Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 25, Issue 4 (2023) ISSN$_{Print}$: 1816-9171

Using Zero Inflation Poisson Regression Model To Analyze Lung Cancer Data

Tahir R. Dikheel

Huyam A. Jouda

University of AL-Qadisiyah, Iraq

Corresponding Author : Huyam A. Jouda

Abstract **:** In this article, the phenomena that follow the models for dealing with numerical data, which take the form of time series, have been dealt with. This type of time series often suffers from the problem of zero inflation. which represents a problem that cannot be overlooked. so a set of methods and methods has been proposed to deal with this problem. and one of the most important models in this case is Poisson's zero-inflated model. It should be noted here that there is a set of methods that are used to estimate the parameters of the Poisson zero-amplified model. including the MLE method and the NLM method.

Thesis problem:

The presence of zero inflation in an integer time series.

Research objective:

The research aims to use the zero-inflated Poisson model to deal with time series data that take the form of counts in the presence of zero inflation and estimate the model parameters using the Hardel model and traditional methods (MLE; NLM) and compare them.

Time Series:

Time series is a collection of random variables over an extended period of uses of time series analysis including forecasting of the economy, sales, the stock market, sports and others.

Fitting a model that describes the time series' structure and offers practical interpretations is the fundamental goal of time series modelling. An application for a fitted model is:

- To draw attention to the essential elements of the time series, such as change points, seasonality, and trend.
- To clarify the relationship between present and past occurrences so that future values of the series can be predicted.

The data aren't always independent, which is a one-way time series analysis that differs from regression analysis. Let (X_1, X_2,⋯ , X_n) be a time series with n elements, denoted as $\{X_t\}_{t=1}^{n}$ The mean of $\{X_t\}_{t=1}^{n}$ $is\ \mu_t = E[X_t]$. The structure of covariance of $\{X_t\}_{t=1}^{n}$ can be shown by its autocovariance function (ACVF). ACVF with h lag at time t can be written as:

$\gamma_x(t, t + h) = Cov(X_t, X_{t+h}) = E(X_t, X_{t+h}) - E(X_t)E(X_{t+h})$ ……………………………….………(1)

the time series $\{X_t\}$ if it meets the following requirements:

a) the mean $E(X_t)$ is the same for each t.

b) for all t and every $h \in \{0,1,2,\cdots\}$the covariance between X_t and X_{t+h} is same. Similarly,

It is argued that a time series $\{X_t\}$ is strictly stationary if ($X_1, X_2, \ldots , X_n$) & ($X_{1+h}, X_{2+h}, \ldots , X_{n+h}$) for all integers, the joint distribution is the same $h > 0$ and $n > 0$. Clearly, weakly stationary is implied by rigidly stationary.

In the case of a (weakly) stationary series $\{X_t\}$, In this setting, the lag h ACVF is independent of t for some of h.

$$\gamma_x(t, t + h) = \gamma_x(0, h)$$

For ease of notation, all ACVFs can utilize the same argument:

$\gamma_x(h) = \gamma_x(0, h)$.

Sometimes it is simpler to examine correlations than covariances. A stationary time series' autocorrelation function (ACF) $\{X_t\}$ is defined as

$$\rho(h) = Corr(X_t, X_{t+h}) = \frac{\gamma(\text{h})}{\gamma(0)}$$

The Cauchy-Schwarz inequality clearly shows that ACFs are between -1 and 1. The impact of series dispersion is eliminated in ACFs. ACFs can be used to compare how dependent various series are.

For series, it’s advantageous to pursue a partial autocorrelation function (PACF). In general, a conditional correlation is a partial correlation. The conditional correlation between X_t & X_{t+h}, $h > 0$, is what is used to (PACF) for a time series between

X_t and X_{t+h}, conditional on $X_{t+1}, X_{t+2}, \dots, X_{t+h-1}$

$$\kappa(h) = Corr(X_t, X_{t+h} | X_{t+1}, \cdots, X_{t+h-1}),$$

after linear prediction for all variables between X_t and X_{t+h}, where the conditional correlation is taken between X_t and X_{t+h}.

The most popular model class in stationary time series analysis is an autoregressive moving average (ARMA) model class. The general (ARMA) model was introduced by (Peter Whittle in 1970). The most current observation in a series is linked to older observations and incorrect forecasts using the ARMA model class. the ARMA(p,q) model contains moving-average terms up to order q as well as autoregressive terms up to order p. It abides by recursion.

$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + z_t - \theta_1 z_{t-1} - \theta_2 z_{t-2} - \cdots - \theta_q z_{t-q}$ ………(2)

where p $and\ q$ non-negative integers. White noises make up the series (z_t) which is frequently thought to have an independent, uniform distribution in time t. the ARMA(p,q) model is also known as a moving-average model (MA(q)) where $p = 0$. Likewise, when $q = 0$, This model is known as an autoregressive model of order p (AR(p)).

We can use model (ARIMA) Autoregressive integrated moving average to illustrate patterns in non-stationary series. The model's components, which consist of d difference operation, q moving average terms and p autoregressive terms, are defined in the form ARIMA (p, d, q), greater formality, operation $\{X_t\}$ if $(1 - B)^d X_t$ is ARMA(*p,q*), when(*p,d,q*) are positive integers is called to be ARIMA (*p,d,q*), $(1 - B)^d$ is the d_{th} operator for order differences.

The models, that belong to the generalized linear family, the negative binomial (NB) regression model and the poisson regression model...etc.

Count Time Series:

There has been significant recent interest in modelling stationary series that have discrete marginal distributions. Often, the discreteness arises in the form of counts taking values in $\{0, 1, 2, \cdot\cdot\cdot\}$. Count series are widely used when describing storm numbers, accident tallies, wins (Yisu Jia , 2018)

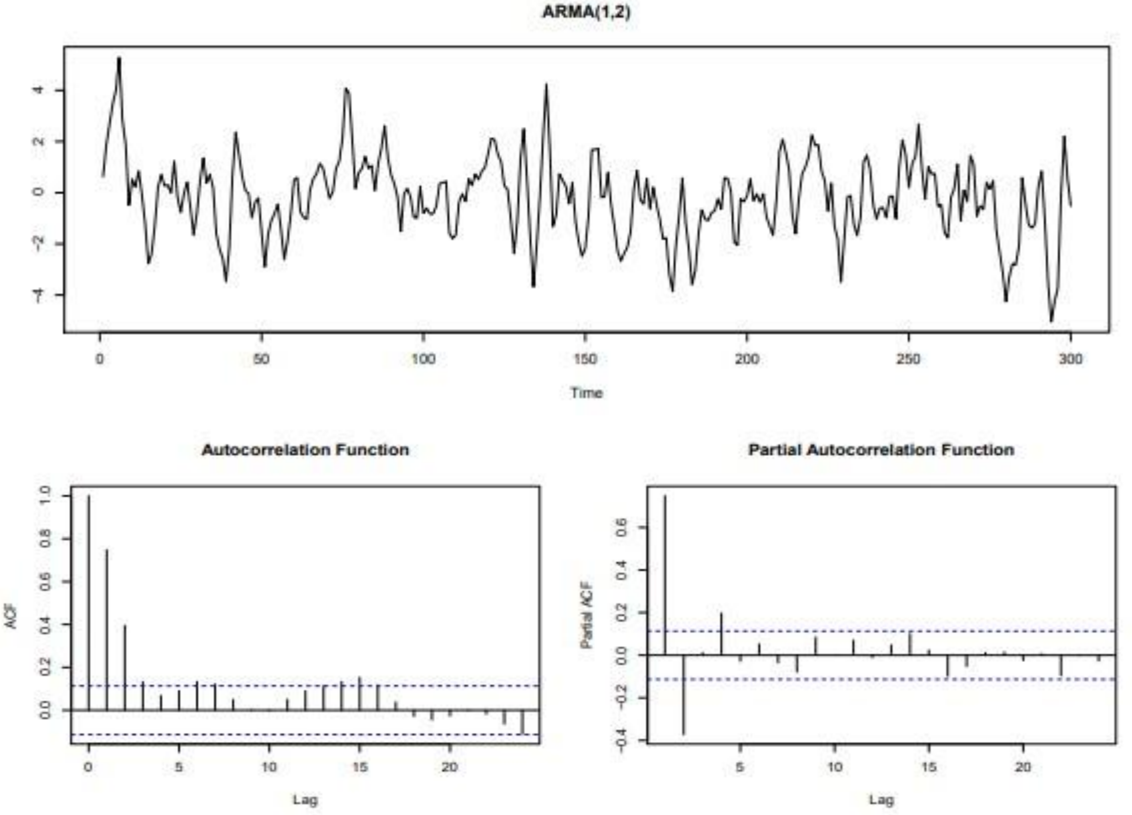

Figure 1: Realization of 300 observations of a ARMA(1,2) with $\emptyset_1 = 0.5$, $\theta_1 = 0.5$, $\theta_2 =$ 0.3.

By a sports team, disease cases, etc. An example of a count time series is shown in Figure 3, and in example of a non-count time series ARMA (1,2) Figure 1, and its (ACF, PACF). It shows the annual number of Atlantic tropical cyclones from 1850 to 2011. The observation at each time is integer-valued, which clearly cannot be normally distributed (Yisu Jia , 2018).

The traditional ARMA/ARIMA model classes work well in describing series with Gaussian marginal distributions. However, no one definitive model class dominates the count series literature. In fact, the autocovariance function of many commonly used count models is deficient in some senses, as described below (McKenzie, E, 1988).

The theory of stationary Gaussian time series is well developed by now. However, there is no known result characterizing autocovariance functions of stationary count series. Elaborating, $\gamma X(\cdot)$ is a symmetric non-negative definite function on the integers, if and only if there exists a stationary Gaussian sequence {Xt} with $Cov(X_t = X_{t+h}) = \gamma X(h)$ for all integers h. Here, non-negative definite (McKenzie, E, (1988).

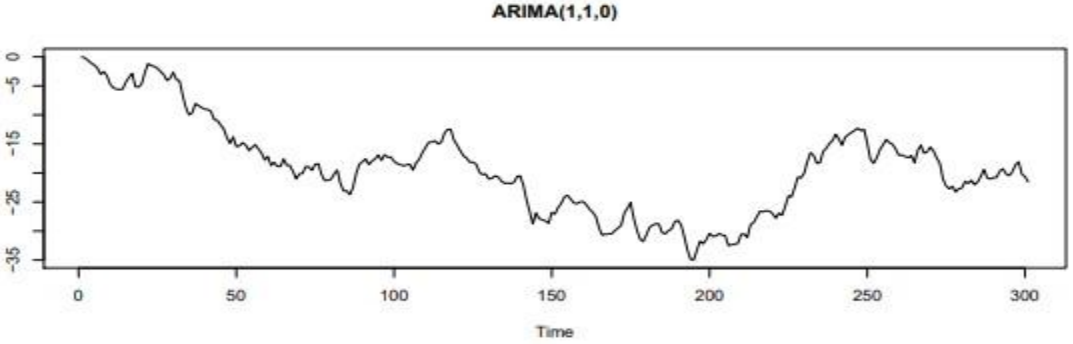

Figure 2(a): realization of 300 observations of an ARIMA (1,1,0) series with $\emptyset_1 = 0.3$.

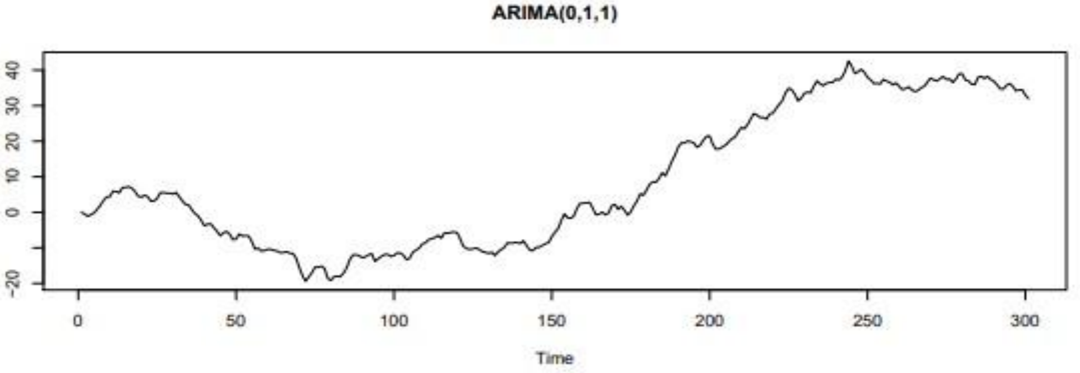

Figure 2(b): realization of 300 observations of an ARIMA (0,1,1) series with $\theta_1 = 0.5$. is defined as

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i \gamma X(t_i - t_j) a_j \geq 0$$

for every choice of $n \in \{1, 2 \ldots, \}$ and real numbers $a_1, \ldots, a_n$. Unfortunately, no analogous result exists for say, a stationary series with Poisson marginal distributions. In fact, restrictions on autocovariance functions of count time series are often more stringent than just non-negative definiteness. For example, it may not be possible to have a stationary count series having a specific marginal distribution that is highly negatively correlated at some lag while the autocovariance function can take on any value between -1 and 1 in a Gaussian process (McKenzie, E, (1988) .

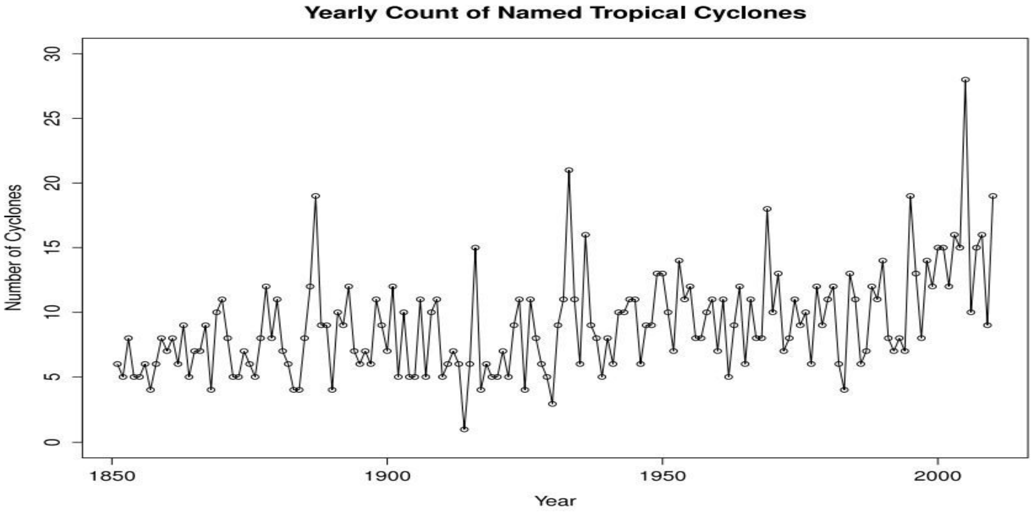

Figure 3: Annual number of Atlantic tropical cyclones from 1850 to 2011.

The Poisson Regression Model:

This kind of count time series is used in this thesis. Poisson distribution models the probability of y, its formula:

$$Pr(Y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!} \qquad (y = 0,1,2,\dots) \text{ ……………………………………………………..(3)}$$

Keep in mind that there is only one parameter used to define the Poisson distribution. This is a rare event's average incidence rate per unit of exposure. parameter μ it can be explained as the risk of a new incident of the event during a given exposure period, t. The probability of y events is given by the relation:

$$Pr(Y = y|\mu, t) = \frac{e^{-\mu t}(\mu t)^{y}}{y!} \qquad (y = 0,1,2,\dots) \text{ …………………………………………………(4)}$$

The likelihood that the mean, variance and Poisson distributions are equal is.

Regular multiple regression is similar to Poisson regression, but the dependent variable (Y) in Poisson regression is a count that is observed and follows(typical references include Heilbron (1989)) Poisson distribution. Thus, the possible values of (Y) are positive integers: (0, 1, 2,3,ect.). It's believed that high numbers are uncommon. In light of the fact that logistic regression also involves a discrete response variable, Poisson regression is comparable to it. However, unlike in logistic regression, the answer is not constrained to particular values (Zeger, S. L, 1988).

The investigation of the relationships between the colony counts of bacteria and various environmental factors and dilutions is one example of an appropriate application of Poisson regression. Another illustration is the quantity of machine breakdowns under various operating circumstances. Another illustration would be crucial data on cancer incidence or newborn mortality in certain demographic groups (Giles, D. E, 2007).

In Poisson regression, we assume that of collection of k regressor variables (X's) determine the Poisson incidence ratio. A formula of this quantity is:

$$\mu = t\ exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \text{ ………………………………………………………………………………….(5)}$$

The regression coefficients $\beta_1, \beta_2, \dots, \beta_k$ a parameters that are unknown and are estimated using the set of data. Their estimates are labeled $b_1, b_2, \dots, b_k$. that notation is used For an observation, the basic Poisson regression model is expressed as

$$Pr(Y_i = y_i|\mu_i, t_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!}$$

Where

$$\mu_i = t_i\ \mu(X_i'\beta) = t_i\ exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) \text{ ……………………………………………………..(6)}$$

In other words, the result follows the Poisson distribution for a particular set of the regressor variables' values (Winkelmann (2000) and Cameron & Trivedi (1998)).

The Poisson model has a number of issues, It includes the Zero-inflation result of the model estimate.

Estimation:

The Maximum Likelihood (ML) estimator:

We observe data $\{(x_i, y_i)|1 \leq i \leq n\}$. The number y_i is a realization of the random variable Y_i. Using independence, the total log-likelihood is given by (Winkelmann, R, 2008):

$$log\ L(y_i, \ldots,, y_n \backslash \beta, x_i, \ldots, x_n = \sum_{i=1}^{n} log\ P(Y_i = y_i \backslash \beta, x_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

With, according to:

$$P(Y_i = y_i \backslash \beta, x_i) = \frac{exp(-\mu_i)\mu_i^{y_i}}{y_i!}$$

And μ_i = exp ($\beta^t x_i$). Write now long $L(\beta)$ is a quick way to express the overall likelihood.
It then follows.

$$log\ L(\beta) = \sum_{i=1}^{n}\{-exp(\beta^t x_i) + y_i(\beta^t x_i) - \log(y_i!)\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(8)$$

Therefore (ML) estimator is of course is given by:

$$\hat{\beta}_{ML} = arg_{\beta} max\ log\ L(\beta)$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = 0$$

With $\hat{y}_i = exp(\beta^t x_i)$ the fitted value of y_i . As is customary, the anticipated fitted value has been used as the estimated value of $E[Y_i|x_i]$. The vector of the residual is orthogonal to the vectors of the explicative variables, according to this first-order requirement (White, H,1982).

The advantage of the maximum likelihood framework is that the cov formula is readily available:

$$cov(\hat{\beta}_{ML}) = \left(\sum_{i=1}^{n} x_i x_i^t \hat{y}_i\right)^{-1}$$

Additionally, Wald tests, Lagrange multiplier tests, and Likelihood Ratio tests can now be used to do hypothesis tests (Cameron & Trivedi ,2013).

Hurdle model:

In contrast to ZI models, hurdle models (Mullahy 1986; Heilbron 1994) can be viewed as a two-component mixture model consisting of a zero mass and the positive observations component following a truncated count distribution, such as truncated Poisson or truncated NB distribution.

Let Yi denote the response of the i th observation, i = 1,···, n , where n denotes the total number of observations. The general structure of a hurdle model is given by:

$$P(Y_i = y_i) = \begin{cases} p_i & y_i = 0 \\ (1 - p_i)\frac{p(y_i;u_i)}{1-p(y_i=0;u_i)} & y_i > 0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

where p_i is the probability of a subject belonging to the zero component; p(yi;μi) represents a probability mass function (PMF) for a regular count distribution with a vector of parameters μi and p(yi = 0;μi) is the distribution evaluated at zero. It can be seen that the positive count is governed by a regular counts distribution as the PMF divided by 1 minus the PMF of this regular counts distribution evaluated at zero.

For example, if the count distribution follows a Poisson distribution, the probability distribution for the hurdle Poisson model is written as:

$$P(Y_i = y_i) = \begin{cases} p_i & y_i = 0 \\ (1 - p_i)\frac{e^{-u_i}u_i{}^{y_i}/y_i!}{1-e^{-u_i}} & y_i > 0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (10)$$

Application:

Lung Cancer:

It's with every five cases of cancer, there is one case in males, and out of every nine cases of cancer in females, there is one case. Lung cancer ranks second in terms of the speed of its spread compared to the rest of the types of cancer that originate in the lungs' cellular structure. Many additional cancers, including breast and kidney cancers, have the potential to disseminate (metastasize) to the lungs. There is no referral for the cancer to be lung disease when this occurs. This is so because the location of the original tumor determines what type of cancer it is and how it is treated. It's divided into two basic types: Small Cell Lung Cancer and Non-Small Cell Lung Cancer (NSCLC) for instance, if breast cancer has spread to the lungs, it be treated as metastatic breast cancer rather than lung cancer

(SCLC). These varieties develop and disperse differently. They are frequently handled differently. Lung cancer statistics are used to highlight the performance methodologies. The writers gathered these statistics from an Iraqi medical facility in Al-Naasiria City, Iraq. They reflect the number of lung cancer patients diagnosed each day in Al-Naasiria City between January 1 and December 31, 2021. One response variable (lung cancer) and bacterial water pollutants (T.P.C.) make up these data. Variable (Y) represents the number of people with lung cancer, and variable (X) represents bacterial water pollutants (T.P.C).

The table 5 represents General statistics of lung cancer data:

Table1 : Statistical indicators

Variables	Means	SD	Cv
X	5.12	2.1428	41.8512
Y	5.1224	2.1599	42.1658

It is clear from the above table that the value of the arithmetic mean for the variable X amounted to 5.12 with a standard deviation of 2.1428 and a coefficient of variation of 41.8512, and the value of the arithmetic mean for the variable Y amounted to 5.1224 with a standard deviation of
2.1599 and a coefficient of variation of 42.1658.

Table 2: Estimation parameter and RMSE for the methods.

		Par			Expar		
method	RMSE	b_0	b_1	b_2	b_0	b_1	b_2
Mle	0.1838	0.9719	-0.0021	-0.2262	2.6429	0.9979	0.7976
hur.count	0.0142	2.0148	-0.0008	0.0401	7.4990	0.9992	1.0418
Nlm	0.0904	1.7298	-0.0247	-0.0215	5.6395	0.9756	0.9787

It is clear from the above table that the hur.count method is the best because it gave the lowest estimate based on the RMSE values. The results proved that water pollution (T.P.C) is one of the causes of lung cancer.

Conclusions:

The time series that takes the numerical form often suffers from the problem of zero inflation, which represents a problem that cannot be overlooked, so a group of methods was proposed to deal with this problem, and one of the most important models, in this case, is the zip model, it should be noted here that There is a set of methods that were used to estimate the parameters of the zip model, including the method of mle, nlm, hur. count, by applying it to real lung cancer data that was collected from the cancer tumor center in Dhi Qar daily for one year, and water pollution data was also collected (Chemical pollution) leading to lung cancer, it was found that chemical pollution is one of the causes leading to lung cancer.

References:

Cameron, A. C., & Trivedi, P. K. (2001). Essentials of count data regression. *A companion to theoretical econometrics*, *331*.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge university press.

Giles, D. E. (2007). Modeling inflated count data. In *MODSIM 2007 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand, Christchurch, NZ* (pp. 919-925).

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, *36*(5), 531-547.

Jia, Y. (2018). *Some Models for Count Time Series* (Doctoral dissertation, Clemson University).

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, *20*(4), 822-835.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, 1-25.

Whittle, P. (1970). Probability. Penguin.

Winkelmann, R. (2000). Seemingly unrelated negative binomial regression. *Oxford Bulletin of Economics and Statistics*, *62*(4), 553-560.

Winkelmann, R. (2008). *Econometric analysis of count data*. Springer Science & Business Media.

I. Zeger, S. L. (1988). A regression model for time series of counts. *Bio*