Bayesian Parameter Estimation for Poisson AR Model

Vladimira Seckarova and Kamil Dedecius

Abstract A popular model in time series analysis is the autoregression model. It explains the current value of a modeled process by means of a weighted average of its past values. In this paper, we focus on two important aspects. Firstly, the modeled variable stands for counts, i.e., it takes values in the set of nonnegative integers. Secondly, we focus on *online* modeling, where the arriving observations sequentially update estimates of the parameters. The emphasis is put on low computational requirements, opening the way towards high-rate real world applications. The paper describes our initial results in this domain. The solution is based on a Poisson autoregression model, where the linear predictor and the modeled variable are canonically linked by the logarithmic function. The adopted Bayesian estimation framework relies on an analytical approximation of the Poisson model by a Gaussian density. A Gaussian prior then provides analytically tractable posterior estimator. Two examples demonstrate the feasibility of the solution. We demonstrate a potential of our approach on the simulated data as well as COVID-19 data and notice advantageous performance of considered methods especially for the use in machine learning.

1 Introduction

Traditional statistical analyses of time series rely on techniques of exponential smoothing and ARMA modeling. In our work, we focus on the autoregression models, where the prediction of the variable of interest is based on a weighted linear combination of its past values. Based on their structure, they allow for modeling

Vladimira Seckarova
School of Mathematics, Monash University, 9 Rainforest Walk, Clayton, Victoria 3800 Australia, e-mail: vladka.seckarova@monash.edu

Kamil Dedecius
Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 182 00 Prague, Czech Republic, e-mail: dedecius@utia.cas.cz

of stationary processes, random walks (potentially with a drift), or oscillating processes. For this substantial flexibility, the autoregression models have found an immense number of applications in virtually all domains of human activity, from humanities over science to technology [4, 15, 21]. Their flexibility has also led to development of numerous modifications and improvements of the basic autoregression model. To name only a few: the autoregression model with exogeneous inputs (ARX) is popular in the control theory [17], the seasonal autoregression (SAR) model used in domains with fixed periodicity of the observed variable, the nonlinear autoregression and its variants [3], or the vector autoregression (VAR) [22]. The autoregression can even represent the temporal evolution of a random process variance, which gives rise to the autoregressive conditional heteroskedasticity (ARCH) models [9]. The combination of various AR models with moving average models then yields the phenomenal class of ARMA models [4, 21].

We focus on autoregression models of count variables. Recall, that the 'classical' AR models assume Gaussian distribution of the error component. Violation of this assumption need not be critical if the counts are high enough. However, if the counts are low, the 'classical' models are prone to failure. In particular, the forecasted values may take negative values. One possible solution is to consider the variable at time t to be connected with its (transformed) past values through a convenient link function. This gives rise to the generalized linear models (GLMs) [20]. If the link function is the natural logarithm, the resulting model is Poisson. For completeness we mention, that the 'classical' AR model is a GLM too – the link function is the identity function.

A common feature of the inferential methods for the Poisson AR model is their offline character. This means, that they are able to estimate the model parameters from a batch of observed data. However, if new observations are acquired, there is no computationally cheap way to update the previous estimates. In combination with the iterative character of the inferential methods, this may be a significant obstacle to the use of the Poisson AR model in quick online applications.

The present paper reports our initial results in the domain of *sequential* modeling of time series of counts. We exploit the results of the second author in Bayesian estimation of the Poisson GLM [7]. However, they are not fully transferable, as the time series domain faces specific issues, e.g., those connected with the stationarity of AR processes, or the invertibility of MA processes. These topics and their impact on estimation are postponed to future research.

2 Poisson autoregression

Let us assume a probability space $(\Omega, \mathscr{A}, P)$ and $\mathscr{T} \equiv \mathscr{N}_0$ a set of discrete time indices. Furthermore, let $(Y_t)_{t\in\mathscr{T}}$ be a discrete-time stochastic process with nonnegative values $y_t \in \mathbb{N}_0$. Let the variable X_t be measurable with respect to the σ-algebra

$$\mathscr{F}_{t-1} = \sigma(Y_\tau; \tau \leq t-1).$$

If it follows from the generative model

$$Y_t|\mathscr{F}_{t-1} \sim \text{Poisson}(\lambda_t) \tag{1}$$

with $\lambda_t = \exp(\nu_t)$ where

$$\begin{aligned} \nu_t &= \beta_0 + \sum_{i=1}^{p} \beta_i \log(L^i Y_t) && (2)\\ &= \boldsymbol{\beta} \boldsymbol{X}_t && (3)\\ &= \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} 1 \\ \log(Y_{t-1}) \\ \vdots \\ \log(Y_{t-p}) \end{bmatrix}, \qquad p \geq 1, && (4) \end{aligned}$$

we talk about the Poisson autoregressive process of order p. The operator L is the lag (or backshift) operator. The resulting series $(Y_1, \ldots, Y_t)$ is the *time series of counts*. $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is the vector of regression coefficients, and $\boldsymbol{X_t} \in \mathbb{R}^{p+1}$ is the regressor. From the properties of the underlying Poisson distribution, it holds

$$\begin{aligned} \mathbb{E}[Y_t|\mathscr{F}_{t-1}] &= \lambda_t = \exp(\boldsymbol{\beta} \boldsymbol{X}_t), && (5)\\ \text{var}[Y_t|\mathscr{F}_{t-1}] &= \lambda_t = \exp(\boldsymbol{\beta} \boldsymbol{X}_t). && (6) \end{aligned}$$

This property is known as the equivariance.

For obvious reasons, the given model admits zero values of Y_t only in the limit case if $\nu_t \to -\infty$, which makes it inappropriate for most real time series. Several modifications have been proposed in the literature to deal with this issue. For instance, the vector $\boldsymbol{X_t}$ is replaced with

$$\boldsymbol{X}'_t = \begin{bmatrix} 1 \\ \log(Y_{t-1} + c) \\ \vdots \\ \log(Y_{t-p} + c) \end{bmatrix}, \qquad \text{or} \qquad \boldsymbol{X}'_t = \begin{bmatrix} 1 \\ \log(\max(Y_{t-1}, c)) \\ \vdots \\ \log(\max(Y_{t-p}, c)) \end{bmatrix}. \tag{7}$$

Both models have been used to model counts, see [23]. In the former case, a positive constant c ("immigration" rate) is added to the measurements in order to suppress zero values. In the latter case, c in the argument $\max(Y_{t-1}, c)$ relates to the probability of $Y_t > 0$ given that $y_{t-1} = 0$ for Poisson distributed random variable.

Mostly, the value $c = 1$, but it is reported that a reasonable values of c do not have a gross impact on modeling performance [10].

The presented model is only one of many possible choices. For further details see book of Kedem and Fokianos [14, Chapter 4], [16], [6].

3 Sequential estimation of model parameters

The Poisson model (1) is a standard generalized linear model with a *canonical* logarithmic link function [18]. It is characterized by the probability density function

$$p(y_t|\mathscr{F}_{t-1}) = \frac{\lambda_t^{y_t}\exp(-\lambda_t)}{y_t!} \tag{8}$$

$$= \frac{\exp(\boldsymbol{\beta}\boldsymbol{X}_t y_t)\exp(-\exp(\boldsymbol{\beta}\boldsymbol{X}_t))}{y_t!}. \tag{9}$$

The estimation of the vector $\boldsymbol{\beta}$ traditionally relies on the maximum likelihood method [10]. As the score equation is nonlinear in respect of the model parameters, the estimator is not analytically tractable and will have to be found using iterative optimization. Typically, iterative reweighted least squares (IRLS) relying on Newton-Raphson procedure are used. This and other MLE-oriented approaches are discussed, e.g., in [19] and [5, Chapter 3.1]. Alternatively, it is possible to infer $\boldsymbol{\beta}$ by means of the Bayesian framework. This domain clearly relies on expensive Markov chain Monte Carlo (MCMC) methods [19]. In either case, the estimation is

- *offline*, i.e., it requires the batch of data and does not allow for sequential updating of estimates by new observations;
- intrinsically *computationally expensive* due the iterative character of existing methods.

In order to perform a true sequential inference of $\boldsymbol{\beta}$ from continually acquired observations, we may take advantage of the method suggested by the second author in [7] and originally inspired by [8]. The transformation, ascribed to Bartlett and Kendall [2], is that any probability density function of the form

$$\theta^{y-1}\exp(-\theta) \tag{10}$$

is approximately proportional to

$$\exp\left(\frac{y}{2}\left(\log(\theta)-\log(y)\right)^2\right), \tag{11}$$

see [8]. This allows to rewrite the densities proportional to (10) as the densities of $N(logy, y^{-1})$. This approximation works relatively well for y large enough. The publication [7] of the second author additionally considers a calibration suppressing the approximation error for low values of y. Since the approximation does not admit zero values of y_t, we adopt the principle presented in Section 2 and add a small constant to it. The resulting variable $\tilde{y}_t = y_t + c$ now allows us to rewrite the Poisson GLM as

$$p(y_t|\mathscr{F}_{t-1}) \propto \exp(\boldsymbol{\beta}\boldsymbol{X}_t\tilde{y}_t)\exp(-\exp(\boldsymbol{\beta}\boldsymbol{X}_t)) \tag{12}$$

$$\approx \sqrt{\frac{\tilde{y}_t}{2\pi}}\exp\left(-\frac{\tilde{y}_t}{2}\left(\log(\tilde{y}_t)-\boldsymbol{\beta}\boldsymbol{X}_t\right)^2\right). \tag{13}$$

It is well known that a convenient prior distribution for $\boldsymbol{\beta}$ obeying the Gaussian likelihood (13) with known variance is a $p+1$-dimensional Gaussian distribution,

$$\pi(\boldsymbol{\beta}|\mathbb{F}_{t-1}) = \frac{1}{\sqrt{(2\pi)^{p+1}|\boldsymbol{\Sigma}_{t-1}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{t-1})\boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{t-1})^{\intercal}\right), \quad (14)$$

where $\hat{\boldsymbol{\beta}}_{t-1} \in \mathbb{R}^{p+1}$ is the prior point estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{(p+1)\times(p+1)}$ is the covariance matrix. The filtration

$$\mathbb{F}_{t-1} = (\mathscr{F}_{\tau})_{\tau=0,\ldots,t-1} \quad (15)$$

encompasses all information about the process up to time $t-1$.

The posterior distribution resulting from the Bayes' theorem reads

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\mathbb{F}_t) &\propto p(y_t|\mathscr{F}_{t-1})\pi(\boldsymbol{\beta}|\mathbb{F}_{t-1}) && (16)\\
&\propto \sqrt{\frac{y_t}{2\pi}} \exp\left(-\frac{y_t}{2}\left(\log(y_t)-\boldsymbol{\beta}\boldsymbol{X}_t\right)^2\right)\\
&\times \frac{1}{\sqrt{(2\pi)^{p+1}|\boldsymbol{\Sigma}_{t-1}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{t-1})\boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{t-1})^{\intercal}\right) && (17)\\
&\propto \frac{1}{\sqrt{(2\pi)^{p+1}|\boldsymbol{\Sigma}_t|}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_t)\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_t)^{\intercal}\right), && (18)
\end{aligned}$$

where the updated hyperparameters $\hat{\boldsymbol{\beta}}_t$ and $\boldsymbol{\Sigma}_t$ follow from a simple algebra. They are given by the formulas

$$\boldsymbol{\Sigma}_t = \left(\tilde{y}_t \boldsymbol{X}_t \boldsymbol{X}_t^{\intercal} + \boldsymbol{\Sigma}_{t-1}^{-1}\right)^{-1}, \quad (19)$$

$$\hat{\boldsymbol{\beta}}_t = \boldsymbol{\Sigma}_t\left(\boldsymbol{\Sigma}_{t-1}^{-1}\hat{\boldsymbol{\beta}}_{t-1} + \tilde{y}_t \log(\tilde{y}_t)\boldsymbol{X}_t\right). \quad (20)$$

The following sections contain application of the above theory to simulated and real data to see whether choice of modifications (7) and/or constant c affects estimates of hyperparameters $\boldsymbol{\beta}$ and predictions.

4 Example I: Simulated data

In this chapter we focus on performance of Poisson autoregressive model (3) of order $p=2$ for 3 modifications given in (7) and the following values of constant c:

$$\text{Model 1 } [c=1]: \ \nu_t = \beta_0 + \beta_1 \log(y_{t-1}+1) + \beta_2 \log(y_{t-2}+1) \quad (21)$$

$$\text{Model 2 } [c=1]: \ \nu_t = \beta_0 + \beta_1 \log(\max(y_{t-1},1)) + \beta_2 \log(\max(y_{t-2},1)) \quad (22)$$

	β_0	β_1	β_2
Model 1	0.233	0.083	0.073
Model 2	0.103	0.05	0.02
Model 3	0.071	0.039	0.022

Table 1 Simulated data example: RMSE of estimates of unknown hyperparameters $\boldsymbol{\beta}$ against preset values $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (1.1, 0.5, 0.1)$.

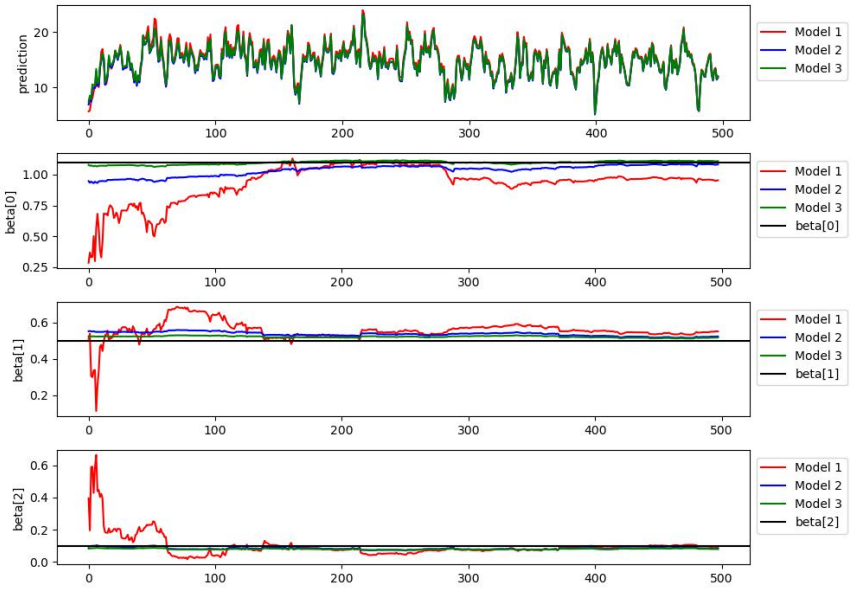

Fig. 1 Simulated data example. Top: Predictions given by models (21), (22), (23). Bottom three plots depict the evolution of estimates of $(\beta_0, \beta_1, \beta_2)$ plotted against values $(1.1, 0.5, 0.1)$ used to generate observations from autoregressive Poisson process of 2^{nd} order.

$$\text{Model 3 } [c = e^{-10}]: \; \nu_t = \beta_0 + \beta_1 \log(\max(y_{t-1}, e^{-10})) + \beta_2 \log(\max(y_{t-2}, e^{-10})) \tag{23}$$

To demonstrate effectiveness of model (3) with modifications (7) we generated first two observations from Poisson distribution with $\lambda = 5$ and the rest of the data from Poisson distribution with updated value of the rate parameter according to modifications (21), (22), (23).

Figure 1 shows the predictions and comparison of estimated values $\hat{\boldsymbol{\beta}}$ to the original values of hyperparameters $\boldsymbol{\beta}$ based on models (21), (22), (23). The prediction trendline as well as the root mean square errors (RMSE) for estimates of $\boldsymbol{\beta}$ are similar for all three models with Model 2 and Model 3 having slightly lower values of RMSE, see Table 1. However, Model 1 estimates only stabilise around time step 150 while the estimates based on Model 2 and Model 3 stabilise early on. Although all three models performed well in terms of predictions, the behaviour of regression co-

	Model 1	Model 2	Model 3
RMSE (whole dataset)	45.26	45.86	45.36
RMSE (train/test)	3.98	3.26	4.36

Table 2 Real data example: RMSE of predictions based on COVID-19 data from the whole year and on testing set when splitting into training set and testing set.

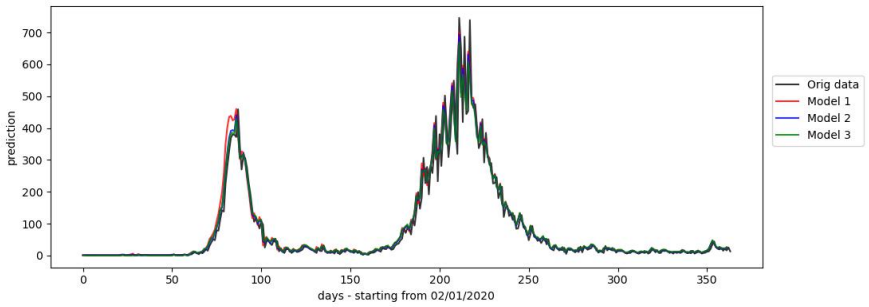

Fig. 2 Black line: numbers of new confirmed COVID-19 cases - raw data. Coloured lines: predictions using models (21), (22), (23) for the whole year.

efficients estimates especially for Model 1 (with possible constraints on the values) is of interest in future research.

5 Example II: Real time series

To demonstrate effectiveness of models (21), (22), (23) on real time series we use data from COVID-19 Google Open Dataset (COD) https://health.google.com/covid-19/open-data/, i.e., we use daily numbers of new confirmed cases in Australia from the first year (365 days) starting from 01/01/2020. Example of another approach dealing with COVID-19 data using Poisson AR model can be found in [1].

5.1 Whole dataset

In this section we let our models run on the whole set of 365 data. Figure 2 shows the original data and trendlines obtained – produced trendlines are very similar and resemble the raw data trendline. RMSE for all three models are very similar, see Table 2.

Figure 3 shows the estimates of unknown $\boldsymbol{\beta}$ at each time step. While the estimates based on Model 2 and Model 3 stabilise early, estimates based on Model 1 only stabilise after $2/3$ of the data has been processed.

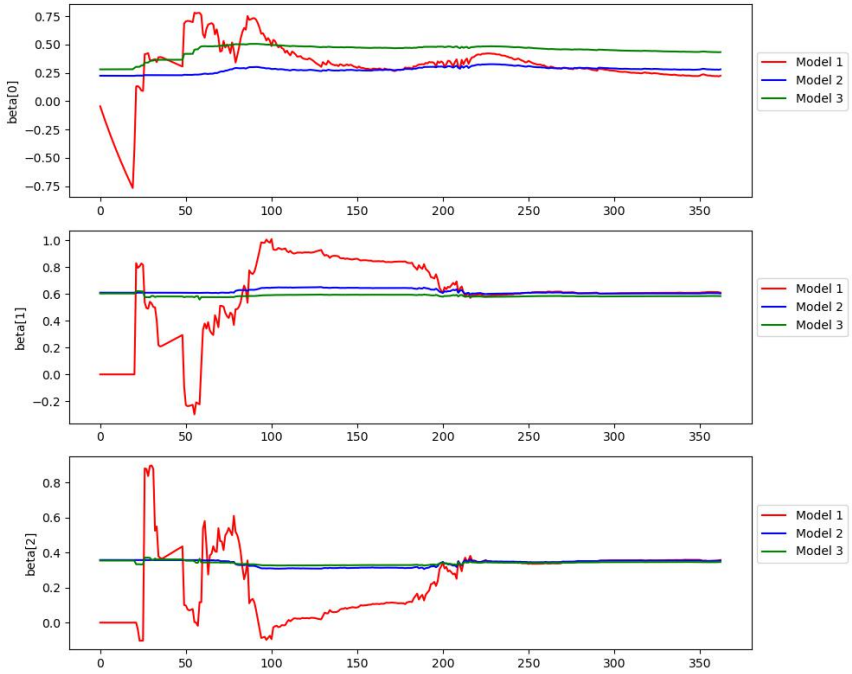

Fig. 3 Estimates of hyperparameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ based on modifications (21), (22), (23) for COVID-19 data from Australia (whole year starting from 01/01/2020).

This is a potential issue especially if we were to use machine learning techniques for $\boldsymbol{\beta}$ estimation in future, e.g., splitting dataset into training part and testing part - we investigate this in the next subsection.

5.2 Training and testing set

In this section we divide COVID-19 dataset into two disjoint subsets - a training set and a testing set. We let all three models run on the first subset of data (training set); last estimated value of hyperpameters $\boldsymbol{\beta}$ is then taken and used to obtain predictions. We then compare the predictions to appropriate subset of raw data (testing set).

Firstly, we use first 80% of data as a training set to obtain an estimate for unknown $\boldsymbol{\beta}$. Since the estimates, especially in Model 1, are stabilised by this time point the predictions for all three considered models are very similar to the values of raw data, see Figure 4.

All three considered modifications perform well in terms of predictions; their RMSE are quite small and not significantly different, see Table 2; however the behaviour of hyperparameter (regression coefficients) estimates deserves further in-

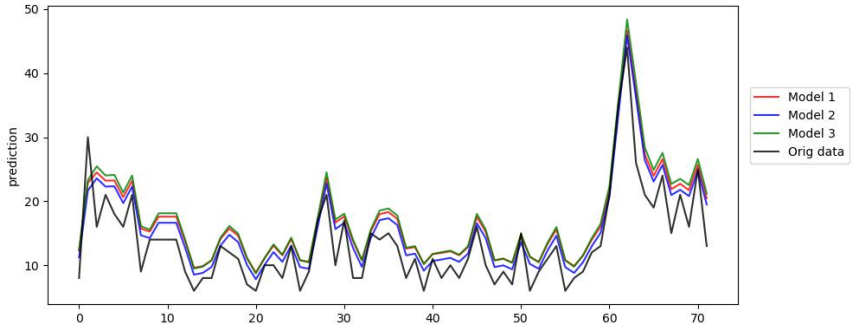

Fig. 4 Predictions based on splitting of the real data into: first 80% training data, the rest 20% testing data.

vestigation with introduction of possible constraints on their values and perhaps inclusion of new modifications.

6 Conclusion and future work

We have presented our initial results in the domain of sequential modeling of count data. This domain becomes increasingly popular due to the omnipresence of count data, and the rapid development of cheap smart appliances, Internet of Things etc. The examples demonstrate that the proposed method is able to reliably estimate the true model parameters. The example focused on a real time series then shows that the method is applicable to real-world problems.

There is a huge research potential in the domain. In particular, we will focus on two issues. First, the real-world count data often contain more zeros than prescribed by the Poisson model. This issue is known as zero-inflation [11]. Second, the equivariance assumption is rather limiting in practical applications. In the GLM domain, there exist more apt models for under- and overdispersion, e.g., the negative binomial or the generalized Poisson models [12, 13].

References

1. Agosto, A., Campmas, A., Giudici, P., Renda, A.: Monitoring COVID-19 contagion growth. Statistics in Medicine **40**, 4150–4160 (2021).
2. Bartlett, M.S., Kendall, D.G.: The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation. Supplement to the Journal of the Royal Statistical Society **8**(1), 128–138 (1946).
3. Billings, S.A.: Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains. John Wiley & Sons (2013)

4. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control, 5 ed. Wiley, Hoboken, NJ, USA (2015)
5. Cameron, A.C., Trivedi, P.K.: Regression Analysis of Count Data. Cambridge University Press (2013).
6. Davis, R. A., Fokianos, K., Holan, S. H., Joe, H., Livsey, J., Lund, R., Pipiras, V., Ravishanker, N.: Count Time Series: A Methodological Review, Journal of the American Statistical Association **116**(535), 1533–1547 (2021).
7. Dedecius, K., Zemlicka, R.: Sequential Poisson Regression in Diffusion Networks. IEEE Signal Processing Letters **27**(1), 625–629 (2020).
8. El-Sayyad, G.M.: Bayesian and Classical Analysis of Poisson Regression. Journal of the Royal Statistical Society: Series B (Methodological) **35**(3), 445–451 (1973).
9. Engle, R.F.: Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica **50**(4), 987–1007 (1982).
10. Fokianos, K., Tjøstheim, D.: Log-Linear Poisson Autoregression. Journal of Multivariate Analysis **102**(3), 563–578 (2011).
11. Haslett, J., Parnell, A.C., Hinde, J., Andrade Moral, R.: Modelling Excess Zeros in Count Data: A New Perspective on Modelling Approaches. International Statistical Review pp. 1–21 (2021).
12. Hilbe, J.M.: Negative Binomial Regression. Cambridge University Press, Cambridge (2007).
13. Hilbe, J.M.: Modeling Count Data. Cambridge University Press (2014).
14. Kedem, B., Fokianos, K.: Regression Models for Time Series Analysis. Wiley (2005).
15. Koop, G.: Bayesian Econometrics. Wiley (2003).
16. Liboschik, T., Fokianos, K., Fried, R.: tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. Journal of Statistical Software **82**(5), 1–51 (2017).
17. Ljung, L., Glad, T.: Modeling and Identification of Dynamic Systems. Prentice Hall (1994).
18. McCullagh, P., Nelder, J.A.: Generalized Linear Models, Second Edition. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall (1989).
19. Myers, R.H., Montgomery, D.C., Vining, G.G., Robinson, T.J.: Generalized Linear Models. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA (2010).
20. Nelder, J.A., Wedderburn, R.W.M.: Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General) **135**(3), 370 (1972).
21. Pollock, D.S.G.: Handbook of Time-Series Analysis, Signal Processing and Dynamics. Academic Press (1999)
22. Reinsel, G.C.: Elements of Multivariate Time Series Analysis. Springer (2003).
23. Zeger, S.L., Qaqish, B.: Markov Regression Models for Time Series: A Quasi-Likelihood Approach. Biometrics **44**(4), 1019–1031 (1988).