The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 359-364

Model Checking for Time-Series Count Data[1]

Sungim Lee[2]

Abstract

This paper considers a specification test of conditional Poisson regression model for time series count data. Although conditional models for count data have received attention and proposed in several ways, few studies focused on checking its adequacy. Motivated by the test of martingale difference assumption, a specification test via Ljung-Box statistic is proposed in the conditional model of the time series count data. In order to illustrate the performance of Ljung-Box test, simulation results will be provided.

Keywords : Time-series count data, conditional model, Martingale difference test, Ljung-Box test

1. Introduction

In the modeling of time series count data, conditional models for y_t given $y_{t-1},\ldots,y_1$ and x_t have been considered by several authors such as Wong(1986), Zeger and Qaquish (1988), and Fokianos (2000). In general, there has been two approaches to the problem of building models from those data. One is conditional models which assume only a correctly specified conditional mean $E(y_t|y_{t-1},\ldots,y_1,x_t)$ together with some appropriate variance and autocorrelation structure. The other is marginal models conditioning only on x_t and not on past outcomes (cf. Fahrmeir and Tutz (2000)). Brumback et al.(2000) unify these two different approaches in terms of the extended generalized linear models (GLMs). For a variety of different models applied in their work, Choi et al. (2003) considered the model selection criteria to select proper model.

Although there has been various conditional models for autocorrelated data, a few studies have applied on a test for the appropriateness a specified model. When we assume the conditional model for nonnormal time series data, a correctly specified conditional mean becomes key part of the model assumption like the standard GLM. In particular, since the use of past outcomes as predictors could explain possible autocorrelation to some degrees, even if

[1] This work was supported by Dankook University Research Fund of 2004.

[2] Full-time lecturer, Department of Information Statistics, Dankook University, Seoul 140-714, Korea
E-mail : silee@dankook.ac.kr

it is *ad hoc* approach, a model's adequacy can be assessed by testing the assumption that the data were generated by the specified conditional mean.

In this paper, following the idea of martingale difference assumption, we will consider Ljung-Box statistic as a test statistic which checks if a conditional mean is correctly specified.

The remainder of this paper is organized as follows. Section 2 defines the Poisson regression model for time series count data and introduces the null hypothesis and defines the test statistic. In Section 3, the simulation results are provided for the performance of test statistics. Conclusions are given in Section 4.

2. Model checking of Poisson regression model for time series count data

We consider a conditional Poisson regression model for time series count data as a extended GLM which has past outcome as predictors in order to explain possible autocorrelation. As it is previously well-explored by Brumback et al.(2000), we assume that the conditional expectation $\mu_t = E(y_t|H_t)$, $t=1,2,\cdots,T$ is of the form

$$\mu_t = \exp(x_t'\beta + H_t'\alpha) \;, \quad Var(y_t|H_t) = \mu_t \tag{2.1}$$

where $H_t = \{x_t, x_{t-1}, \cdots, x_1, y_{t-1}, y_{t-2}, \cdots, y_1\}$ is the history of past observations and of present and past covariates at time t. This model in (2.1) is formally identical to that of generalized linear model for independent observations. When we consider the conditional first moment of y_t appeared in (2.1), we have

$$E(y_t - \mu_t|H_t) = 0. \tag{2.2}$$

Let us define $u_t = y_t - \mu_t$. Then, u_t satisfies a martingale difference sequence process given H_t. A martingale difference sequence is defined as a process that has constant mean (usually zero) given some information set which typically includes just its past values. Let $\hat{\beta}$ and $\hat{\alpha}$ be the conditional least squares estimator of β and α. The estimation can be carried out using a iteratively reweighted least squares algorithm. For details related to estimation, refer to Section 3 of Brumback et al.(2000) and references cited therein. If the conditional mean in (2.2) is correctly specified, we can expect that the residual, $\hat{u}_t = y_t - \hat{\mu}_t$ satisfy martingale difference assumption. This implies that the Poisson time series data can be modeled appropriately. As dealt with in Econometrics (Durlauf (1991), Anderson (1993), and Hong (1996), etc.), the common way of testing this property has been testing that the process is uncorrelated. Hence, the test statistic typically employed has been based on the sample autocorrelations. In this paper, we will employ the Ljung-Box test (cf. Ljung and Box, 1978) for testing that a process u_t is uncorrelated.

Therefore, the considered null hypothesis is represented by

$$H_0:\ u_t\text{'s are uncorrelated.} \tag{2.3}$$

Based on the residual $\hat{u}_t$ we can calculate the residual autocorrelation at the lag h as

$$\hat{r}(h)=\frac{\sum_{t=1}^{n-h}(\hat{u}_t-\bar{\hat{u}})(\hat{u}_{t+h}-\bar{\hat{u}})}{\sum_{t=1}^{n}(\hat{u}_t-\bar{\hat{u}})^2}\ ,\qquad h=1,2,\cdots,k \tag{2.4}$$

where $\bar{\hat{u}}=\sum_{t=1}^{n}\hat{u}_t/n$ and k is a positive integer. As similarly shown in Kim et al.(2004), the Ljung-Box test statistic

$$Q_n(k)=n(n+2)\sum_{h=1}^{k}(n-h)^{-1}r^2(h) \tag{2.5}$$

is approximately $\chi^2(k)$ under the model with the assumption in (2.3). The following theorem ensures the result in (2.5).

Theorem. Suppose that $\{y_t\}$ satisfies that

$$y_t-\mu_t=u_t,\ \ \mu_t=\exp(x_t'\beta+H_t\alpha)$$

where $\{u_t\}$ forms a stationary martingale difference sequence with $Var(u_t|H_t)=\mu_t<\infty$. If $\hat{\beta}$ and $\hat{\alpha}$ are an estimator of β and α such that

$$\sqrt{n}(\hat{\beta}-\beta)=O_p(1),\ \ \sqrt{n}(\hat{\alpha}-\alpha)=O_p(1)$$

then

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}\hat{u}_t\,\hat{u}_{t+h}=\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}u_t\,u_{t+h}+o_p(1)$$

where $\hat{u}_t=y_t-\exp(x_t'\hat{\beta}+H_t'\hat{\alpha})$.

Proof. For the proofs, we only show that the terms in (i), (ii), and (iii) are $o_p(1)$.

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}\hat{u}_t\,\hat{u}_{t+h} &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}(y_t-\hat{\mu}_t)(y_{t+h}-\hat{\mu}_{t+h})\\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}\big((y_t-\mu_t)+\mu_t-\hat{\mu}_t\big)\big((y_{t+h}-\mu_{t+h})+\mu_{t+h}-\hat{\mu}_{t+h}\big)\\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}(y_t-\mu_t)(y_{t+h}-\mu_{t+h})\\
&\qquad +\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}(y_t-\mu_t)(\mu_{t+h}-\hat{\mu}_{t+h})\quad (i)\\
&\qquad +\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}(y_{t+h}-\mu_{t+h})(\mu_t-\hat{\mu}_t)\quad (ii)\\
&\qquad +\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}(\mu_t-\hat{\mu}_t)(\mu_{t+h}-\hat{\mu}_{t+h})\quad (iii)\,.
\end{aligned}$$

Take $\sqrt{\mu_t}\varepsilon_t = y_t - \mu_t$. Then, the term in (i) can be rewritten as

$$(i) = \frac{1}{n}\sum_{t=1}^{n-h}\varepsilon_t\sqrt{\mu_t}\cdot\sqrt{n}(\hat{\mu}_t - \mu_t).$$

Here,

$$\begin{aligned}\sqrt{n}(\hat{\mu}_t - \mu_t) &= \sqrt{n}\{x_t(\hat{\beta}-\beta) + H_t(\hat{\alpha}-\alpha)\}\mu_t \\ &\quad + \frac{1}{2}\{x_t(\hat{\beta}-\beta) + H_t(\hat{\alpha}-\alpha)\}^2\mu_t + \cdots\end{aligned}$$

For bounded regressors, we can prove that $\sup_t\sqrt{n}(\hat{\mu}_t - \mu_t) = O_p(1)$. Then, term (i) converges to zero as $n\to\infty$. Similarly as in (i), the terms (ii) and (iii) can be shown to be $o_p(1)$.

3. Simulation studies

In this section, we will investigate the performance of the Ljung-Box statistic in (2.5) through simulations. We consider first the true model (2.1) with $H_t = y_{t-1}$, that is,

$$y_t = \mu_t + \sqrt{\mu_t}u_t, \tag{3.1}$$

$$\mu_t = \exp(0.3x_t + 0.1y_{t-1}) \tag{3.2}$$

where u_t is iid observations from N(0,1) and x_t is generated from uniform (0,2).

For the empirical size and power at the significance level $\alpha = 0.05$ of test statistic in (2.5), 1,000 samples of size 100, 200, and 500 are generated from the model in (3.1). In each simulation, 200 observations are discarded to remove initialization effects. Five values for k 3, 10, 15, 20, and 30, were considered. For the power of the test, the data is generated by

$$\mu_t = \exp(0.3x_t + 0.1y_{t-1} - 0.2y_{t-2}). \tag{3.3}$$

As a result, Table 1 shows the significance level and power of the test statistics for the three kinds of sample size. We can see that for fixed k the size distortions tend to decrease as n increases whereas the observed significance level is close to 0.05 at $k=20$ for each n In particular, the significance levels show less variability for $k=15$ or 20. Concerning the power of the test, we observe that it yields good powers as n increases.

So far, we have investigated the model in (2.1) with H_t which include lagged values of y_t as covariates and consider the power when the data are generated the different lagged values. In order to analyze the performance of the power against the different alternative models which replace the past outcome in covariates by residual type, we will calculate the test powers for the following models.

<Table 1> Empirical sizes under the model (3.2) and powers of $Q_n(k)$ when the data are generated from the model (3.3) and the model (3.2) is fitted.

	Empirical size					Empirical power				
k	3	10	15	20	30	3	10	15	20	30
$n=100$	0.022	0.039	0.046	0.054	0.070	0.500	0.335	0.303	0.285	0.281
$n=200$	0.030	0.038	0.047	0.051	0.058	0.852	0.659	0.581	0.529	0.472
$n=500$	0.029	0.048	0.044	0.052	0.057	0.965	0.883	0.826	0.765	0.706

$$\text{Model 1.}\quad \mu_t=\exp(0.3x_t+0.1(y_{t-1}-\exp(x_{t-1}\beta))). \tag{3.4}$$

$$\text{Model 2.}\quad \mu_t=\exp(0.3x_t+0.1(\log(\max(y_{t-1},0.1))-x_{t-1}\beta)). \tag{3.5}$$

<Table 2> Powers of the test when the data are generated from Models 1 and 2 in (3.4)-(3.5) and in each case, Model (3.2) is fitted with n=200 and $\alpha=0.05$

k	3	10	15	20	30
Model 1	0.582	0.555	0.544	0.541	0.539
Model 2	0.563	0.530	0.521	0.515	0.513

Table 2 shows the power of test statistics applied to the residual autocorrelations from the models (3.4) and (3.5) after fitting an model (3.2). Compared to Table 1, we can see that overall they yield similar powers.

4. Conclusions

As previously investigated by Pena and Rodriguez (2002) in ARMA(p,q), the significance level and power of Ljung-Box statistic tend to be somewhat different according to the degrees of autocorrelation. In the simulation studies of this paper, we use rather weak correlation between the present and past outcome. Hence, considering the property of Ljung-Box test, our results seems quite encouraging. In addition, we need to note that the test statistic yields reasonable powers even if the model in (3.2) has much similarities with the models in (3.4) and (3.5). From our results we can conclude that the Ljung-Box test can be employed as an appropriate diagnostic tool for model specification test although the proof of asymptotics (2.5) needs to be more rigorously derived. We will consider the proof where growing regressors will be of interest without rather strong assumption of boundness in future study. Also, it would be worthwhile to extend this test to the wider class of GLMs.

References

[1] Kim, E.H., Ha, J.C., Jeon, Y.S., and Lee, S.Y.(2004). Ljung-Box test in unit root AR-ARCH model. *The Korean Communications in Statistics* vol. 11. No.2. 323-327.

[2] Choi, Y.H., Lee, S., and Lee, S.Y.(2003). Generalized liner model with time series data, *The Korean Journal of Applied Satistics*, vol. 16, 365-376.

[3] Anderson, T.W. (1993), Goodness of fit tests for spectral distributions, *Annals of Statistics*, vol. 21, 830-847.

[4] Brumback, B.A., Ryan, L.M., Schwartz, J.D., Neas, L.M., Stark, P.C, and Burge, H.A. (2000). Transitional regression models, with application to environmental time series. *Journal of American Statistical Association*, vol. 95, 16-27.

[5] Durlauf, S. N. (1991), Spectral based testing of the martingale hypothesis, *Journal of Econometrics*, vol. 50, 355-376.

[6] Fahrmeir, L., and Tutz, G.(2001). *Multivariate statistical modelling based on generalized linear models*, New-York: Springer-Verlag.

[7] Fokianos, K. (2000). Truncated poisson regression for time series of counts. *Scandinavian Journal of Statistics*. vol. 28. 645-659.

[8] Hong, Y. (1996). Consistent testing for serial correlation of unknown form. *Econometrica*, vol. 64, 837-864.

[9] Ljung, G.M. and Box, G.E.P.(1978). On a measure of lack of fit in time series models. *Biometrika*, vol. 65, 297-303.

[10] Pena, D. and Rodriguez, Julio. (2002). A powerful portmanteau test of lack of fit for time series. *Journal of American Statistical Association*, Vol. 97, 601-610.

[11] Wong, W. H. (1986). Theory of partial likelihood. *Annals of Statistics*, vol. 14, 88-123.

[12] Zeger, S. L. and Qaquish, B. (1988). Markov regression models for time series: A Quasi-Likelihood Approach. *Biometrics*, Vol. 44, 1019-1031.

[Received January 2005, Accepted March 2005]