion and presence of excess zeros in the data.

REFERENCES

[1]. M. Ridout, "Models for count data with many zeros," 1998.
[2]. A. C. Cameron, Regression analysis of count data, vol. 53, 2013.
[3]. W. M. E. P. a. S. E. C. Gardner, "“Regression analysis for counts and rates: Poisson, Over-dispersed Poisson and Negative Binomial," vol. 118, pp. 392-404, 1995.
[4]. S. E. A. R. G. W. Saffari, "“Handling of Over-Dispersion of Count Data via Truncation using Poisson Regression Model”, Journal of," vol. 1, August 2011..
[5]. V. M. F. a. B. W. Shankar, "Effect of roadway geometrics and environmental factors on rural freeway accident frequencies," vol. 27, no. 3, p. 371–389, 1995.
[6]. D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," Taylor & Francis, vol. 34, pp. 1-14, 1992.
[7]. M. G. N. B. M. S. C. K. L. N. Chipeta, " “Zero adjusted models with applications to analysing helminths count data”,," vol. 7, pp. 1-11, 2014.
[8]. W. S. C. J. Liu, " “Count Data Models in SAS®”,Statistics and Data Analysis, SAS Global Forum," p. 371, 2008.
[9]. J. D. Lewsey, "The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status," Wiley Online Library, vol. 32, no. 3, pp. 183-189, 2004.
[10]. J. Mullahy, "Specification and testing of some modified count data models," Elsevier, vol. 33, pp. 341-365, 1986.
[11]. Y. Min, "Random effect models for repeated measures of zero-inflated count data," Statistical modelling, vol. 5, pp. 1-19, 2005.
[12]. J. M. Miller, "Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation," 2007.
[13]. C. D. Desjardins, "Evaluating the performance of two competing models of school suspension under simulation-the zero-inflated negative binomial and the negative binomial hurdle," 2013.