Iraqi Journal of Statistical Sciences (22)2012
PP[15-25]

Re-sampling Techniques in Count Data Regression Models

Zakariya Y. Algamal [*] Khairy B. Rasheed[**]

Abstract

Modeling count variables is a common task in many application areas such as economics, social sciences, and medicine. The classical Poisson regression model for count data is often used and it is limited in these disciplines since count data sets typically exhibit overdispersion, so negative binomial regression can be used. We use a jackknife- after-bootstrap procedure to assess the error in the bootstrap estimated parameters. The method is illustrated through two real examples. The results suggest that the jackknife- after- bootstrap method provides a reliable alternative to traditional methods particularly in small to moderate samples.

Keywords: Poisson regression, Overdispersion, Negative binomial regression, Bootstrap, Jackknife- after- Bootstrap

أساليب إعادة المعاينة في نماذج انحدار بيانات العد

المستخلص :

تعد عملية نمذجة المتغيرات القابلة للعد من المهام المهمة في العديد من المجالات منها الاقتصادية والعلوم الاجتماعية والطبية . غالبا ما يستخدم نموذج انحدار بواسون لنمذجة مثل هذا النوع من البيانات ويكون هذا النموذج غير ملائم عندما يعاني من مشكلة (Overdispersion)وعليه سوف يستخدم نموذج انحدار ثنائي الحدين السالب .

[*] Lecturer / Statistics and Informatics Dept. / Computer Science and Mathematics college / Mosul University / Mosul / IRAQ. E-mail: zak.sm_stat@yahoo.com

[**] Lecturer / Statistics and Informatics Dept. / Computer Science and Mathematics college / Mosul University / Mosul / IRAQ. E-mail: Khairy_line@yahoo.com

Received Date 29/5/2011 Accept Date 1/3/2012

وقد استخدمنا في هذا البحث أسلوب (Jackknife- after- Bootstrap) لتقييم الخطأ الذي يحصل عند تقدير المعلمات باستخدام الـ (Bootstrap) في انحدار بواسون ، إذ استخدمنا مجموعتين من البيانات الحقيقية لتوضيح الأسلوب المستخدم وقد وضحت النتائج بان استخدام أسلوب (Jackknife- after- Bootstrap) يمكن التعويل عليه مقارنة بالطرائق التقليدية وخاصة عند أحجام العينات الصغيرة والمتوسطة .

1- Introduction

Researchers spend much of their time counting things, numbers of symptoms, placements, and so on. Count variables indicate the number of times a particular event occurs to each case such as number of hospital visits per year , number of divorces per city (Orme & Combs-Orme, 2009). Count variable is an integer and can range from 0 through $+\infty$. Two common distributions are used often to model the count variable; they are Poisson and negative binomial distributions.

When the response variable (y) is a count variable and we fit the linear regression model using ordinary least squares (OLS) method, then we may have several problems. First, the usual assumption that the errors are normally distributed fails, since (y) is typically non normal. Second, OLS estimators also assume a homoscedastic error structure, this is problematic if (y) is a count variable. Third, if the errors are really hetroscedastic, the standard error estimates produced by OLS are biased (Demairs, 2004). Jackknife and bootstrap re-sampling techniques are designed to estimate standard errors, bias, confidence intervals, and prediction error. The bootstrap is a re-sampling method that draws a large collection of samples from the original data. It is used to select the observation randomly with replacement from the original data sample, and jackknife is generated by sequentially deleting single datum from the original sample (Efron and Tibshirani, 1993). We use a jackknife- after-bootstrap procedure to assess the error in the bootstrap estimated parameters. The method is illustrated through two real examples. In

sections 2, 3, and 4 we described the Poisson regression model, negative binomial regression model, and overdispersion, respectively. In section 5, the use of the jackknife-after-bootstrap was discussed. The analytical examples are given in section 6 where two real data sets were used. Finally, section 7 shows the conclusions.

2- Poisson Regression Model

Poisson Regression Model (PRM) is a technique which allows to model response variable that describs count data. It is often applied to study the occurrence of small number of counts as a function of a set of explanatory variables (Cameron & Trivedi, 1998). The PRM relates the probability function of a response variable (y) to a vector of explanatory variables (x) (Winkelmann , 2008) , more formally , the PRM assumes that the response variable (y) drawn from a Poisson distribution with mean and variance (μ) . The p.d.f of (y) is :

$$f(y_i \mid \mu) = \frac{e^{-\mu}\mu^{y_i}}{y_i!} \quad , y_i = 0,1,2,.... \quad ...(1)$$

The Poisson distribution is unimodal and skewed to the right over the possible values 0,1,2,… . It has a single parameter $\mu > 0$, which is both its mean and its variance , that is (known as equidispersion) (Agersti,2006) :

$$E(y_i \mid x) = Var(y_i \mid x) = \mu \quad ....(2)$$
$$\text{or } E(y) = Var(y) = \mu$$

With PRM the mean μ is explained in terms of explanatory variables (x) via an appropriate link function. The popular choice for the link function is the log link, that is:

$$\mu = E(y_i \mid x) = Exp(x'\beta) \quad ...(3)$$

Where (β) is a ($k*1$) vector of parameters, and (x) is a ($k*1$) vector of explanatory variables. Taking the exponential of ($x\beta$) forces (μ) to be positive which is necessary since count only ($\beta_0 = 0$) or positive (Long &

Freese, 2001) (De Jong & Heller, 2008). So, the multiple PRM can be written as:

$$\hat{\mu} = \mathrm{Exp}(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k) \quad ...(4)$$

or equivalently:

$$\ln \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k \quad ...(5)$$

The parameters (β) can be estimated by using the maximum likelihood method (m.l.). The standard error of the estimated parameters is:

$$Se(\hat{\beta}_{PRM}) = \{\hat{\sigma}^2 [\sum_{i=1}^{n} \mathrm{Exp}(x'\hat{\beta}) x_i' x_i]^{-1}\}^{1/2} \quad ...(6)$$

3- Overdispersion and Underdispersion

The key assumption of the PRM is that the conditional mean equals the conditional variance i.e. $E(y_i \mid x) = Var(y_i \mid x)$.In many applications this assumption has not met. If $E(y_i \mid x) < Var(y_i \mid x)$, respectively $E(y_i \mid x) > Var(y_i \mid x)$,then we speak about overdispersion, respectively underdispersion. The PRM does not allow for overdispersion (Cameron & Trivedi , 1998)

4- Negative Binomial Regression Model

The negative binomial regression model (NBRM) is the most commonly used alternative to the (PRM) when it has overdispersion problem (Winkelmann, 2008).

Under the Poisson distribution, the mean, μ_i , is assumed to be constant or homogeneous within the class . By assuming the specific distribution for (μ_i) to be a gamma with mean $E(\mu_i) = \theta_i$ and variance $Var(\mu_i) = \theta_i^2 v_i^{-1}$, and $y_i \mid \mu_i$ to be a poisson with conditional mean $E(y_i \mid \mu_i) = \mu_i$, it can be shown that the marginal distribution of y_i follows a negative binomial distribution with p.d.f :

$$f(y_i)=\frac{\Gamma(y_i+v_i)}{\Gamma(y_i+1)\Gamma v_i}\left(\frac{v_i}{v_i+\theta_i}\right)^{v_i}\left(\frac{\theta_i}{v_i+\theta_i}\right)^{y_i} \quad ...(7)$$

where the mean is $E(y_i)=\theta$ and the variance is $Var(y_i)=\theta_i+\theta_i^2 v_i^{-1}$, This is called the negative binomial I .

From regression analysis of count data the most common implementation of the negative binomial is called negative binomial II model (NB_2). By letting $v_i=\theta_i a^{-1}$, this time with mean $E(y_i)=\theta_i$ and variance $Var(y_i)=\theta_i(1+a)$. If a equals zero , then the mean and variance will be equal , resulting the distribution to be a poisson . If $(a>0)$, the variance will exceed the mean and the distribution allows for overdispersion as well . So, the p.d.f. is:

$$f(y_i)=\frac{\Gamma(y_i+\theta_i a^{-1})}{\Gamma(y_i+1)\,\Gamma(\theta_i a^{-1})}\left(\frac{a^{-1}}{1+a^{-1}}\right)^{\theta_i a^{-1}}\left(\frac{1}{1+a^{-1}}\right)^{y_i} \quad ...(8)$$

or

$$f(y_i)=\frac{\Gamma(y_i+\theta_i a^{-1})}{\Gamma(y_i+1)\,\Gamma(\theta_i a^{-1})}\left(\frac{1}{a+1}\right)^{\theta_i a^{-1}}\left(\frac{1}{1+a^{-1}}\right)^{y_i} \quad ...(9)$$

where $\theta_i=Exp(x_i'\beta)$ (Cameron & Trivedi, 1998),(Greene, 2008).

5- Jackknife after Bootstrap Procedure

The use of the bootstrap and the jackknife re-sampling methods is gradually increasing nowadays, due to increasing computer power. The basic idea of bootstrapping is to generate a large number of samples by randomly drawing observations with replacement from the original data set , and to recalculate a statistic for each bootstrap sample , whereas the jackknife is generated by sequentially deleting single datum from the original sample (Efron & Tibshirani , 1993) .

Jackknife After Bootstrap (JAB) method was proposed by (Efron,1992) to investigate the effect of a single observation in bootstrap, where Efron pointed out that the bootstrap estimates have two distinct sources of

variance, they are: sampling variability, due to the fact that we have only a sample of size n rather than the entire population, and bootstrap resampling variability, due to the fact that we take only B bootstrap samples rather than an infinite number. Suppose we have drawn (B) bootstrap samples and calculate the standard error of the regression parameter , $Se_B(\hat{\beta})$, we would like to have a measure of the uncertainty in $Se_B(\hat{\beta})$. The JAB method provides a way of estimating the standard error of $Se_B(\hat{\beta})$, $Se_{JAB}(Se_B(\hat{\beta}))$, using only information in our B bootstrap samples . The jackknife estimate of standard error of $Se_B(\hat{\beta})$ involves two steps :

For $i = 1,2,...,n$, leave out data point i and re-compute $Se_B(\hat{\beta})$ and called the results $Se_{B(i)}(\hat{\beta})$.

Define

$$Se_{JAB}(Se_B(\hat{\beta})) = [\frac{n-1}{n}\sum_{i=1}^{n}(Se_{B(i)}(\hat{\beta}) - Se_{B(.)}(\hat{\beta}))^2]^{1/2} \quad ...(10)$$

Where $Se_{B(.)}(\hat{\beta}) = \frac{\sum_{i=1}^{n} Se_{B(i)}(\hat{\beta})}{n}$ (Efron & Tibshirani , 1993)

In each i , there are some bootstrap samples in which that the data point , say x_i , does not appear , and we can use those samples to estimate $Se_{B(i)}(\hat{\beta})$. Let C_i denote the indices of the bootstrap samples that don't contain data points x_i , and there are B_i such samples , then :

$$Se_{B(i)}(\hat{\beta}) = [\frac{\sum(\hat{\beta}_B - \bar{\hat{\beta}}_B)^2}{B_i}]^{1/2} \quad ...(11)$$

$$\bar{\hat{\beta}}_B = \frac{\sum \hat{\beta}_B}{B_i}$$

6- Analytical Examples

In order to use the PRM and NBRM we deal with two real data sets . All the results done using S-plus 6.1 program.

6-1- Example 1

In this example we fit the PRM. The response variable represents the number of dead cocks. Three explanatory variables are considered , they are : age of the cocks in days , the quantity of the feed in kilogram (kg) and the temperature (AL-Suliaman, 1995). The sample size was (62), the results are shown in table (1).

Table(1):The results of PRM of the number of dead cocks .

parameters	$\hat{\beta}_{(PR)}$	S.e($\hat{\beta}$)
$\hat{\beta}_0$	0.84865017	0.8317
$\hat{\beta}_1$	-0.0325	0.01139
$\hat{\beta}_2$	0.4558	0.133
$\hat{\beta}_3$	0.0236	0.0236

The fitted PRM is :

$$\hat{y}_i = \text{Exp}[0.848 - 0.032\,(\text{age}) + 0.455\,(\text{feed}) + 0.032\,(\text{temp})]$$

The bootstrap (B) and the jackknife-after-bootstrap (JAB) results are shown in table (2).

Table (2): The Bootstrap and JAB results of example (1) (B=10,000)

Parameters	B			JAB
	$\hat{\beta}_{(B)}$	bias	S.e($\hat{\beta}_{(B)}$)	S.e($\hat{\beta}_{(JAB)}$)
$\hat{\beta}_0$	0.69375	-0.1549	1.09784	0.4005
$\hat{\beta}_1$	-0.0311	0.00137	0.01352	0.00498
$\hat{\beta}_2$	0.4459	-0.0099	0.13276	0.0499
$\hat{\beta}_3$	0.03624	0.0039	0.03021	0.01089

Figure (1) shows the observations with large influence on $s.e(\hat{\beta}_{(B)})$, where observations (17 and 25) have large influence on the intercept , observations (35 , 47) on $\hat{\beta}_{1(age)} = x_1$, observations (47 , 62) on $\hat{\beta}_{2(feed)} = x_2$, and observations (17 , 25) on $\hat{\beta}_{3(temp)} = x_3$

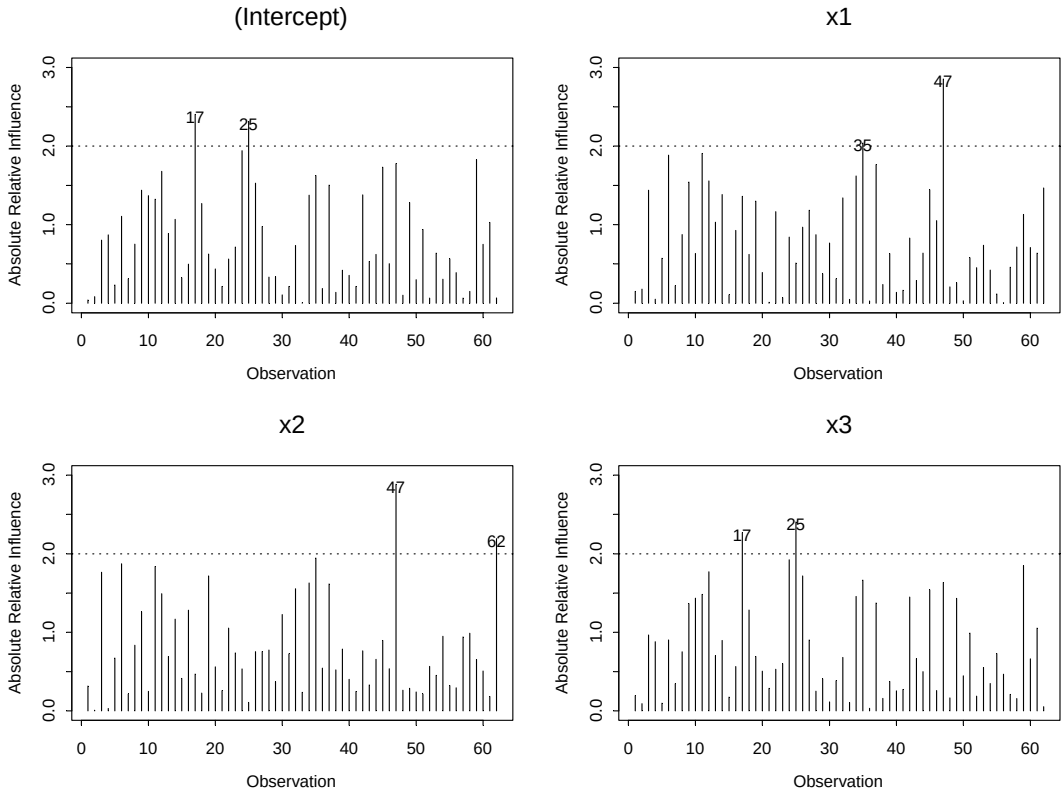

Figure (1): JAB influence of the number of dead cocks parameters.

6-2- Example 2

In this example we use gala data from faraway package (Faraway , 2006) The data describe the relationship between the number of plant species and several geographic variables is of interest, where n is 30. Species: The number of plant species found on the island, Endemics: The number of endemic species, Area: The area of the island (km $^ 2$), Elevation: The highest elevation of the island (m), Nearest: The distance from the nearest island (km), Scrnz: The distance from Santa Cruz island (km), and Adjacent: The area of the adjacent island (km^2). In this example we can't

fit the PRM because we have overdispersion problem. So the best alternative model is NBRM, table (3) shows the results of NBRM.

Table (3): The results of NBRM of the Gala example

parameters	$\hat{\beta}_{(NB)}$	S.e($\hat{\beta}$)
$\hat{\beta}_0$	2.5093	0.2058
$\hat{\beta}_1$	0.0475	0.01
$\hat{\beta}_2$	-0.0003	0.00023
$\hat{\beta}_3$	0.00022	0.00099
$\hat{\beta}_4$	-0.00312	0.0087
$\hat{\beta}_5$	0.00063	0.002
$\hat{\beta}_6$	0.0000022	0.0002

The fitted NBRM is:

$$\hat{y}_i = Exp[2.509 + 0.0475\,(Endemics) - 0.0003(Area) + 0.00022(Elevation) - 0.00312(Nearest) + 0.00063(Scrnz) + 0.0000022(Adjacent)]$$

The bootstrap (B) and the jackknife-after-bootstrap (JAB) results are shown in table (4).

Table (4): The Bootstrap and JAB results of example (2) (B=10,000)

parameters	B			JAB
	$\hat{\beta}_{(B)}$	Bias: $\hat{B}_{(B)} - \hat{B}$	S.e($\hat{\beta}_{(B)}$)	S.e($\hat{\beta}_{(JAB)}$)
$\hat{\beta}_0$	2.4024	-0.106	0.3206	0.111
$\hat{\beta}_1$	0.05762	0.01	0.0204	0.0084
$\hat{\beta}_2$	-0.0011	-0.00083	0.0015	0.0024
$\hat{\beta}_3$	0.00002273	-0.00019	0.00126	0.00045
$\hat{\beta}_4$	-0.0027	0.000357	0.013	0.0047
$\hat{\beta}_5$	0.00077	0.00013	0.003129	0.0015
$\hat{\beta}_6$	0.000018	-0.000003	0.001321	0.00175

Figure (2) shows the cases with large influence on $\mathbf{S.e(\hat{\beta}_{(B)})}$, where cases (18 and 27) have large influence on the intercept , case (27) has large influence on $\mathbf{\hat{\beta}_1(Endemics)}$, $\mathbf{\hat{\beta}_3(Elevation)}$, $\mathbf{\hat{\beta}_4(Nearest)}$, and $\mathbf{\hat{\beta}_5(Scruz)}$ are influenced by cases (19) , (15) , and (30) , respectively .

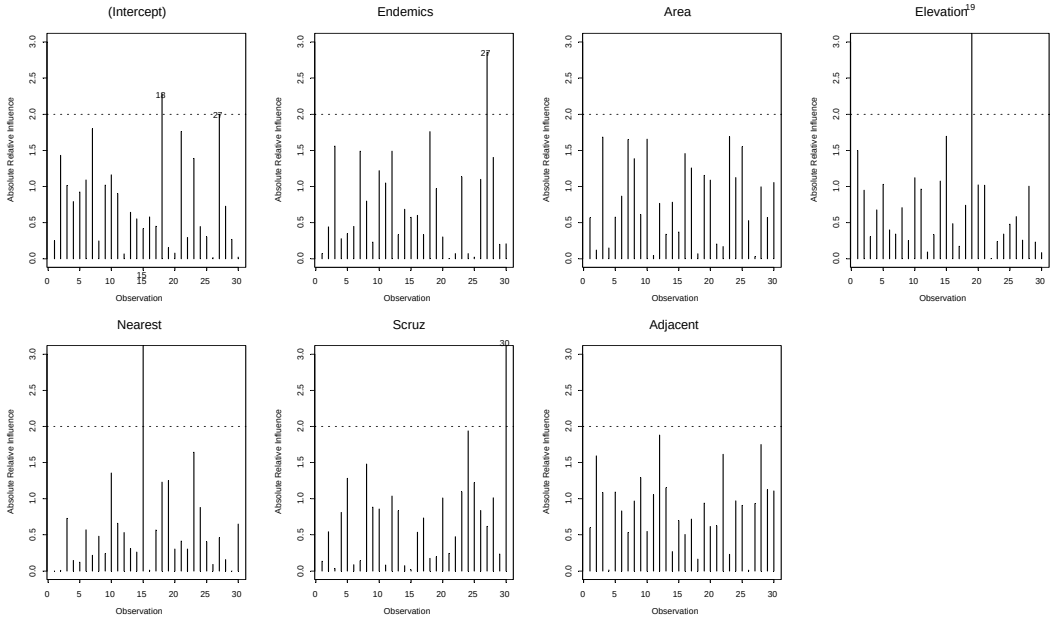

Figure (2): JAB influence of the Gala example parameters.

7- Conclusion

As a result, we conclude that the $S.e(\hat{\beta}_{(JAB)})$ is less than the $S.e(\hat{\beta}_{(B)})$ for all PRM and NBRM parameters. The results suggest that the bootstrap re-sampling provides a reliable alternative to traditional methods and JAB procedure provides a good measure of diagnosis for bootstrap. We see that from figure (1) that cases (17 and 25) have large influence on both intercept and temp., whereas the case (47) has large influence on the age and feed. The cases (35) and (62) have large influence on age and feed respectively. From figure (2), no case has large influence on the Area and Adjacent. One case has influence, 27, 19, 15, and 30 on the Endemics,

Elevation, Nearest, and Scruz, respectively. Finally, the cases (27 and 18) have influence on the intercept.

8- References

1- السليمان ،مثنى صبحي(1995) ،"التقدير باستخدام النماذج الخطية واللاخطية"، رسالة ماجستير ، كلية الإدارة والاقتصاد ، جامعة الموصل ، غير منشورة .

2- Agresti, A., (2006) "An Introduction to Categorical Data Analysis" 2nd ed., John Wiley & Sons, Inc., New Jersey.
3- Cameron, A., C., and Trivedi, P., K., (1998) " Regression Analysis of Count Data", Cambridge University Press, New York.
4- De Jong, P. & Heller, G. Z., 2008, "Generalized Linear Models for Insurance Data", Cambridge University Press.
5- Demaris, A., (2004) " Regression with Social Data", John Wiley & Sons, Inc., New Jersey.
6- Efron, B. and Tibshirani, R., (1993), "An introduction to the bootstrap", Chapman and Hall, New York.
7- Efron, B.,(1992) " Jackknife-after-Bootstrap Standard Errors and Influence Functions" Journal of Royal Statistics, Soc. B 54, pp.83-127.
8- Faraway, J., J., (2006),"Extending the Linear Model with R, Generalized Linear, Mixed Effects and Nonparametric Regression Models", Chapman & Hall/CRC, Florida.
9- Greene, W.,(2008) " Functional Forms for the Negative Binomial Model for Count Data", Economic Letters, Vol. 99, pp.585-590.
10-Long, J., S., and Freese, J.,(2001) " Regression Models fpr Categorical Dependent Variables Using STATA" , Stata Press Publication, Texas.
11-Orme, J., G. and Combs-Orme, T.,(2009) " Multiple Regression with Discrete Dependent Variables", Oxford University Press,Inc., New York.
12-Winkelmann, R., (2008) "Econometric Analysis of Count Data", 5th ed., Springer-Verlag Berlin Heidelberg, Heidelberg.