Research Article

Test Statistics for Dispersion Parameter in Poisson Regression and Generalized Poisson Regression Models

Veeranun Pongsapukdee[1*], Pairoj Khawsittiwong[1] and Maysiya Yamjaroenkit[2]

[1]*Department of Statistics, Faculty of Science, Silpakorn University, Nakhon Pathom 73000, Thailand*
[2]*Department of Mathematics and Computing, Faculty of Science and Technology, Phetchaburi Rajabhat University, Phetchaburi, Thailand*
[*]*Correspondence author. Email address: veeranun@silpakorn.edu, veeranun@hotmail.com*

Received March 15, 2015; Accepted September 3, 2015

Abstract

Two symmetrical distributed test statistics, called Z_m and Z_{0_New} are proposed and their goodness-of-fit tests are compared with other available five test statistics: Wald-t, Score test, $Z_{\hat{\bar{\mu}}}$, $Z_{\bar{Y}}$, and Z_0, for overdispersion in Poisson regression model versus generalized Poisson model. Five thousand data sets in each condition of overdispersion parameters and sample sizes are simulated to perform the assessment of the models' fits using those statistics, concerning the coverage probability and power of tests. Results show that the Z_m test performs closely as good a $Z_{\hat{\bar{\mu}}}$ and $Z_{\bar{Y}}$tests but it tend to be better than the others when the sample size is large. Even if the Z_{0_New} test has the largest power; however, in consideration for coverage probability and power of tests, the Z_m test probably be more reliable. The Z_m test statistic is interesting not only in its simplest form, with the reasonable coverage probability and power but also in its robust property of using median that needs fewer assumptions for its parent distribution.

***Key Words*:** Goodness-of-fit statistics; Coverage probability; Power of tests; Generalized linear models

Introduction

Statistical modeling is a well known process for analyzing the count data. In the case of Poisson response with at least one continuous explanatory variable, Poisson regression modeling is often fitted in the analysis as a basis for categorical data analysis (Frome et al., 1973; Frome, 1983; Yamjaroenkit and Pongsapukdee, 2012; Pongsapukdee, 2012; Agresti, 2013). In theory, data of the Poisson distribution should have its mean equal to its variance. However, in practice, the corresponding mean or observed variance of the data seldom meets this assumption and tends to be larger than the corresponding mean. Thus, count data might display substantial extra-Poisson variation (or overdispersion) and that the appropriate statistical modeling, tackling of this problem must be concerned in advance.

In the case of Poisson response with at least one continuous explanatory variable, Poisson regression modeling procedure is commonly used in the analysis of data. For Poisson model, data arising from groups or individuals are often statistically dependent within the group, therefore, the observed variance of data may not equal to the corresponding mean. The excess variability is called overdispersion. In this problem, the ordinary Poisson regression modeling is not adequate in analysis of data; however, generalized Poisson models (Consul and Jain, 1973), which is using the negative binomial distribution instead of the Poisson distribution to account for the overdispersion, will be more appropriated. Several test statistics and methods have been suggested for dealing with such problem in order to determine the overdispersion of data; for example, Wald-t test (Wang and Famoye, 1997), Score test, $S(\hat{B})$(Yang, et al., 2009), Z_0 test (Böhning, 1994), and $Z_{\hat{\bar{\mu}}}$ and $Z_{\bar{Y}}$ tests (Yamjaroenkit and Pongsapukdee, 2012).

Moreover, Lawless, 1978; Dean and Lawless, 1989; and Dean, 1992 have noted that in certain circumstances and problems, the asymptotic distributions used with the

ISSN 1905-9159

usual tests such as a likelihood ratio test may provide misleading results, as they tend to underestimate the evidence against the base model. Statistical models for describing and analyzing the over dispersed Poisson data and Poisson rates usually are generalized Poisson regression models (Consul and Jain, 1973; Consul, 1989; Famoye, 1997) which includes the Poisson regression model (Cameron and Trivedi, 1998) as nested or a special case model. Therefore, testing for extra-Poisson variation can be applied and performed by fitting a longer model or more comprehensive model such as a generalized Poisson regression model (for overdispersion case) versus fitting a shorter model such that a Poisson regression model. This is corresponding to the test for a goodness-of-fit for a reduction to the simple or shorter model, using a test for overdispersion. In other words, if the null model is not rejected or no overdispersion is significance, a Poisson regression model is adequate. However, if the null model is rejected, the overdispersion problem is evidence, thus a generalized Poisson regression model is a better fit model for analyzing the data and handle the problem in this case.

In this article we propose two newly developed test statistics, Z_m and Z_{0_New} of testing for overdispersion of count data and empirically investigate the results under the null Poisson regression model versus the alternative generalized Poisson regression model. If no evidence of overdispersion occurs, Poisson regression model is good fitted. But if overdispersion occurs, generalized Poisson regression model is more appropriated. For considering the efficiency of the test statistics, the study based on the power of the tests and the coverage probabilities are compared with those from other tests: Wald-t test, Score test, Z_0 test, $Z_{\bar{\mu}}$ and $Z_{\bar{Y}}$ tests. The response data are simulated under the generalized Poisson regression model with the single explanatory variable generated from continuous Uniform (0, 1). Data analyses using the rewritten macro run with SAS® enterprise version 5.1 by performing 5,000 data sets in each condition of the dispersion parameters and the sample sizes of 30, 50, 100, 200, and 500, respectively.

Models and test statistics

Generalized Linear Models and Generalized Poisson regression models

Poisson regression models and Generalized Poisson regression models are widely used in many areas of scientific applications and social science researches. These models are a branch of the Generalized Linear Models or GLMs (Nelder and Wedderburn, 1972) which concerns an exponential family distributed response, systematic linear combination, and link functions. The log link function and the categorical response variable of interest make the Poisson regression and the generalized Poisson regression models distinct from a linear regression model which depends only on a normal distributed response. These two GLMs models can also be applied by extending to correlated data sets using Generalized Linear Mixed Models or GLMMs including some nonlinear modeling and random effects modeling (McCulloch and Searle, 2001).

The usual Poisson loglinear model (Agresti, 2013), with the positive mean of Poisson distribution and an explanatory variable X, has the form: $\log\mu(x) = \alpha + \beta x$, where α, β are model parameters. This model satisfies the exponential relationship $\mu(x) = \exp(\alpha + \beta x)$ and a 1-unit increase in X_j has a multiplicative impact of $e^{\beta j}$. The mean at $X_j + 1$ equals the mean at X_j multiplied by $e^{\beta j}$. The log mean is the natural parameter for the Poisson distribution. When at least one of the explanatory variable, X's, is continuous variable, the above model is often called "a Poisson regression model". In addition, when events of a certain type occur over time, space, or some other index of size, it is usually more relevant to model the rate at which they occur than modeling thecount data and the above model is often modified to Poisson regression for rates (Agresti, 2013). For more complicated Poisson loglinear modeling such that generalized loglinear models (Salee and Pongsapukdee, 2013) and the generalized Poisson regression models, often we need to check for the overdispersion problems (Agresti, 2013).

In contrast, when the overdispersion situation is common in the modeling of counts, and the variance is larger than its mean. The generalized Poisson regression model (Consul and Jain, 1973), with taking the negative binomial distribution instead of the Poisson distribution to account for the overdispersion, is the better fit model. Therefore the generalized Poisson model is often considered to use in most modeling of overdispersion situation compared with the Poisson regression models after testing that it is a better alternative by using several test statistics (Yang, et al., 2009; Pongsapukdee, 2012; Agresti, 2013).

Test statistics

Seven test statistics consist of Wald-t test (Wang and Famoye, 1997), Score test, $S(\hat{\beta})$ (Yang,et al., 2009), Z_0 test (Böhning,1994), $Z_{\bar{\mu}}$ and $Z_{\bar{Y}}$ tests (Yamjaroenkit & Pongsapukdee, 2012), and the two newly proposed test statistics, Z_m test and the Z_{0_New} test are all considered. All of which the formulae are stated in the following (1) - (7).

Five former statistics

The Wald-t (cited in Yang, et al., 2009; Wang and Famoye, 1997) has the form

$$t = \frac{\hat{\varphi} - \varphi}{SE\left(\hat{\varphi}\right)}. \quad (1)$$

where, $\hat{\varphi}$ is the estimated value of the dispersion parameter, φ , $SE(\hat{\varphi})$ is the standard error of $\hat{\varphi}$.

The Score test or $S(\hat{\beta})$ (Yang, et al., 2009) has the form in (2).

$$S\left(\hat{\beta}\right) = \left(\sqrt{2\sum_{i=1}^{n}\hat{\mu}_i^2}\right)^{-1} \sum_{i=1}^{n}\left(\left(y_i - \hat{\mu}_i\right)^2 - y_i\right) \quad (2)$$

where, $\hat{\mu}$ is the estimated value of y, from the corresponding model.

Following the test statistic of Böhning (Böhning, 1994) under the null hypothesis is true: H_0: $\varphi = 0$, Z_0 has the form in (3).

$$Z_o = \frac{S^2 - \overline{Y}}{\sqrt{2\overline{Y}^2/(n-1)}} = \sqrt{\frac{n-1}{2}}\left(\frac{S^2}{\overline{Y}} - 1\right) \quad (3)$$

where, S^2 is the sample variance, and $\overline{Y}$ is from the sample mean of Poisson observed counts.

The test for dispersion parameter $\varphi = 0$ is corresponding to the Poisson model in null hypothesis versus the alternative hypothesis, H_1: $\varphi > 0$ which is in contrast corresponding to the generalized Poisson regression model.

Extending the Böhning's statistic by using the $\overline{Y}$ from its predicted mean of the fitting generalized Poisson regression model under study, that is $\hat{\mu}$, we obtain the estimated variance, S^2, which can be rewritten in the form of $S^2 = \overline{\hat{\mu}}\left(1 + \hat{\varphi}\,\overline{\hat{\mu}}\right)^2$, the estimated variance of the generalized Poisson regression model from sample data. Hence, the test statistic, $Z_{\overline{\hat{\mu}}}$, has the form in (4).

$$Z_{\overline{\hat{\mu}}} = \sqrt{\frac{n-1}{2}}\left(\frac{\overline{\hat{\mu}}\left(1 + \hat{\varphi}\,\overline{\hat{\mu}}\right)^2}{\overline{\hat{\mu}}} - 1\right) = \sqrt{\frac{n-1}{2}}\left(\left(1 + \hat{\varphi}\,\overline{\hat{\mu}}\right)^2 - 1\right) \quad (4)$$

where, φ is the dispersion parameter, $\overline{\hat{\mu}}$ denotes the predicted mean of the generalized Poisson regression model from sample data, and $\hat{\varphi}$ is the estimated value of the dispersion parameter from sample data. The resulting plots of the $Z_{\overline{\hat{\mu}}}$ distribution obtained from simulation studies indicate that the test statistic $Z_{\overline{\hat{\mu}}}$ has approximately asymptotic standard normal (Yamjaroenkit and Pongsapukdee, 2012; Yamjaroenkit, 2012). An alternative to $Z_{\overline{\hat{\mu}}}$ called $Z_{\overline{Y}}$ using $\overline{Y}$ or the sample mean of the response variable instead of the $\overline{\hat{\mu}}$ in the equation (4) we obtain another test statistic in (5).

$$Z_{\overline{Y}} = \sqrt{\frac{n-1}{2}}\left(\left(1 + \hat{\varphi}\,\overline{Y}\right)^2 - 1\right) \quad (5)$$

where, $\overline{Y}$ denotes the sample mean of the response variable Y, and $\hat{\varphi}$ is the estimated value of the dispersion parameter from sample data. The resulting plot of the distribution of $Z_{\overline{Y}}$ also has approximately asymptotic standard normal (Yamjaroenkit and Pongsapukdee, 2012).

Two proposed statistics

In this article we propose two new test statistics, called Z_m test and Z_{0_New} test, of which the forms are similar to (5) but with new corresponding estimators and are stated in (6) – (7), respectively.

$$Z_m = Z_{median} = \sqrt{\frac{n-1}{2}}\left(\left(1 + \hat{\varphi}\,\Delta\right)^2 - 1\right) \quad (6)$$

where, Δ denotes the sample median of the response variable Y,

$\hat{\varphi}$ denotes the estimated value of the dispersion parameter.

$$Z_{0\_New} = \sqrt{\frac{n-1}{2}}\left(\frac{S^2}{\Delta} - 1\right) \quad (7)$$

where, Δ denotes the sample median of the response variable Y, and S^2 denotes the sample variance.

Simulation and statistical analyses

The response Y data are generated from the Poisson model, $\log \mu = 2 + 0.5x$, corresponding with $\mu = \exp(2 + 0.5x)$ and that the generalized Poisson distributed mean is $\frac{\hat{\mu}}{(1+\varphi\mu)^2}$. Whereas, the explanatory variable is generated from continuous U(0, 1) for each condition of sample sizes (n = 30, 50, 100, and 200), and the corresponding overdispersion parameters (φ = 0, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.20). Each data set is simulated and processed for each condition of each Poisson regression probability model ($\varphi = 0$) and each generalized Poisson regression probability model ($\varphi > 0$). The hypotheses model is H_0: $\varphi = 0$, which the case

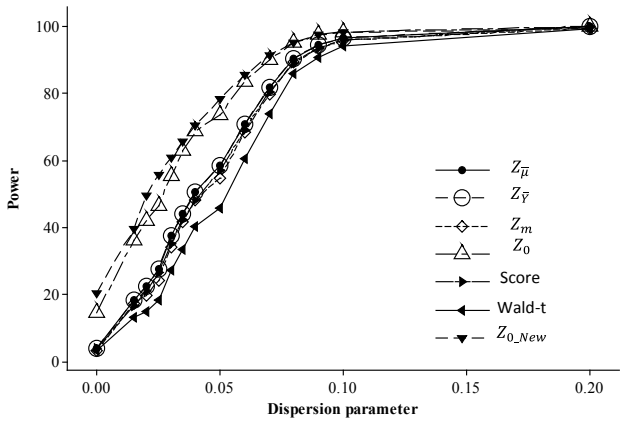

Figure 1 Power plots for overdispersion among seven tests for n=30

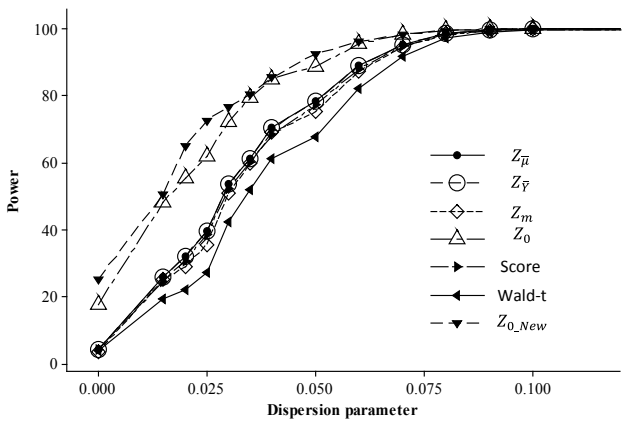

Figure 2 Power plots for overdispersion among seven tests for n=50

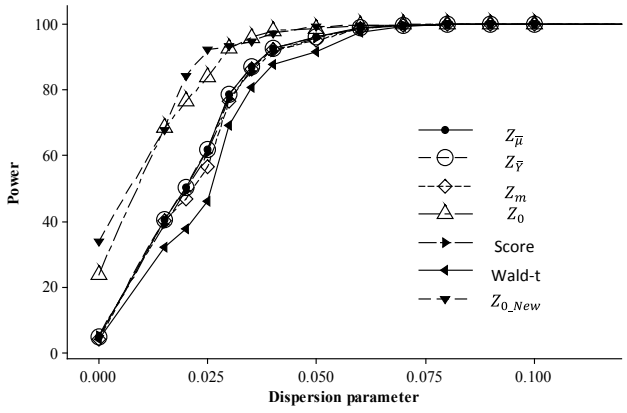

Figure 3 Power plots for overdispersion among seven tests for n=100

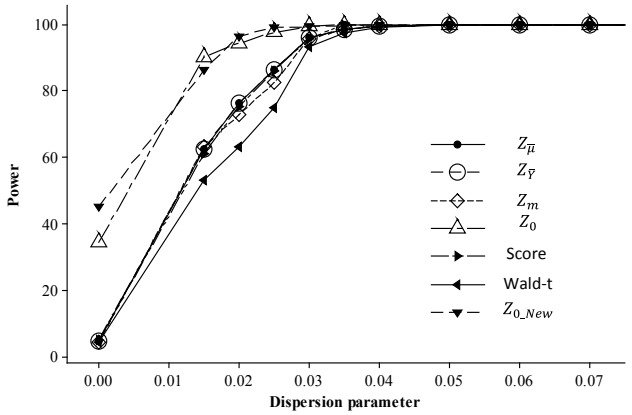

Figure 4 Power plots for overdispersion among seven tests for n=200

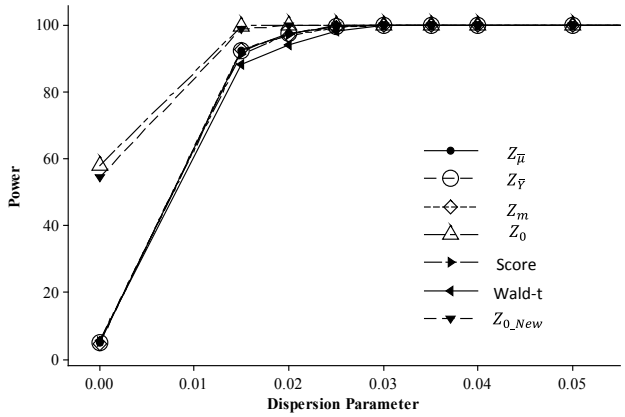

Figure 5 Power plots for overdispersion among seven tests for n=500

of no overdispersion situation is or the Poisson regression model is true, versus $H_1: \varphi > 0$, which is the case of overdispersion situation or the generalized Poisson model is true, and then from testing, H_1 should be accepted. The corresponding critical regions are $Z_\alpha = 1.645$, $Z_{1-\alpha} = -1.645$. In this analysis, the coverage probability and the power of testsof the statistics: the Wald-t, the Score testor $S(\hat{\beta})$, the $Z_{\tilde{\mu}}$ and $Z_{\bar{Y}}$, the Z_0, and two proposed test statistics, Z_m test and Z_{0_New} are all calculated and investigated through the statistical modeling process. Simulations, using the authors' rewritten macro run with SAS® enterprise version 5.1, were processed repeatedly for 5,000 sets in each condition of sample sizes and overdispersion parameters.

Results

The results for the coverage probability ($\varphi = 0$) show that the Score test seems to be appropriate in every sample size. However, this score statistic still involves some outliers and that its power of the tests tends to have less power than that of the Wald-t test; especially, when sample size is large (Yamjaroenkit and Pongsapukdee, 2012). Similarly, the $Z_{\tilde{\mu}}$ and $Z_{\bar{Y}}$ statistics likely dominate all other statistics (except the score test) in terms of the coverage probability ($\varphi = 0$), but for their power of the tests ($\varphi > 0$) these two statistics tend to give smaller power than those from Z_0 and Z_{0_New}. When sample size is large, Z_{0_New} give better coverage probability than the Z_0. When considering only the power it is clear that Z_{0_New} dominate all other tests (Figures 1-5).

Even if, the statistics, Z_0 and Z_{0_New}, perform similarly in possessing higher power of the tests ($\varphi = 0$) than all the rest but they would still appear to produce too big coverage probability values for all sample sizes. Therefore, the Z_m test is probably a reasonable statistic, even if when sample size is small, its coverage probability is still interesting with its converging to 0.05 and its performance is much improved not only in term of the coverage probability but also in term of power of the test, especially when the sample size is large. Thus, this statistic possesses the robust property and that its coverage probability which is approaching the significance level of $\alpha = 0.05$. Meanwhile, its power is also approaching 100%, consistently (Figures 1-5).

Furthermore, the plots of the power of the tests parameters classified by the sample size, the bigger sample size we use the much more power we obtain; especially, from those of Z_m and Z_{0_New} tests. Results also show that when the sample size is increasing the proposed robust tests, Z_m, performs approximately as good as $Z_{\tilde{\mu}}$ and $Z_{\bar{Y}}$ but it tend to be better than others. It can be seen that the coverage probability and the power from the Z_m test are posing most reasonable plots among all tests (Figures 4-5). Thus, in practice the implementation with potential safety test, we recommend to use the Z_m test, particularly for the large sample size. Hence, when concerning both power and the coverage probability, the Z_m test probably be more preferable due to not only in testing goodness-of-fit for overdispersion with its robustness of using median that needs less assumption about any distribution, but also its simplicity form as well as it can be applied for both continuous and discrete explanatory variables as well.

Discussion and conclusion

Two new symmetrical distributed tests called Z_m and Z_{0_New} are proposed and their goodness-of-fit tests for overdispersion are competitively investigated. All results among seven goodness-of-fit test statistics: Wald-t, Score test, $Z_{\tilde{\mu}}$, $Z_{\bar{Y}}$, Z_0, Z_m and Z_{0_New} are compared for overdispersion tests under the Poisson regression model versus the generalized Poisson regression model in terms of the coverage probability ($\varphi = 0$) and the power of tests ($\varphi > 0$). It indicates that the test statistics that can provide both the reasonable coverage probability and the power of tests probably be Z_m test. The Z_m test, even if it cannot dominate other tests with both the coverage probability and power of test; however, its power of the test is approaching to approximately 100 % (or 99.88%) and exactly 100 % when $n = 200$, $\varphi = 0.06$ and 0.07, respectively. Beside this, Z_m test is the simplest statistic and still be appropriate for both cases of either discrete or continuousexplanatory variable. Results also indicate that its power is properly increasing consistently as the sample size is increased. Then, this test statistic would be used more safely in testing goodness-of-fit for overdispersion and should be considered for implementation in practice; especially for large sample. In future research, it may further study among the test statistics' asymptotic distributions and their relative efficiencies, particularly with an extension to models which consist of both discrete and continuous variables that often appear in most practices.

References

Agresti, A. (2013) *Categorical Data Analysis*, 3rd ed., John Wiley & Sons, New York, pp. 115-130.

Böhning, D. A (1994) Note on test for Poisson overdispersion. *Biometrika* 81: 418-419.

Cameron, A. C. and Trivedi, P. K. (1998) *Regression Analysis of Count Data*, Cambridge University Press, Cambridge, pp. 87-146.

Consul, P. C. (1989) *Generalized Poisson Distribution: Properties and Application*, Dekker, Inc.New York, pp. 377-385.

Consul, P. C. and Jain, G. C. (1973) A generalized of the Poisson distribution. *Techonometrics* 15(4): 791-799.

Dean, C. B. (1992) Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*87: 451-457.

Dean, C. and Lawless, J. F. (1989) Test for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* 84: 467-471.

Famoye, F. (1997) Restricted generalized Poisson regression model. *Communication in Statistics – Theory and Method* 22: 1335-1354.

Frome, E. L. (1983) The analysis of rates using Poisson regression models. *Biometrics* 39: 665-674.

Frome, E. L., Kutner, M. H. and Beauchamp, J. J. (1973) Regression analysis of Poisson distributed data. *Journal of the American Statistical Association* 68: 935-940.

Lawless, J. F. (1987a) Negative binomial regression models.*Canadian Journal of Statistics* 15: 209-226.

McCulloch, C. E. and Searle, S. (2001) *Generalized, linear, and mixed models*, John Wiley & Sons, New York, pp. 220-238.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society A* 135(3): 370-384.

Pongsapukdee, V. (2012) *Analysis of categorical data: Theories and applications with GLIM, SPSS, SAS and MTB*, 3rd ed., Silapakorn University Press, Nakhon Pathom, pp. 203-209, 406-418.

Salee, W. and Pongsapukdee, V. (2013) Odds prediction of drought category using loglinear models based on SPI in the northeast of Thailand. *Silpakorn University Science and Technology Journal* 7(1): 32-40.

Wang, W. and Famoye, F. (1997) Modeling household fertility decisions with generalized Poisson regression. *Journal of Population Economics* 10: 273-283.

Yamjaroenkit, M. and Pongsapukdee, V. (2012) Tests of dispersion parameter in generalized Poisson regression models. In *Proceeding of Silapakorn University International Conference on Academic Research and Creative Arts: Integration of Art and Science*. The Art and Culture Center Commemorating the 6th Cycle Birthday Anniversary of His Majesty The King, Silpakorn University, NakonPathom, Thailand.

Yamjaroenkit, M. (2012) *Tests of Dispersion Parameter in Generalized Poisson Regression Models*. M.Sc. thesis, Department of Statistics, Silpakorn University.

Yang, Z., Hardin, J. W. and Addy, C. L. (2009) A score test for overdispersion in Poisson regression based on the generalized Poisson-2 model. *Journal of Statistical Planning and Inference* 139: 1514-1521.