Zero Inflated models vs Hurdle models in modelling auto-insurance claims in an emerging market: A case study of Nigeria

Mary I.Akinyemi* Abisola A.Rufai†

Abstract

Count data occur naturally in a number of disciplines ranging from economics and the social sciences to finance as well as medical sciences. Most count data are plagued with over-dispersion and excess zeros making it difficult to model them with vanilla linear models. Different models have been proposed to capture this peculiarity in count data viz.: Classical models such as the generalized Poisson regression model and the negative binomial regression model have been used to model dispersed count data. Hurdle and zero-inflated models are also said to be able to capture over-dispersion and excess zeros in count data. In this paper, we compare the performance of Poisson and Negative Binomial hurdle models, zero-inflated Poisson and Negative Binomial models, classical Poisson and Negative Binomial regression models as well as the zero-inflated compound Poisson generalized linear models to modelling frequency of auto insurance claims in a typical emerging market. The model parameters are estimated using the method of maximum likelihood. The models performances are compared based on their information criteria (AIC and BIC) and the Gini index. The zero-inflated compound Poisson generalized linear models out performed the other models considered.

Key Words: Count data ; Zero Inflated models; Hurdle models ; Gini Index

1. Introduction

Count data occur naturally in a number of disciplines ranging from economics and the social sciences to finance as well as medical sciences. Naturally, a typical data set containing the number of insurance claims made over a period is considered as count data (see (Hidayat and Pokhrel, 2010), (Cameron and Trivedi, 1996) and (Famoye and Singh, 2006)). Modelling the number of claims is a crucial part of insurance pricing. Count regression analysis allows identification of risk factors and prediction of the expected frequency claims based on the type of policy taken out and the characteristics of the policy holders. Most insurers would calculate the premium by combining the expected claim amount with the conditional expectation of the number of claims given the risk characteristics. Some insurers may also consider experience rating when setting the premiums, so that the number of claims reported in the past can be used to improve the estimation of the conditional expectation of the number of claims for the following year (Boucher and Guillen, 2008).

Over the years, Insurers gradually amassed sizeable longitudinal information on their policy holders, this somewhat availability of data has allowed research in this area to expand so that the literature on count regression analysis has grown considerably in the past years. (Boucher and Guillen, 2008) in their paper addressed panel count data models in the context of insurance, to showcase the advantages of using the information on each policy holder over time for modelling the number of claims. They argue that new panel data models presented in their work allow for time dependence between observations and are closer to the data generating process that one can find in practice.

Most count data are plagued with over-dispersion and excess zeros making it difficult to model them with vanilla linear models.

*Department of Mathematics, University of Lagos, Lagos Nigeria
†Department of Mathematics, University of Lagos, Lagos Nigeria

Different models have been proposed to capture this peculiarity in count data, (Ozmen and Famoye, 2007) apply the Poisson, NB, GP, ZIP and ZIGP to zoological data set where the count data may exhibit evidence of many zeros and over-dispersion. by modelling the number of C. caretta hatchlings dying from exposure to the sun. (Ismail and Zamani, 2013) fitted negative binomial and generalized Poisson regression models to Malaysian OD claim count data and zero-inflated negative binomial and zero-inflated generalized Poisson regression models were fitted to the German healthcare count data.(Gurmu, 1998) applies generalised hurdle models suitable for the analysis of over-dispersed or under dispersed count data allowing for asymmetric departures from the binary logit model to Medicaid utilisation data. (Shi and Valdez, 2014) investigate alternative approaches to constructing multivariate count models based on the negative binomial distribution. They considered two different methods of modelling multivariate claim counts using copulas. The first one works with the discrete count data directly with the mixture of max-id copulas that allows for flexible pairwise association as well as tail dependence. The second one employs elliptical copulas to join continuitized data while preserving the dependency among original counts. The empirical analysis looks into an insurance portfolio from a Singapore auto insurer where claim frequency of three types of claims (third party property damage, own damage, and third party bodily injury) are considered. The results demonstrate the superiority of the copula based approaches over the common shock model.

Nigeria was named one of the emerging economies in 2014 along with Mexico, Indonesia and Turkey (BBC, 2014). Recently, the Nigerian economy has been badly hit by election uncertainty coupled with a huge dip in oil prices and religious insurgency. Although, the Nigerian economy seems to rock immensely under the new leadership, our staggering population projected at over 180 million still makes us an attractive destination for consumer goods and services especially new and used automobiles. The Nigerian road use laws (http://www.highwaycode.com.ng/iv-vehicle-insurance.html) stipulate that an automobile user shall take out either third party or comprehensive insurance policies, so that most people typically subscribe to some insurance scheme majorly for statutory reasons. However, what we observed is that most automobile users do not make claims even if they can legitimately make one.

This study compares the performance of Poisson and Negative Binomial hurdle models, zero-inflated Poisson and Negative Binomial models, classical Poisson and Negative Binomial regression models as well as the zero-inflated generalized compound Poisson models to modelling frequency of auto insurance claims in Nigeria. The model parameters are estimated using the method of maximum likelihood. The models performances are compared based on their information criteria (AIC and BIC) and the Gini index compares the lift of a model against another model.

The rest of this paper is structured as follows: In Section 2, we discuss the models considered, we give useful details regarding the Zero Inflated models in Section 2.2 and that of the Hurdle models in Sections 2.3. Section 3 describes the data used. The results are presented in Section 4. Finally, Section 5 concludes.

2. Methods

2.1 Generalised Linear Models (GLM)

Consider a set of n observations $y_i \dots y_n$ and a vector of regressors $\boldsymbol{x}$. The Generalised Linear Model (**GLM**) is given as: $f(\lambda) = \beta_0 + \beta_1 * X_1 + \beta 2 * X_2 + \cdots + \beta_n * X_n$ with variance function $Var(Y) = \phi * V(\lambda)/w$ The conditional distribution of $y_i \mid x$ is a linear

exponential family with pdf

$$f(y:\lambda,\phi)=\exp\left(\frac{y\cdot\lambda-v(\lambda)}{\phi}+w(y,\phi)\right) \tag{2.1.1}$$

where λ is the canonical parameter that depends on the regressors via a linear predictor and ϕ is a dispersion parameter that is often known. GLMs calculate the coefficients that maximize likelihood, and w is the weight that each record gets in that calculation. The GLM Variance Function is determined by the distribution e.g.

- Normal: $V(\lambda)=1$
- Poisson: $V(\lambda)=\lambda$
- Gamma: $V(\lambda)=\lambda^2$

The following assumptions are made on generalized linear models

- Target variable Y does not depend on the value of Y for any other record, only the predictors
- Distribution of Y is a member of the exponential family of distributions
- Variance of Y is a function of the mean of Y
- $f(\lambda)$ is linearly related to the predictors. The function $f(.)$ is called the link function
- The functions $v(\cdot)$ and $w(\cdot)$ are known and determine the member of the family of distributions used. The exponential family of distributions include the following: Normal, Poisson, Gamma, Binomial, Negative Binomial, Inverse Gaussian, Tweedie.

Common choices for GLM log link functions include:

- Identity:$f(\lambda)=\lambda$
- Log: $f(\lambda)=\ln(\lambda)$
- Logit: $f(\lambda)=\ln(1\lambda)$

Generalised linear models allow us to quantify uncertainty in parameter estimates e.g.the Wald's confidence interval for mean of parameter. The Wald Chi Square test can also be used to test for the significance of an individual parameter in the model.

2.2 Zero Inflated models

Zero-inflated models have been proposed as a class of models more capable of dealing with excess zeros in count data than the classical GLMs ((Mullahy, 1986); (Lambert, 1992)). They are two-component mixture models combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial. Thus, there are two sources of zeros: zeros may come from both the point mass and from the count component. For modeling the unobserved state (zero vs. count), a binary model is used: in the simplest case only with an intercept but potentially containing regressors (Zeileis et al., 2008). Formally, Zero-inflated models mix a point mass at zero $I_0(y)$ and a count distribution $f_{count}(y;x,\beta)$. The probability of observing a zero count is inflated with probability $\pi=f_{zero}(0;x,\gamma)$:

$$\begin{aligned} f_{zeroinfl}(y;x,z,\beta,\gamma) &= f_{zero}(0;x,\gamma)\cdot I_0(y) \\ &\quad +(1-f_{zero}(0;x,\gamma))\cdot f_{count}(y;x,\beta) \end{aligned} \tag{2.2.1}$$

Where $I(\cdot)$ is an indicator variable and the unobserved probability π of belonging to the point mass component is modelled by a binomial GLM $\pi = g^{-1}(z^\intercal\gamma)$. The corresponding regression equation for the mean is

$$\mu_i = \pi_i \cdot 0 + (1 - \pi_i) \cdot \exp(x_i^\top \beta)$$

using the canonical log link. The vector of regressors in the zero-inflation model z_i and the regressors in the count component x_i need not to be distinct in the simplest case, $z_i = 1$ is just an intercept. The default link function $g(\pi)$ in binomial GLMs is the logit link, but other links such as the probit are also available. The full set of parameters of β, γ, and potentially the dispersion parameter ϕ (if a negative binomial count model is used) can be estimated by ML. Inference is typically performed for β and γ, while ϕ is treated as a nuisance parameter even if a negative binomial model is used.

2.3 Hurdle models

The hurdle model was originally proposed by (Mullahy, 1986). They are two-component models: A truncated count component, such as Poisson, geometric or negative binomial, is employed for positive counts, and a hurdle component models zero vs. larger counts. For the latter, either a binomial model or a censored count distribution can be employed (Zeileis et al., 2008).

Hurdles models combine a count data model $f_{count}(y; x, \beta)$ and a zero hurdle model $f_{zero}(y; x, \gamma)$. The models are such that $f_{count}(y; x, \beta)$ is left truncated at $y = 1$ and $f_{zero}(y; x, \gamma)$ is right truncated at $y = 1$:

$$f_{hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{zero}(0; x, \gamma) & \text{if } y = 0, \\ (1 - f_{zero}(0; x, \gamma)) \cdot \frac{f_{count}(y;x,\beta)}{1-f_{count}(0;x,\beta)} & \text{if } y > 0. \end{cases} \tag{2.3.1}$$

The model parameters β, γ, and potentially one or two additional dispersion parameters ϕ (if f_{count} or f_{zero} or both are negative binomial densities) are estimated by ML, where the specification of the likelihood has the advantage that the count and the hurdle component can be maximized separately. The corresponding mean regression relationship is given by

$$\log(\mu_i) = x_i^\top \beta + \log(1 - f_{zero}(0; z_i, \gamma)) - \log(1 - f_{count}(0; x_i, \beta))$$

using the canonical log link. For interpreting the zero model as a hurdle, a binomial GLM is probably the most intuitive specification. Another useful interpretation arises if the same regressors $x_i = z_i$ are used in the same count model in both components $f_{count} = f_{zero}$: A test of the hypothesis $\beta = \gamma$ then tests whether the hurdle is needed or not.

2.4 Model evaluation

We compare model performances by employing the following penalised measures:

1. Akaike Information Criterion (AIC): It penalizes the loglikelihood for additional model parameters. AIC provides an asymptotically unbiased estimator of the expected Kullback discrepancy between the generating model and the fitted approximating model. it is computed as follows:

   $$AIC = -2 \ln f(y \mid \hat{\theta}_k) + 2k$$

   (Schwarz, 1978).Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

2. Bayesian Information Criterion (BIC): It also penalizes the loglikelihood for additional model parameters, however this penalty increases as the number of records in the dataset increases. BIC provides a large-sample estimator of a transformation of the Bayesian posterior probability associated with the approximating model. it is computed as

$$BIC = -2\ln f(y \mid \hat{\theta}_k) + k\ln n$$

(Schwarz, 1978) and (Kass and Raftery, 1995). Given again a set of candidate models for the data, the one with the lowest BIC is preferred.

Note 2.1. It is noteworthy that AIC and BIC feature the same goodness-of-fit term, however, the penalty term of BIC is more stringent than the penalty term of AIC. (For $n \geq 8$, $k\ln n$ exceeds $2k$.) Consequently, BIC can be too restrictive and tends to favor smaller models than AIC.

We evaluate the model lift using the *Gini Index* before we discuss the Gini Index, we will briefly define the *Lorenz curve*

2.4.1 The Lorenz curve

For a given population, let y be personal income, x a pre-specified level of income, $F(x)$ a fraction of the population with $y \leq x$ with density function $f(x) = F^{'}(x)$. Furthermore, denote the average income (assuming all income is negative) by $\bar{y} = \int_0^\infty yf(y)dy$. The *The lorenz function* is a function $L : [0,1] \to \mathcal{R}$, satisfying,

$$P = F(x) \implies L(P) = \frac{\int_0^x yf(y)dy}{\bar{y}}$$

. Where P is a proportion of the said population (Atkinson and Bourguignon, 2000) and (Aghion and Durlauf, 2005).

The *Lorenz curve* is simply the graph of $(P, L(P))$.

2.4.2 Gini Index

The Gini index (also called the Gini coefficient or the Gini ratio) is defined on the basis of the Lorenz curve and is a measure of the degree of income inequality in society. The Gini index is defined as

$$Gini = \frac{\int_0^1 (P - L(P))dP}{1/2}$$

.

A high value of Gini means high degree of inequality in the distribution of income. If everybody had the same income, then the Lorenz curve would coincide with the 45^o line and the Gini index would be zero. In the context of this paper, a high value of Gini means a model has a higher lift than the other and if the models were the same, the Lorenz curve would coincide with the 45^o line on the axis of the plot and the Gini index would be zero(Atkinson and Bourguignon, 2000).

3. Data

The data consists of 616 policies issued between 2011-2015 by an indigenous insurance company. The attributes available for consideration from the data source include: year policy was taken our (year), gender of the policy holder, class of the car (private or commercial), premium, insurance type (third party or comprehensive) and the number of claims.Table 1 below shows the summary of the data considered.

Table 1: Descriptive statistics of the data

Attribute	Factors	Frequency	Percent
Year	2011	72	11.69
	2012	101	16.40
	2013	158	25.65
	2014	200	32.47
	2015	85	13.80
Gender	Male	526	85.39
	Female	90	14.61
Motor class	Private	478	77.6
	Commercial	138	22.4
Insurance type	Third party	17	2.76
	Comprehensive	599	97.24
Claims	0	565	91.72
	1	48	7.79
	2	2	0.32
	3	1	0.16

From Table 1 we see that the highest number of policies were taken out between 2013 and 2014, more than 85% of these policies were taken out by men. We also observed that most of the customers who took out policies took them out on their private cars and there was a preference for comprehensive insurance policies (>97%). Furthermore, we observed that $> 91\%$ of the policy holders made no claims so that the data does have many zero (i.e. it is zero inflated).

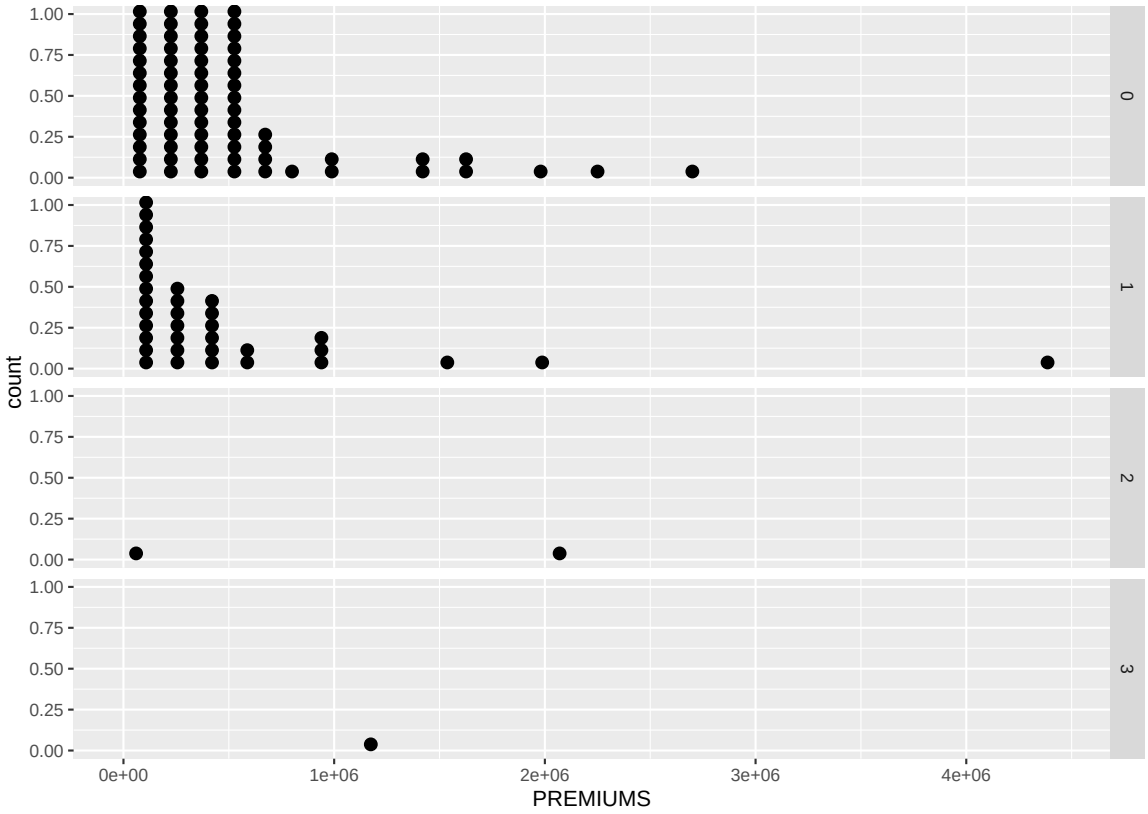

Figure 1: Premium amount by number of claims

In addition, we observe from Figure1 that the bulk of the customers with the no claims fall within the lower average premium bracket. Furthermore

Since data consists of 97% comprehensive and 3% third party insurance policy holders and none of the third party insurance policy holders made any claims in the time period considered, we base the analysis on comprehensive insurance policy holders only.

4. Results

The Gini index (Table 2) and corresponding asymptotic standard errors (Table 3) were computed based on the ordered Lorenz curve (Figure 2) for each of the 7 models considered. It was observed that the GLM-type models had the least performance.

Table 2: Gini Index scores

	poisreg	negbin	zeropois	zeronegbin	hurdlepois	hurdlenegbin	zicglm
Poisson	0.00	14.21	16.99	16.95	20.65	19.51	19.80
Negative binomial	-11.12	0.00	16.30	16.30	21.19	20.19	18.03
Zero inflated poisson	-6.04	-5.05	0.00	-5.58	0.59	0.16	5.84
Zero inflated negative binomial	-6.01	-5.06	5.60	0.00	0.60	0.13	5.81
Hurdle poisson	-12.28	-13.49	7.79	7.81	0.00	7.06	10.92
Hurdle negative binomial	-11.37	-12.73	9.30	9.33	-4.70	0.00	11.90
Zero inflated generalised compound poisson	-4.50	-2.21	2.11	2.16	1.58	2.08	0.00

We observed from Table 2 (Greyed areas represent better lift)) that the zero inflated models as well as the hurdle models have better lift than the classical Possion and Negative Binomial models. Furthermore, the classical Negative binomial model also has better lift than the classical Poisson model. The zero inflated models also have better lift than the

hurdle models. The zero inflated generalised compound poisson model outperforms all the other models. In addition, according to the "min-max" argument, the selected best model is the Zero inflated generalised compound poisson model (ZIGCPM).

Table 3: Corresponding standard error for Gini Index scores

	poisreg	negbin	zeropois	zeronegbin	hurdlepois	hurdlenegbin	zicglm
Poisson	0.00	8.30	8.35	8.36	7.74	8.15	7.86
Negative binomial	8.71	0.00	8.50	8.50	7.79	7.86	8.20
Zero inflated poisson	8.48	8.56	0.00	7.91	8.81	8.52	7.65
Zero inflated negative binomial	8.48	8.56	7.91	0.00	8.78	8.52	7.65
Hurdle poisson	7.82	7.79	8.68	8.66	0.00	7.75	7.95
Hurdle negative binomial	8.19	7.85	8.40	8.40	7.90	0.00	7.90
Zero inflated generalised compound poisson	8.08	8.26	7.93	7.92	8.21	8.14	0.00

- Figure 2 is the plot of the ordered Lorenz curves for the data.

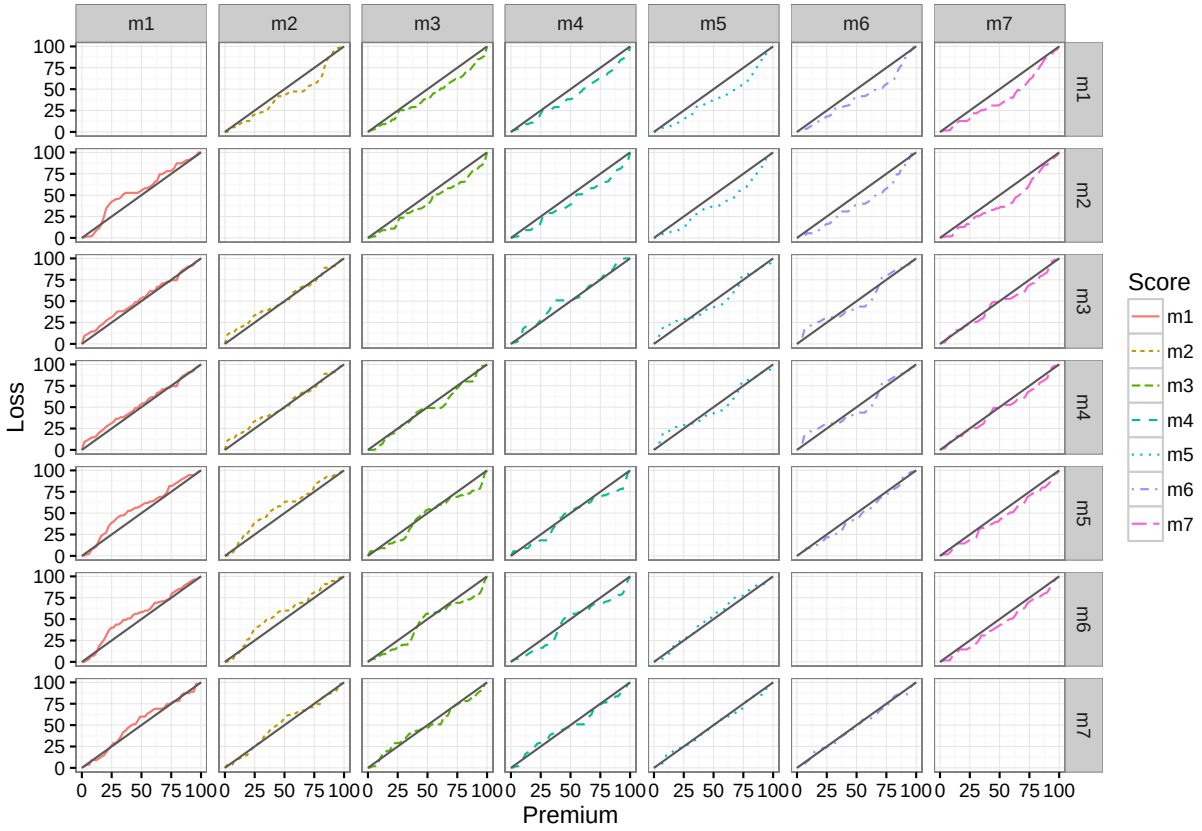

Figure 2: Lorenz curve

The results of the model selection criteria (Akaike information criteria (AIC) and Bayesian information criteria (BIC)) are presented in Table 4. The results of the AIC and BIC criteria for the models agree with that of the Gini index as the ZIGCPM still shows up as the best model since it has the smallest AIC and BIC.

Table 4: AIC and BIC results for each model

	AIC	BIC
Poisson	370.10	392.08
Negative binomial	371.66	398.03
Zero inflated poisson	374.11	413.67
Zero inflated negative binomial	376.11	413.67
Hurdle poisson	372.91	408.07
Hurdle negative binomial	374.07	407.23
Zero inflated generalised compound poisson	234.93	265.70

5. Summary

Zero Inflated models have been proposed as better models for modelling count data with excess zeros. This work applied the Classical Poisson and Negative binomial, Zero inflated Poisson and Negative binomial models and Hurdle Poisson and Negative Binomial models as well as Zero Inflated Generalised Compound Poisson model to auto insurance claims data from a typical indigenous Nigerian insurance company. The results selected the Zero Inflated Generalised Compound Poisson model as the optimal model.

References

Aghion, P. and Durlauf, S., editors (2005). *Handbook of Economic Growth*, volume 1. Elsevier, 1 edition.

Atkinson, A. and Bourguignon, F., editors (2000). *Handbook of Income Distribution*, volume 1. Elsevier, 1 edition.

BBC (2014). The mint countries: Next economic giants? *International Journal of Environmental Research and Public Health*.

Boucher, Jean-Philippe, M. D. and Guillen, M. (2008). Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions. *Variance*, 2(1):132–162.

Cameron, A. C. and Trivedi, P. K. (1996). Count data models for financial data. In *Handbook of Statistics*, pages 363–391. Elsevier, North-Holland.

Famoye, F. and Singh, K. P. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4(1):117–130.

Gurmu, S. (1998). Generalized hurdle count data regression models. *Economics Letters*, 58(3):263 – 268.

Hidayat, B. and Pokhrel, S. (2010). The selection of an appropriate count data model for modelling health insurance and health care demand: Case of indonesia. *International Journal of Environmental Research and Public Health*, 7(1):9–27.

Ismail, N. and Zamani, H. (2013). Estimation of claim count data using negative binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

Ozmen, I. and Famoye, F. (2007). Count regression models with an application to zoological data containing structural zeros. *Journal of Data Science*, 5(4):491–502.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shi, P. and Valdez, E. A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, 55:18 – 29.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software*, 27(1):1–25.