PHYSICAL SCIENCES

PROPORTIONAL EFFECT OF OUTLIERS ON OVER-DISPERSION

[1]Akeyede, I[*]., [2]Saleh, I. M. and [3]Babalola, O. A.

[1]Department of Mathematics, Federal University Lafia, P.M. B 146, Lafia, Nigeria.
[2]Department of Mathematics, Federal University Lafia, P.M. B 146, Lafia, Nigeria.
[3]National Population Commission, Osogbo, Osun State
[*]Corresponding Author: E-mail: imamakeyede@gmail.com

Date Manuscript Received:10/06/15 Accepted:06/12/15 Published: December 2015

5

ABSTRACT

The impact of outlier on analysis of time series data in causing over-dispersion was examined. The problem of over-dispersion is central to all General Linear Models (GLM's) having discrete responses. If the estimated dispersion after fitting is not near the expected values, then the data may be over dispersed. One of the causes of over-dispersion is outlier. Outlier is a data which is unusual with respect to the group of data in which it is found. In this paper, data were simulated based on poison model using SPSS and first analysed to see whether the estimated parameters is unbiased of the fixed parameters. Thereafter, two different values of outliers, 10's and 20's were introduced to different percentages of the generated data and then analysed using the STATA package to observe the effect of the outliers being introduced on the data for small, moderate and large samples. The data simulated were replicated 300 times for all categories. The averages of the results were computed. The results showed that the higher the percentage of outliers the more the over-dispersion occurs in the models and the larger the sample size the less the over-dispersion.

Keywords: Outliers, Over-Dispersion, Simulation,

INTRODUCTION

An outlier is an observation that lies outside the overall pattern of a distribution (Moore and Mccabe, 1999). A convenient definition of an outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile. It can also occur when comparing the relationship between two set of data. According to Oxford Dictionary of Statistics (2008), outlier is an observation that is very different to other observations in a set of data. It is a data value which is unusual with respect to the group of data in which it is found. It may be a single isolated value far away from all others, or a value which does not follow the general pattern of the rest. Usually the presences of outliers indicate some sort of problem. This can be a case which does not fit the model under study or an error measurement. Outliers are often easy to spot in histograms. Since the most common cause of outlier is recording error, it is sensible to search for outliers by means of summary statistics and plots of the data before conducting any detailed statistical modeling or analysis. If there is only a single outlier present, then an effective test is the Dixon test. For data from a normal distribution, the test statistic of the Grubbs test, suggested by Grubbs (1969), could be used. The Rosner test for multiple outliers relies on ordering the n observation interms of their distance from the overall mean. Certain statistical estimators are able to deal with statistical outliers and are robust while others cannot deal with them. A typical example is the case of median, that can deal with outliers well, since it would not matter whether the extreme point is far away or near the other data points, as long as the central value is unchanged. The mean on the other hand, is affected by outliers as it increases or decreases in value depending on the position of the outlier. According to Hardin and Hilbe (2007), presence of outliers in data set may rise to apparent over-dispersion. Over-dispersion is a phenomenon that occurs with data fitted using the binomial, poison or negative binomial distribution. If the estimated dispersion after fitting is not near the assumed values, then the data may be overdispersed, the value is greater than the expected value. It is underdispersed, if the value is less than expected. It is generally caused by positive correlation between responses or by excess variation between response probabilities or counts. It also arises when there are violations in the distributional assumptions of the data (Breslow, 1990). The problem with over-dispersion is that it may cause underestimation of standard errors of the estimated coefficient vector. A variable may appear to be a significant predictor when in fact it is not. Usman and Oyejola (2013) emphasized that apparent over-dispersion may arise from any of the following:

(i) The model omits important explanatory predictors
(ii) The data contain outliers.
(iii) The data contain excess zero.
(iv) The model fails to include enough interaction terms.
(v) A predictor needs to be transformed (to the log or some other scale).

The assumed linear relationship between the response and the link function and predictor is misspecified. (Hardin &Hilbe 2007) A model may be overdispersed if the value of the Pearson (or $?^2$) statistics divided by the degree of freedom is greater than 1.0. The quotient of either is called the dispersion. Small amounts of over-dispersion are of little concern; however, if the dispersion statistics is greater than 1.25 for moderate size models, then a correction may be warranted. Models with large numbers of observations may be overdispersed with a dispersion statistics of 1.05 (Hilbe 2007). This study therefore examined the effect of proportion of outliers and sample size in causing over-dispersion to set of data

MATERIALS AND METHODS

Proportional impact of outliers were studied by creating simulated data set for small, moderate and large samples which were taken to be 20, 50 and 100 respectively. For each sample size, we introduced 1, 2 and 3 different sets of outliers out of each of the values 20, 50 and 100 of the response y_i following the idea of Usman and Oyejola (2013). These were replicated 300 times. For instance, the numbers of values of outliers introduced in each sample represent 5, 10 and 15 percent of the observations for the sample size of 20. The values of y_i simulated range from 0 to 9.

Two sets of outliers were introduced into generated data. In the first set we added 10 to the first, first and second, and first, second and third respective values of y_i randomly in the different data generated. While in the second set, we added 20 the same way. Each constructed data set entails a specific cause of the over-dispersion observed in the display of the model output stated as follows;

Constant (β_o) = 0.9 and β_1 = 0.2, β_2 = -0.5, β_3 = 0.6 are coefficients of the predictors. $t = 0, 1, \ldots, \infty$, and $i = 1, 2, \ldots, 300$

Results Output of Sample Size of 20 without Outliers using Stata Codes

glmy$_i$ x$_{1i}$ x$_{2i}$ x$_{3i}$, family(Poisson) link(identity) nolognonrtolerance i=1, …, 300

A sample output of the above code is given as linear models. No. of obs = 20

```
Optimization: ML                  Residual df    = 16
                                  Scale parameter = 1
Deviance  =15.967703              (1/df) Deviance =  0.997981
Pearson = 14.876987               (1/df) Pearson  =0.929811
Variance function: V(u) = u       [Poisson]
Link function: g(u) = u           [Identity]
                                  AIC      = 2.6256856
Log likelihood   = -30.63208      BIC      = -57.9049
```

y_i |	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
x_{1i} |	0.2053205	0.0283715	1.8968	0.51	0.249713	0.3609276
x_{2i} |	-0.5042415	0.0338103	-1.5563	0.10	-0.670508	0.537975
x_{3i} |	0.6024644	0.0304156	3.3466	0.23	0.342851	0.4620779
Cons |	1.0000265	0.0300299	3.5087	0.03	0.947569	1.06528404

The data simulated for 300 replications were computed and the average of the outputs were computed and presented in the table 1-6. However, we would expect that the parameter estimates would equal the values we assigned them and that the Pearson dispersion statistics, defined as the Pearson statistics divided by the model degree of freedom, would less than 1.0. Note the Pearson dispersion statistics in the above model is 0.90798 with parameter estimates approximating the values we specified. Furthermore, a value of outlier was introduced in the generated data above, in this case we added 10 to the fist data. The codes used in the STATA to generate the responses with an outlier introduced on the same set of predictors yield the output is given below.

Results Output of Sample Size of 20 with an Outlier '10' using Stata Codes

geny$_i$ = *y*

replacey$_i$ = *y*$_i$+ *10 in 1/1*

glmy$_i$ *x*$_{1i}$ *x*$_{2i}$ *x*$_{3i}$, *family(Poisson) link(identity) nolognonrtolerancei=1, …, 300*

A sample output of the above code is given as linear models. No. of obs = 20

```
Optimization    : ML              Residual df    = 16
                                  Scale parameter = 1
Deviance =  21.936224             (1/df) Deviance = 1.371014
Pearson  = 32.374656              (1/df) Pearson = 2.023416
                                  AIC      = 2.805261
Log likelihood  = -38.630618      BIC      = -5514.594
```

yiCoef.	Std. Err.	z	P>|z|	[95% Conf. Interval]		
x1 |	.3427364	.0334277	10.25	0.001	.2772193	.4082534
x2 |	-.5451369	.0453168	-12.03	0.003	-.6339563	-.4563175
x3 |	.3167947	.030732	10.31	0.000	.2565612	.3770283
cons |	1.124692	.0344106	32.68	0.001	1.057248	1.192135

From the above result, the parameter estimates are significantly different from the parameters fixed for the model having a responsey$_i$, i.e. y with the first responses having 10 added to the value y. the Pearson dispersion statistics, however, has doubled to a value of 2.0234. The AIC and BIC Statistics are also inflated. Given a small number of observations, a value of 2.0234 indicates a serious over-dispersion, of course, we understand that the source of the over-dispersion result from the 10-outlier. Adding another 10's counts to the observations we already made to the first observations produce multiple over-dispersion (see table 1-6 for the results). More so another value of outlier was introduced in the generated data.In this case, we added 20 to the first data. The codes used in the STATA to generate the responses with an outlier introduced on the same set of predictors yield the output given below.

Results Output of Sample Size of 20 with an Outlier '20' using Stata Codes

genyi = y

replaceyi = yi+20 in 1/1

glmyi x1i x2i x3i, family(Poisson) link(identity) nolognonrtolerancei=1, …, 300

A sample output of the above code is given as linear models. No. of obs = 20

```
Optimization    : ML              Residual df     =    16
                                  Scale parameter =     1
Deviance =  37.79224              (1/df) Deviance = 2.362015
Pearson  =  58.03245              (1/df) Pearson  = 3.627025
                                  AIC      = 5.002345
Log likelihood  = -38.4567137     BIC      = -11.0712
```

yiCoef	Std. Err.	z	P>|z|	[95% Conf. Interval]		
x1 |	.3754891	.0245773	11.21	0.007	.3.012568	.78956321
x2 |	-.7896587	.0675321	-10.53	0.010	-.7.231562	-.6853217
x3 |	.4517123	.0564325	12.45	0.000	.0457852	.5876543
cons |	2.154321	.0312543	13.69	0.023	1.453278	1.786549

When another value of outlier was introduced in the generated data, i.e 10 added to the first data. The Pearson index increased to 3.627025 which is seriously overdispersed. Also there is a change in the parameter estimates and the AIC and BIC criteria increase to 5.002345 and -11.0712 respectively. The effect of 10% of the observation constituted outlier is remarkable.

RESULTS AND DISCUSSION

ANALYSIS WITH VALUE OF 10'S ADDED TO SOME PERCENTAGE OF RESPONSE

The analysis of data when 10 was added to the first, first and second, and first, second and third data could be seen clearly in the table 1-6. The data were simulated for each three sample size under consideration and analysed. Then a value of 10 was introduced to 5%, 10% and 15% of 20, 50 and 100 observations and they were analysed using the Stata code. Each set of simulations were replicated 300 times, the average of the results were taken and displayed in table 1-3 as follows.

Table 1: Proportional Effect of Outliers in Analysis for Data Set of Sample Size 20

Percentage Index	Pearson Likelihood	Log-	AIC	BIC	B_1	B_2	B_3	Constant
0%	0.9298	-30.6322.8052-57.9050.2053-0.5042			0.6025		1	
5%	1.5673	-36.9933.2997-23.6720.3126-0.76540.41361.3218						
10%	2.3426	-37.8923.6712-21.7650.4312-0.71120.46781.5432						
15%	2.5630	-39.8974.2314-19.8650.45210.78560.56321.6754						

Table 2: Proportional Effect of Outliers in Analysis for Data Set of Sample Size 50

Percentage Index	Pearson Likelihood	Log-	AIC	BIC	B_1	B_2	B_3	Constant
0%	0.9006	-21.7662.3312-77.125		0.2103-0.50020.61251.0012				
5%	1.5632	-36.8763.1007-53.6740.3451-0.67340.42371.1432						
10%	2.4321	-42.6743.6588-34.7780.4654-0.69870.48911.5651						
15%	2.6126	-49.9884.0654-23.7650.4897			0.867540.54671.7733			

Table 3: Proportional Effect of Outliers in Analysis for Data Set of Sample Size 100

Percentage Index	Pearson Likelihood	Log-	AIC	BIC	B_1	B_2	B_3	Constant
0%	0.8731	-20.98762.3001-82.1350.2189 -0.49820.6521 1.0066						
5%	1.5632	-34.90253.0992-54.0020.3500 -0.6434 0.4743 1.2318						
10%	2.6003	-41.77783.6172-35.7980.4672-0.6532 0.49981.5832						
15%	2.6715	-42.1723.7145 -25.087 0.49510.7864 0.56321.9865						

Tables 1-3 show the effect of percentage of outliers introduced when compared with those with 0% outlier. It was observed that the Pearson's index increases with percentage increase in the number of outliers, hence, over-dispersion occurred as all values of the index are greater than 1.0 for all categories of percentage of outlier introduced and sample sizes. Also from the results all the parameters estimates are significantly different from the fixed parameters. The increase in AIC and BIC information criteria when percentage of outliers increase indicate worse model from smaller to higher number of outliers. It was also observed that the parameters' estimates, Pearson Index and Log-Likelihood increased while AIC and BIC decreased when sample size was increased for different percentage of outliers introduced.

ANALYSIS WITH VALUE OF 20'S ADDED TO SOME PERCENTAGE OF RESPONSE

The data were simulated for each three sample sizes under consideration and analysed. Then a value of 20 was introduced to 5%, 10% and 15% of 20, 50 and 100 observations and they wereanalysed using the Stata code. Eachset of data was replicated 300 times for different percentage of outliers. The average values of the estimated parameters Pearson index, Log-Likelihood and the Information criteria were taken and presented in table 3-6 as follows.

Table 4: Proportional Effect of Outliers in Analysis of Data Set with Sample Size 20

Percentage Index	Pearson Likelihood	Log-	AIC	BIC	B_1	B_2	B_3	Constant
0%	0.92981	-30.6322.8052-57.9050.2053-0.50420.60251.0000						
5%	3.0239	-37.9814.0078-12.8970.2328-0.96670.43211.3254						
10%	3.9876	-50.8985.2245-7.87690.3897-0.90070.43691.3367						
15%	5.8976	-51.7896.8965-2.89650.1567-0.89760.51562.8976						

Table 5: Proportional Effect of Outliers in Analysis for Data Set of Sample Size 50

Percentage Index	Pearson Likelihood	Log-	AIC	BIC	B_1	B_2	B_3	Constant
0%	0.9006	-21.766	2.3312	-77.1250.2103-0.50020.61251.0012				
5%	3.4321	-91.0454.0156-79.9880.4116-0.7654-0.56781.339						
10%	3.9876	-123.685.2245-20.8980.3814-0.78960.59871.9876						
15%	5.1065	-151.676.1897-21.7770.4292-0.79980.60022.8764						

Table 6: Proportional Effect of Outliers in Analysis for Data Set of Sample Size 100

Percentage Index	Pearson Likelihood	Log-	AIC	BIC	B_1	B_2	B_3	Constant
0%	0.9236	-121.76	2.1267	-89.7680.2007-0.51120.61151.0102				
5%	0.9256	-255.923.0651-67.9870.2145-0.54320.67521.0795						
10%	0.9285	-187.903.1236-55.7690.2276-0.54780.67861.2435						
15%	1.8931	-176.443.6751-43.7860.22580.22580.68952.2367						

Tables 3-6 show the effect of percentage of outliers introduced when compared with those with 0% outlier. In this case, outlier 20 was added to 5%. 10 and 15% of data simulated. It was observed that, the Pearson's index increases with percentage increase in the number of outliers, hence, over-dispersion occurred as all values of the index are greater than 1.0 for all categories of percentage of outlier introduced and sample sizes. Also from the results all the parameters estimates are significantly different from the fixed parameters. The increase in AIC and BIC information criteria when percentage of outliers increase indicate worse model from smaller to higher number of outliers.

CONCLUSION

Tables 1-6 show the effect of percentage of outliers introduced when compare with those with 0% outlier. It was observed that the Pearson's index increases with increase in the number of outliers, hence, over-dispersion occurred as all values of the index are greater than 1.0 for all categories of percentage of outliers introduced. Also from the results all the parameters' estimates are significantly different from the model. The increase in AIC and BIC information criteria respectively, when percentage of outliers increase, indicate worse model from smaller to higher number of outliers. It was also observed that, the parameters' estimates, Pearson Index and Log-Likelihood increased while AIC and BIC decreased when sample size was increased for different percentage of outliers introduced. Therefore, the outlier has little effect on the model with increase in the sample size and indeed there is little over-dispersion.

The study concluded that the higher the number of outliers in a set of data the higher the over-dispersion especially at smaller sample sizes. However if the sample size increases with the same number of outliers there will be little over-dispersion. It is therefore recommended that if outliers are present in a data set the sample size should be increased.

REFERENCES

Breslow, N. E. (1990). Tests of hypotheses in Overdispersed Poisson regression and other quasilikelihood models. *Journal of American Statistical Association, 85*, 565-571.

Cox, D. R. (1983).Some Remarks on Over-dispersion. *Biometrica*, 70, 269-274

Grubbs, F. E. (1969).Procedure for detecting outlying observations in samples. *Technometrics*, 11, 1-21.

Hilbe, J. M. (2008).Negative Binomial Regression. Cambridge University Press, United Kingdom. Pp. 51-61

Hardin, J. W. & Hilbe, J. M. (2007).Generalized Linear Models and Extensions. Seco Edition, AStata Press Publication StataCorp LP, College Station, Texas. Pp. 49-50, 165-166, 221-222.

Lambert, D. &Roader, K. (1995).Over-dispersion diagnostics for generalized linear models. *Journal of American Statistical Association*, 90, 1225-1236.

Moore, D.S. & McCabe, G. P. (1999). Introduction to the practice of Statistics, 3rd edition, W. H. Freeman, New York: Pp. 571-578.

Oxford Dictionary of Statistics (2008).*A* Dictionary of Statistics, Second edition revised. Oxford University Press, USA.

Usman, M. & Oyejola, B. A. (2013).Models for Count Data in the Presence of Outliers and / or Excess Zero. *Mathematical Theory and Modeling*, 3(7), 94-103