Diagnostic in Poisson Regression Models

Zakariya Y. Algamal
Lecturer / Statistics and Informatics Dept. / College of Computer Science and Mathematics / University of Mosul / Mosul-IRAQ

Abstract

Poisson regression model is one of the most frequently used statistical methods as a standard method of data analysis in many fields. Our focus in this paper is on the identification of outliers, we mainly discuss the deviance and Pearson χ^2 as diagnostic statistics in identification. Simulation and real data are presented to assess the performance of the diagnostic statistics.

Keywords: Poisson regression, Deviance, Pearson χ^2, outliers.

1- Introduction

Poisson regression models have received much attention in econometrics and medicine literature as model for describing count data that assume integer values corresponding to the number of events occurring in a given interval. The Poisson regression model is the most basic model, where the mean of the distribution is a function of the explanatory variables. This model has the defining characteristic that the conditional mean of the outcome is equal to the conditional variance (Long, 1997).

Outliers are observations that do not follow the statistical distribution of the majority of the data. Outlier detection is a primary step in regression analysis and has attracted enormous attention in the literature over many years including Cook and Weisberg (1982), and Rousseeuw and Leroy (1987). There are a number of different statistics used by statistician to ordinary least squares regression. Leverage in Poisson regression is assessed by the hat values h_i. DFBETA and DFFIT are helpful for detecting influence in Poisson regression. DFBETA is calculated by finding the difference in an estimate before and after a particular observation is removed. The same in DFFIT except the calculating

difference will be in predicted values.

The deviance plays an important role in assessing the fit of the model and statistical tests for parameters in the model, and also provides one method for calculating residuals that can be used for detecting outliers (Jong and Heller, 2009). Guria and Roy (2008) used the deviance for detecting outliers in logistic regression. Our focus in this paper is on the identification of outliers, we mainly discuss the deviance and Pearson χ^2 as diagnostic statistics in identification. The structure of the paper is the following. We briefly present in section 2 the estimation of the Poisson regression parameters for both deleted and undeleted observation. In section 3, we introduced the deviance and Pearson χ^2 criteria to detect the outliers. Simulation and real data examples are covered in section 4 and 5 respectively. Section 6 shows the conclusion.

2-Background and Notation for the Poisson Regression Models

In Poisson regression model, hereafter PR, the number of events **y** has a Poisson distribution with a conditional mean that depends on individual characteristics according to the structural model

$$\theta_i = E(y_i|x_i) = Exp(x_i\beta) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Taking the exponential of $x\beta$ forces the expected count μ to be positive, which is required for the Poisson distribution (Long, 1997). If a discrete random variable y follows the Poisson distribution, then

$$p(Y = y) = \frac{e^{-\theta}\ \theta^{y}}{y!} \quad , y = 0,1,2,....(2)$$

In order to estimate the PR estimator, we use the maximum likelihood estimation. By taking the log-likelihood with $\theta_i = Exp(x_i\beta)$, we get

$$\log L(y_1, y_2, ..., y_n \mid \beta, x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \log p(Y_i = y_i \mid \beta, x_i) \quad \ldots\ldots(3)$$

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ - Exp(x_i'\beta) + y_i (x_i'\beta) - \log(y_i !) \right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots...(4)$$

where (β) is a $(k*1)$ vector of parameters, and (x) is a (n*(k+1)) matrix of explanatory variables The maximum likelihood estimator is then defined as:

$$\hat{\beta}_{ML} = \arg\max_{\beta} \log L(\beta)$$

……..(5)

So, the maximizing value for β is found by computing the first derivative of the (4) and setting it equal to zero

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} [y_i - Exp(x_i'\beta)]x_i = 0$$

…….(6)

The second derivatives, Hessian matrix, is given below

$$\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} Exp(x_i'\beta)\, x_i x_i'$$

……(7)

Since the equation (6) is nonlinear in β, one must use an iterative algorithm. A common choice that work well is the Newton-Raphson method as

$$\beta_{ML}^{(m+1)} = \beta_{ML}^{(m)} - (H^{(m)})^{-1} S^{(m)}$$

…….(8)

Where S is the first derivative of the log-likelihood, H is the second derivatives, Hessian matrix, and m is the number of iterations (Winkelmann, 2008), (Yan and Su, 2009). To see the influence of the deletion of the k^{th} observation on the PR, we consider the log-likelihood function as following:

$$\log L(\beta)_k = \sum_{i=1,\neq k}^{n} \{ - Exp(x_i'\beta) + y_i (x_i'\beta) - \log(y_i !)\}$$

……..(9)

Then

$$S^{(k)} = \sum_{i=1,\neq k}^{n} [y_i - Exp(x_i'\beta)]x_i$$

……...(10)

$$H^{(k)} = -\sum_{i=1,\neq k}^{n} Exp(x_i'\beta)\, x_i x'$$

……(11)

Starting with an initial solution then the Newton-Raphson become

$$\beta_{ML}^{(m+1)(k)} = \beta_{ML}^{(m)(k)} - (H^{(k)})^{-1} S^{(k)}$$

…….(12)

3- Single Case Deletion Diagnostic

To show the amount of change in PR estimates that would occurred if the k^{th} observation is deleted. Two diagnostic statistics are proposed, change in deviance and change in Pearson χ^2 to detect the outliers. Such diagnostic statistics are one that examine the effected of deleting single case on the overall summary measures of fit. Let χ_p^2 denotes the Pearson χ^2 statistics and $\chi_p^{2(k)}$ denotes the statistic after the case k is deleted. Using one-step linear approximations given by Pregibon (1981), it can be shown that the decrease in the value of the χ_p^2 statistic due to deletion of the k^{th} case is:

$$\Delta \chi_p^{2(k)} = \chi_p^2 - \chi_p^{2(-k)} \quad , k = 1,2,3,.....,n$$

……(13)

The χ_p^2 is defined as (McCullagh and Nelder, 1989)

$$\chi_p^2 = \sum_{i=1}^{n} (y_i - Exp(x_i'\beta))^2 / \text{var}(\hat{\mu})$$

……(14)

And the χ_p^2 for the k^{th} deleted case is

$$\chi_p^{2(-k)} = \sum_{i=1,\neq k}^{n} (y_i - Exp(x_i'\beta))^2 / \text{var}(\hat{\mu})$$

……(15)

where $\text{var}(\hat{\mu})$ is the estimated variance function. The one-step linear approximation for change in deviance when the k^{th} case is deleted is

$$\Delta D^{(k)} = D - D^{(-k)}$$

……(16)

Because the deviance is used to measure the goodness of fit of a model, a substantial decrease in the deviance after the deletion of the k^{th} observation is indicate that is observation is a misfit.

The deviance of the PR model with and without the k^{th} observation is respectively (Jong and Heller, 2009)

$$D = 2 \sum_{i=1}^{n} \{ y_i \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i) \}$$

……..(17)

where $\hat{\mu}_i = Exp(x_i'\hat{\beta})$

$$D^{(-k)} = 2 \sum_{i=1,\neq k}^{n} \{ y_i \log(\frac{y_i}{\hat{\mu}_i^{(k)}}) - (y_i - \hat{\mu}_i^{(k)}) \}$$

……(18)

A large value of $\Delta D^{(k)}$ indicates that the k^{th} observation is an outlier.

4- Simulation Study Results

A simulation study was conducted to investigate the behavior of the deviance and chi-square Pearson diagnostic statistics under various modeling scenarios. We considered data simulated from a PR with sample size n and p explanatory variables for the cases $(n,p) = (10,1),(25,2),\text{and}(50,2)$. The first case represents a simple PR with x following uniform $[0,1]$ distribution and $\beta = (0,1)$. The second and third case represent a multiple PR with x_1 and x_2 have uniform $[0,1]$ distribution and $\beta = (0,1,1)$. The percentage of contamination was set to be $10\%, 4\%, \text{and } 4\%$ respectively in order to make one or two observations from the response variable sever from shift-mean outlier (the 10^{th}

observation from the first case, the 25^{th} observation from the second case, and observations 29 and 48 from the third case). For brevity, the $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ are presented only in summary in table 1 for the first case, while the rest case results are shown in figure 2 and figure 3. All computations are done by using R program for windows.

Table (1): The $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ statistics for case one

Observation	$\Delta D^{(k)}$	$\Delta \chi_p^{2(k)}$
1	0.1407	0.1323
2	0.687	0.5736
3	0.8742	0.7124
4	0.0044	0.0044
5	4.032	2.212
6	1.8051	1.466
7	3.912	2.2017
8	0.1007	0.0953
9	0.38001	0.4014
10	18.642	31.386

The deviance for the full model was (25.37). The 10^{th} observation will be outlier since it has $\Delta D^{(k)} = 18.642$ and $\Delta \chi_p^{2(k)} = 31.386$. Figure 1 shows the $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ for this case.

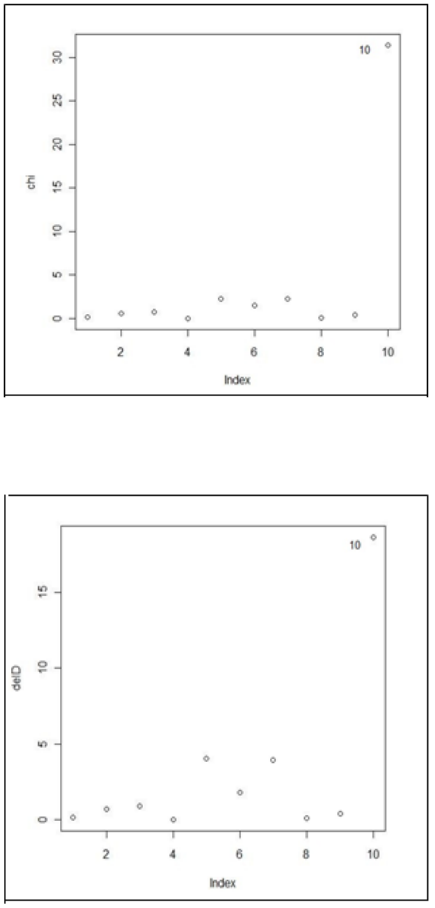
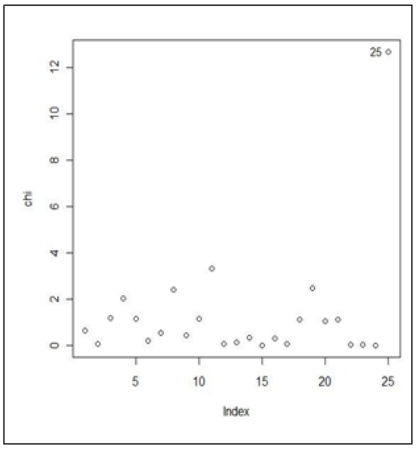


Figure (1): $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ for the first case.



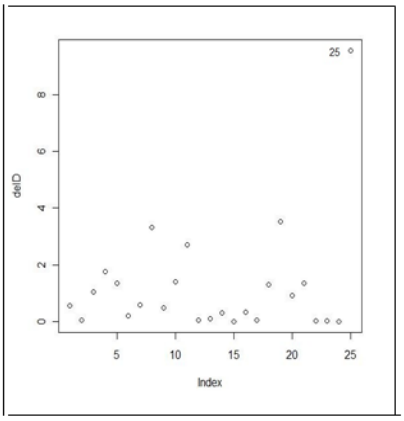
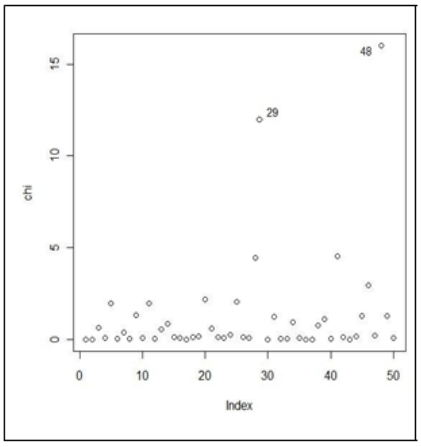


Figure (2): $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ for the second case.

From figure (2) we can conclude that the observation 25 will be outlier since its deletion will decrease the deviance and Pearson χ^2 by (10.3) and (13.4). Again from figure (3), we can considered observations 29 and 48 are outliers since they have large ΔD and $\Delta \chi_p^2$ among the rest of the observations.



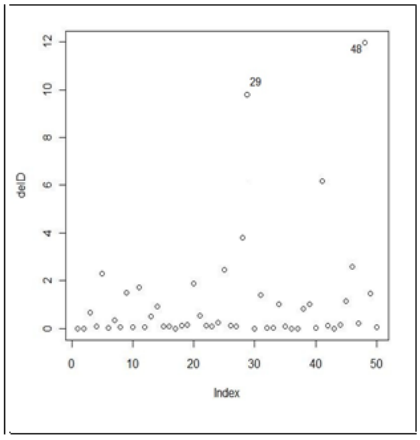


Figure (3): $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ for the third case.

5-Numeriacal Results

The performance of the delta deviance and delta Pearson χ^2 diagnostic statistics was studied in a real data example. Andersen (2008) described data

for Canadian Equality, Security, and Community Survey of 2000. He used only Quebec respondents in the analysis where $(n = 949)$. The response variable is the number of voluntary associations to which respondents belonged. The explanatory variables are gender (with women as the reference category), Canadian born (the reference category is "not born in Canada"), and language spoken in the home (divided into English, French, and other, with French coded as the reference category). He used Cook's distance which indicates that there are two observations (31 and 786) may be particularly problematic. Also, he pointed that the analysis of the DFBETA indicates that the influence of these two observations is largely with respect to the effect of Canadian born variable. To assess our diagnostic statistics performance, we used this example. Table (2) shows the $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ only for the two observations (31 and 786), where the full model deviance is 2427.632. Figure 4 shows the overall look about our analysis. All computations are done by using R program for windows.

Table (2): The $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ for the real data

Obser vation	$\Delta D^{(k}$	$\Delta \chi_p^{2($
31	28.0 0042	63.5 904
786	20.8 13	43.5 95

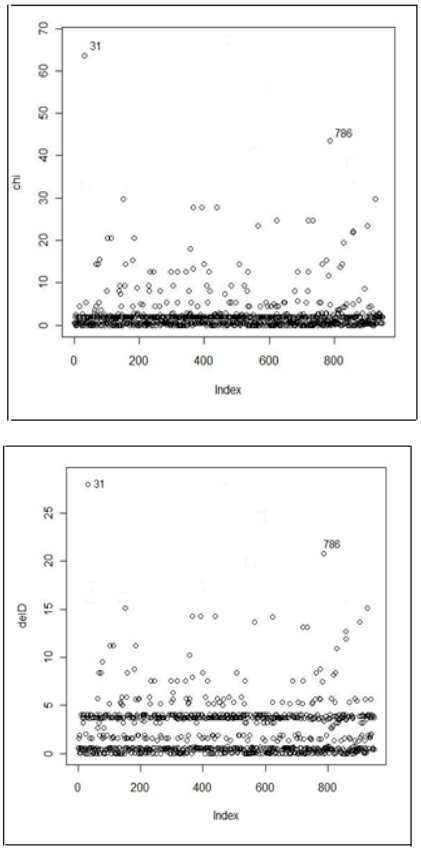


Figure (4): $\Delta D^{(k)}$ and $\Delta \chi_p^{2(k)}$ for the real data.

6- Conclusion

All of the diagnostic statistics described in this paper use one-step approximations to measure the effect of single case deletion on the Poisson regression model parameters. As we can see from figure 1 that the observation 10 considered outlier by our diagnostics statistics and this

corresponds the fact that case 10 has already sever from shifting in mean comparing with the rest observations. The same with observation 25 from figure 2 and observations 29 and 48 from figure 3.

As mentioned in section 5 and from figure 4, the observations 31 and 786 considered outliers using delta deviance and delta Pearson χ^2 diagnostic statistics, this is the same decision that made by Andersen (2008). Here we could conclude that our diagnostic statistics well done in identifying outliers.

7- References

1- Andersen, R., (2008), "Modern Methods for Robust Regression", SAGE Publication, Inc., USA.

2- Cook, R.,D., and Weisberg, S., (1982), "Residuals and Influence in Regression", Chapman & Hall, NY.

3- Guria, S. and Roy, S., S., (2008), "Diagnostics in Logistic Regression Models", Journal of the Korean Statistical Society, 37, pp.89-94.

4- Jong, P. and Heller, G., Z., (2009), "Generalized Linear Models for Insurance Data", Cambridge University Press, UK.

5- Long, J., S., (1997), "Regression Models for Categorical and Limited Dependent Variables", SAGE Publication, Inc., USA.

6- McCullagh, P. & Nelder, J.A., 1989,"Generalized Linear Models", 2nd ed., Chapman and Hall Inc., London.

7- Pregibon, D., (1981), "Logistic Regression Diagnostics", The Annals of Statistics, Vol.9, No.4, pp.705-724.

8- Rousseeuw, P., J., and Leroy, A., M., (1987), "Robust regression and Outliers Detection", John Wiley & Sons, Inc., NY.

9- Seber, G., A., F., and Lee, A., J., (2003),"Linear Regression Analysis", 2nd ed., John Wiley & Sons, Inc., New Jersey.

10- Winkelmann, R., (2008), "Econometric Analysis of Count Data", 5th ed., Springer-Verlag Berlin Heidelberg.
11-Yan, X. and Su, X., G., (2009), "Linear Regression Analysis, theory and computing", World Scientific Publishing Co. Pte. Ltd.