Applied Mathematical Sciences, Vol. 4, 2010, no. 13, 623 - 635

On Estimating Parameters of Censored Generalized Poisson Regression Model

Marie M. Mahmoud and Mahmoud M. Alderiny

Dept. Of Statistics & Mathematics
Faculty of Comm., Tanta University, Egypt
marie_tanta@yahoo.com

Abstract

When the sampling variance of a count variable Y is significantly greater or less than that predicted by an expected probability distribution, Y is said to be over- or underdispersed, respectively. A natural way to analyze regular count data is to use a Poisson regression(PR) model where the Poisson mean can be modeled as a function of linear predictors through the log link function in a generalized linear model (GLM) setting. The generalized Poisson regression (GPR) model is a generalization of the standard (PR) model. When the dispersion parameter $\alpha = 0$, the probability function reduces to the PR model. When $\alpha > 0$, the GPR model represents count data with overdispersion and when $\alpha < 0$, the GPR model represents count data with underdispersion. In this paper a random sample of workers selected from workers of Shebeen Alkom textile industry in Egypt, 2007. The data in the sample has information on many variables including dependent variable, and nine independent variables. The Censored generalized Poisson regression (CGPR) model is considered for identifying the relationship between the dependent, and the previous independent variables. Based on the test for the dispersion parameter and the goodness-of-fit measure for the dependent variable, the (CGPR) model performs as good as or better than the other regression models.

Keywords: Generalized Poisson regression; Mixture model; Maximum likelihood estimation (MLE); The Censored generalized Poisson regression(CGPR) model, Goodness-of-fit measure

1. Introduction

The Poisson regression (PR) model has been widely used for the analysis count data. The Poisson distribution was first used in regression context by letting the mean parameter μ in the Poisson distribution depend on some covariates (Frome, Kutner, and Beauchamp, 1973). When the mean and variance are equal, this is called equidispersion (Cameron and Trivedi, 1998 p. 4). The assumption of equality of the mean and variance of the Poisson model may not hold for some real life applications as count data often show variations where the sample mean may be greater than or smaller than the sample variance. When the sample variance is greater than the sample mean, this is referred to as overdispersion. Underdispersion occurs when the sample variance is less than the sample mean.

Famoye (1993) derived the generalized Poisson regression (GPR) model from the generalized Poisson distribution introduced by Consul and Jain (1973). These distribution can handle count data that is underdispersed, overdespersed and equidispersed. The (GPR) model was applied by Famoye (1993) to data on the number of faults in rolls of fabric studied previously by Hinde (1982).

Wang & Famoye (1997) analyzed data set on fertility from Michigan Panel Study of income Dynamics (PSID) by using Poisson Regression (PR), generalized Poisson Regression (GPR) and censored generalized Poisson Regression (CGPR) model. From set of 5500 household they selected married women aged between 18 and 40 who are not head of households and with nonnegative family income. In this paper the dependent variable, the total number of children up to 17 years old in family, is nonnegative integer ranging from zero to nine in the sample. Although the sample mean less than the sample variance of the dependent, they suggested that the data may be equi-dispersed and thus either the (PR) or the (GPR) model will be adequate for analyzing the data. The purpose of using (CGPR) model is to demonstrate censoring and not to show which independent variable is significant, such that about 4.22% of the sample have dependent variables, so that all values of considered.

Famoye & Wang (2004) suggested the use of censored generalized Poisson regression (CGPR) model. They applied the (CGPR) model to a data set on fertility from the Michigan Panel Study of Income Dynamics and compared it to the Poisson, truncated Poisson, (GPR), and censored Poisson regression models. The censored generalized Poisson regression (CGPR) model illustrated how estimates of parameters can be greatly improved if censoring is considered in data set that is overdispersed or underdispersed.

Cameron& Trivedi (1998) were suggested models allowing for censoring count data. These models are required when the sample variance is different from the sample mean and the response Poisson variable y_i belong to restricted range, while the explanatory variables $(x_{i1}, x_{i2}, ..., x_{i,p-1})$ are always observed.

2. Censored Generalized Poisson Regression(CGPR) Model

Suppose a count dependent variable Y_i is a generalized Poisson random variable and affected by $p-1$ explanatory variables, $(x_{i1}, x_{i2}, ..., x_{i,p-1})$. The generalized Poisson regression (GPR) model derived by Famoye (1993) is that the distribution of Y_i, conditional on explanatory variables $(x_{i1}, x_{i2}, ..., x_{i,p-1})$, and it is defined by

$$p(Y_i = y_i \mid \mathbf{x}_i) = f(y_i) = \frac{\mu_i}{1+\alpha\mu_i}\left(\frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right)^{y_i-1} * \exp\left[\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right]\frac{1}{y_i!}, \quad y_i = 0,1,... \qquad (2.1)$$

Where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{i,p-1})'$ is a $p \times 1$ dimensional vector, $\mu_i > 0$ is the conditional mean of Y_i on $\mathbf{x}_i$.One specification that is mostly used for the mean parameter μ_i is the exponential specification, it is given by

$$E(Y_i \mid \mathbf{x}_i) = \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), \qquad (2.2)$$

The conditional variance is given as

$$Var(Y_i \mid \mathbf{x}_i) = \sigma^2_{y|X} = \mu_i(1+\alpha\mu_i)^2 \qquad (2.3)$$

Where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})'$ is a p-dimensional vector of regression parameters. The parameter α is a measure of dispersion. When $\alpha = 0$, the (GPR) model in (2.1) reduces to the (PR) model. For $\alpha > 0$, the (GPR) model can used to fit overdispersed count data. When $\alpha < 0$, the (GPR) model can used to fit underdispersed count data.

For some observations in data set, the value of Y_i may be censored. If no censoring occurs for the *i*th observation, $Y_i = y_i$. However, If censoring occurs for the *i*th observation, Y_i is at least equal to y_i i.e. $Y_i \geq y_i$.When the data are censored, the distribution that applies to the sample data is derived by using binary variable d_i, this variable is defined as:

$$\begin{cases} d_i = 1 \ \ if \ \ Y_i \geq y_i \\ d_i = 0 \ \ otherwise \end{cases} \qquad (2.4)$$

then (CGPR) model introduced by Famoye and Wang (2004) is given by

$$p(y_i, d_i \mid \mathbf{x}_i) = [f(y_i)]^{1-d_i}\left[1 - \sum_{j=0}^{y_i-1} f(j)\right]^{d_i} \qquad (2.5)$$

The (CGPR) model (2.5) includes $(p+1)$ parameters arrayed in the vector, $\boldsymbol{\Phi} = (\boldsymbol{\beta}', \alpha)'$, and it can be estimated by using maximum likelihood method.

3. Parameter Estimation

The maximum likelihood method is used to estimate the parameter vector $\boldsymbol{\Phi} = (\boldsymbol{\beta}', \alpha)'$. The likelihood function of (CGPR) model (2.5) (Famoye and Wang, 2004) is given by

$$L(\boldsymbol{\beta}, \alpha; y_i) = \prod_{i=1}^{n} \left\{ [f(y_i)]^{1-d_i} \left[1 - \sum_{j=0}^{y_i - 1} f(j) \right]^{d_i} \right\} \qquad (3.1)$$

and the log-likelihood function is

$$LL(\boldsymbol{\beta}, \alpha; y_i) = \sum_{i=1}^{n} [(1-d_i)\log(f(y_i))] + \sum_{i=1}^{n} \left[d_i \log\left(1 - \sum_{j=0}^{y_i - 1} f(j) \right) \right] \qquad (3.2)$$

By using the probability function given by (2.1) in the log-likelihood function in (3.2), we obtain

$$LL(\boldsymbol{\beta}, \alpha; y_i) = \sum_{i=1}^{n} \left\{ (1-d_i) \left[y_i \log\left(\frac{\mu_i}{1+\alpha\mu_i} \right) + (y_i - 1)\log(1+\alpha y_i) - \right. \right.$$

$$\left. \left. \frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i} - \log(y_i!) \right] \right\} + \sum_{i=1}^{n} \left\{ d_i \log\left[1 - \sum_{j=0}^{y_i - 1} f(j) \right] \right\} \qquad (3.3)$$

The likelihood equations for estimating $\boldsymbol{\beta}$ and α are obtained by taking the partial derivations of (3.3) and setting them equal to zero. Thus, we obtain

$$\frac{\partial LL(\boldsymbol{\beta}, \alpha; y_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ (1-d_i) \left(\frac{y_i - \mu_i}{(1+\alpha\mu_i)^2} \right) \mathbf{x}_i \right\} - \sum_{i=1}^{n} \left\{ d_i \frac{\sum_{j=0}^{y_i - 1} \left(\frac{\partial f(j)}{\partial \boldsymbol{\beta}} \right)}{\left(1 - \sum_{j=0}^{y_i - 1} f(j) \right)} \right\} = 0 \qquad (3.4)$$

$$\frac{\partial LL(\boldsymbol{\beta}, \alpha, y_i)}{\partial \alpha} = \sum_{i=1}^{n} \left\{ (1-d_i) \left(\frac{-\mu_i y_i}{(1+\alpha\mu)} + \frac{y_i(y_i - 1)}{1+\alpha y_i} - \frac{\mu_i(y_i - \mu_i)}{(1+\alpha\mu)^2} \right) \right\} -$$

$$\sum_{i=1}^{n} \left\{ d_i \frac{\sum_{j=0}^{y_i - 1} \left(\frac{\partial f(j)}{\partial \alpha} \right)}{\left(1 - \sum_{j=0}^{y_i - 1} f(j) \right)} \right\} = 0 \qquad (3.5)$$

where,

$$\frac{\partial f(j)}{\partial \boldsymbol{\beta}} = f(j) \frac{(j - \mu_i)}{(1+\alpha\mu_i)^2} \mathbf{x}_i \quad , \qquad (3.6)$$

$$\frac{\partial f(j)}{\partial \alpha} = f(j)\left(-\frac{j\mu_i}{(1+\alpha\mu)} + \frac{j(j-1)}{1+\alpha j} - \frac{\mu_i(j-\mu_i)}{(1+\alpha\mu)^2} \right) \quad (3.7)$$

The above likelihood equations are non-linear in parameters $\boldsymbol{\beta}$ and α. These equations are solved simultaneously by using an iterative algorithm. The Statistical Analysis Software (SAS9.1, 2002-2003) can be used to carry out Newton-Raphson method for solving these equations. The initial estimate of $\boldsymbol{\beta}$ and α may be taken as the corresponding final estimates of $\boldsymbol{\beta}$ and α from fitting a generalized Poisson regression model to the data.

On taking the second partial derivatives of (3.3), the Fisher`s information matrix $I(\boldsymbol{\beta}, \alpha)$ can be obtained by taking the expectations of minus the second derivatives. The inverse of $I(\boldsymbol{\beta}, \alpha)$ matrix will provide the variances of the maximum likelihood estimates. The variance of the maximum likelihood estimates can also be obtained from Hessian matrix, H, which is a square matrix of order $(p+1)$. The entries of the Hessian matrix, denoted by the second order partial derivatives of (3.3), and given by

$$H(\Phi) = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} \left(\frac{\partial^2 LL(\boldsymbol{\beta}, \alpha, y_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right) & \left(\frac{\partial^2 LL(\boldsymbol{\beta}, \alpha, y_i}{\partial \boldsymbol{\beta} \partial \alpha} \right) \\ \left(\frac{\partial^2 LL(\boldsymbol{\beta}, \alpha, y_i}{\partial \boldsymbol{\beta} \partial \alpha} \right) & \left(\frac{\partial^2 LL(\boldsymbol{\beta}, \alpha, y_i}{\partial \alpha^2} \right) \end{bmatrix} \quad (3.8)$$

where

$$H_{11} = \frac{\partial^2 LL(\boldsymbol{\beta}, \alpha; y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \left\{ (1-d_i) \left(\frac{1-\alpha\mu_i + 2\alpha y_i}{(1+\alpha\mu_i)^3} \right) \mu_i \mathbf{x}_i \mathbf{x}_i' \right\} -$$

$$\sum_{i=1}^{n} \left\{ d_i \left[\frac{\left(1 - \sum_{j=0}^{y_i-1} f(j) \right) \sum_{j=0}^{y_i-1} \left(\frac{\partial^2 f(j)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right) + \sum_{j=0}^{y_i-1} \left(\frac{\partial f(j)}{\partial \boldsymbol{\beta}} \right) \sum_{j=0}^{y_i-1} \left(\frac{\partial f(j)}{\partial \boldsymbol{\beta}'} \right)}{\left(1 - \sum_{j=0}^{y_i-1} f(j) \right)^2} \right] \right\} \quad , \quad (3.9)$$

$$H_{12} = \frac{\partial^2 LL(\boldsymbol{\beta}, \alpha; y_i)}{\partial \boldsymbol{\beta} \partial \alpha} = -2\sum_{i=1}^{n} \left\{ (1-d_i) \frac{\mu_i (y_i - \mu_i)}{(1+\alpha\mu_i)^3} \mathbf{x}_i \right\} -$$

$$\sum_{i=1}^{n} \left\{ d_i \left[\frac{\left(1 - \sum_{j=0}^{y_i-1} f(j) \right) \sum_{j=0}^{y_i-1} \left(\frac{\partial^2 f(j)}{\partial \boldsymbol{\beta} \partial \alpha} \right) + \sum_{j=0}^{y_i-1} \left(\frac{\partial f(j)}{\partial \boldsymbol{\beta}} \right) \sum_{j=0}^{y_i-1} \left(\frac{\partial f(j)}{\partial \alpha} \right)}{\left(1 - \sum_{j=0}^{y_i-1} f(j) \right)^2} \right] \right\} \quad , \quad (3.10)$$

$$H_{22} = \frac{\partial^2 LL(\boldsymbol{\beta}, \alpha; y_i)}{\partial \alpha^2}$$
$$= \sum_{i=1}^{n} \left\{ (1-d_i) \frac{\mu_i^2 y_i}{(1+\alpha\mu_i)^2} - \frac{y_i^2 (y_i - 1)}{(1+\alpha y_i)^2} + \frac{2\mu_i^2 (y_i - \mu_i)}{(1+\alpha\mu_i)^3} \right\} -$$
$$\sum_{i=1}^{n} \left\{ d_i \left[\frac{\left(1 - \sum_{j=0}^{y_i - 1} f(j)\right) \sum_{j=0}^{y_i - 1} \left(\frac{\partial^2 f(j)}{\partial \alpha^2} \right) + \sum_{j=0}^{y_i - 1} \left(\frac{\partial f(j)}{\partial \alpha} \right)^2}{\left(1 - \sum_{j=0}^{y_i - 1} f(j)\right)^2} \right] \right\}, \quad (3.11)$$

$$H_{21} = H_{12}' . \quad (3.12)$$

$$\frac{\partial^2 f(j)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = f(j) \left[- \left(\frac{1 - \alpha\mu_i + 2\alpha j}{(1+\alpha\mu_i)^3} \right) \mu_i + \left(\frac{(j - u_i)}{(1+\alpha\mu_i)^2} \right)^2 \right] \mathbf{x}_i \mathbf{x}_i', \quad (3.13)$$

$$\frac{\partial^2 f(j)}{\partial \boldsymbol{\beta} \partial \alpha} = f(j) \left\{ \left(\frac{-2\mu_i (j - \mu_i)}{(1+\alpha\mu_i)^3} \right) + \left(\frac{(j-u_i)}{(1+\alpha\mu_i)^2} \right) \left(\frac{-\mu_i j}{(1+\alpha\mu_i)} + \frac{j(j-1)}{(1+\alpha j)} - \frac{\mu_i (j - \mu_i)}{(1+\alpha\mu_i)^2} \right) \right\} \mathbf{x}_i \quad (3.14)$$

$$\frac{\partial^2 f(j)}{\partial \alpha^2} = f(j) \left\{ \left(\frac{\mu_i^2 j}{(1+\alpha\mu_i)^2} - \frac{j^2 (j-1)}{(1+\alpha j)^2} + \frac{2\mu_i^2 (j - \mu_i)}{(1+\alpha\mu_i)^3} \right) + \left(\frac{-\mu_i j}{(1+\alpha\mu_i)} + \frac{j(j-1)}{(1+\alpha j)} - \frac{\mu_i (j - \mu_i)}{(1+\alpha\mu_i)^2} \right)^2 \right\} \quad (3.15)$$

When the Hessian matrix is evaluated at maximum likelihood estimates, $\hat{\boldsymbol{\Phi}} = (\hat{\boldsymbol{\beta}}', \hat{\alpha})'$, and negative of its inverse taken, then the variance-covariance matrix denoted by $S^2\{\hat{\boldsymbol{\beta}}, \hat{\alpha}\} = [-H(\hat{\Phi})]^{-1}$ is obtained.

4. Goodness-of-fit Statistics

For testing the goodness-of-fit of (CGPR) model, we can be applied the likelihood ratio to test the hypothesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0 \quad (4.1)$$

and the likelihood ratio test has the intuitive form

$$LR = -2\left(LL(\hat{\beta}_0, \hat{\alpha}; y_i)_R - LL(\hat{\boldsymbol{\beta}}, \hat{\alpha}; y_i)_U \right) \quad (4.2)$$

where $LL(\hat{\boldsymbol{\beta}}, \alpha_k; y_i)_U$ & $LL(\hat{\beta}_0, \alpha_k; y_i)_R$ are the computed log-likelihood function with using the unrestricted and restricted model. Under the null hypothesis (4.1), the test statistics LR in (4.2) follows a χ^2 distribution with $(p-1)$ degrees of freedom.

5. Test for regression coefficients and dispersion parameter

The maximum likelihood estimates $\hat{\boldsymbol{\Phi}}=(\hat{\boldsymbol{\beta}}',\hat{\alpha})'$ maximize the log-likelihood function (3.3). If the (CGPR) model has been specified correctly, then $\hat{\boldsymbol{\Phi}} = (\hat{\boldsymbol{\beta}}',\hat{\alpha})'$ is consistent for $\boldsymbol{\Phi}=(\boldsymbol{\beta}',\alpha)'$ and the asymptotic normality result $\sqrt{n}(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}) \rightarrow N\left[0, \left[-E\left((1/n)I(\hat{\boldsymbol{\beta}},\hat{\alpha})\right)\right]^{-1}\right]$ Thus inference on the regression coefficients and dispersion parameter, α, can be made. The (CGPR) model reduces to the censored Poisson regression model when the dispersion parameter $\alpha = 0$. To assess the adequacy of the (CGPR) model over the censored Poisson model, we can test the hypothesis

$$H_0 : \alpha = 0 \quad \textbf{against} \quad H_a : \alpha \neq 0 \qquad (5.1)$$

This is to test for the significance of the dispersion parameter α. The presence of the dispersion parameter α in the (CGPR) model is justified when the null hypothesis $H_0 : \alpha = 0$ is rejected. The test statistics for testing null hypothesis (4.3) is given by

$LR_\alpha = -2\left(LL(\hat{\boldsymbol{\beta}}; y_i)_R - LL(\hat{\boldsymbol{\beta}},\alpha; y_i)_U\right)$ (5.2) When the null hypothesis (4.2) is true, the likelihood ratio test statistic, LR_α, in (4.4) is approximately chi-squire distributed with one degree of freedom. Also, to test the significance of coefficient of explanatory variable x_j, β_j, $j = 1,2,..., p-1$, the hypothesis denoted as

$$H_0 : \beta_j = 0 \quad \textbf{against} \quad H_a : \beta_j \neq 0 \qquad (5.3)$$

and the test statistics for testing null hypothesis (4.5) is given by

$$Z = \frac{\hat{\beta}_{j_{mle}}}{S(\hat{\beta}_{j_{mle}})} \qquad (5.4)$$

Where $\hat{\beta}_{j_{mle}}$ is the maximum likelihood estimate of coefficient β_j, $S(\hat{\beta}_{j_{mle}})$ is the standard error of these estimation, determined from the estimation of the variance covariance matrix, $S^2(\hat{\boldsymbol{\beta}},\hat{\alpha})$. Under the null hypothesis (4.5) the test statistic Z is approximately standard normal distributed.

6. Description of worker's absent days data

A random sample of size 80 workers selected from workers of Shebeen Alkom textile industry in Egypt, 2007. A questionnaire was formed in order to obtain the data about dependent and independent variables. The dependent variable data includes the number of absent days for worker, y, in September month which was selected randomly. The independent variables data contains daily wage by Egyptian pound x_1, worker's age x_2, experience years x_3, family size x_4, number of rooms in worker's house x_5, motivations system x_6, { $x_6 = 1$ if system is suitable, $x_6 = 0$ if system is not

suitable}, department x_7 ,{ x_7 =1 if worker belong to textile department, x_7 =0 if worker belong to other department}, industry trips x_8 ,{ x_8 =1 if worker associates to industry trips, x_8 = 0 if worker not associate to industry trips}, and the reason for working in industry x_9 , { x_9 =1 if there is no other source for income, x_9 =0 if there is other source for income}.Table (1) shows the means and the standard deviations for quantitative variables data, $(y, x_1, x_2, x_3, x_4, x_5)$.

Table (1)

Quantitative Variables	mean	St.dev.
Number of absent days y	2.725	1.518
Worker's wage x_1	3.871	1.008
Worker's age x_2	43.013	6.530
Experience years x_3	24.525	6.882
Family size x_4	6.375	1.694
number of rooms in house x_5	3.338	0.967

The dependent variable, the number of absent days in September month, is nonnegative integer ranging from zero to eight in the sample. From table (1), the sample variance of number of absent days y ,2.304, is less than the sample mean, 2.753, then we expected that underdispersion occurs and dispersion parameter, α , may be negative. About 3.75% of the sample has dependent variable $y_i \geq 6$, and then we consider all values $y_i \geq 6$ as censored and applied censored Poisson regression and CGPR models to the data. Table (2) displays the observed frequencies and the percentiles for the levels of qualitative variables data, (x_6, x_7, x_8, x_9).

Table (2)

Var.		count	percentile
x_6	Motivations system is suitable	35	44%
	Motivations system is not suitable	45	56%
x_7	Belong to textile department	43	54%
	Belong to other department	37	46%
x_8	Worker associates to industry trips	15	19%
	Worker not associate to industry trips	65	81%
x_9	Existing other source for income	60	75%
	Not Existing other source for income	20	25%

7.Results and conclusions

In comparing the sample mean 2.725 of the dependent variable to its sample variance 2.304, and about 3.75% of the sample has dependent variable $y_i \geq 6$, the data suggests a case of under-dispersion and considers all values of $y_i \geq 6$ as censored. The parameter estimates and their standard errors using the CPR and the CGPR models are given in table (3).

	CPRM				CGPRM			
Parameters	**Estim.**	**S.E**	**Z**	**p. v.**	**Estim.**	**S.E**	**Z**	**p. v.**
β_0	**1.169**	**0.636**	**1.839**	**0.066**	**1.175**	**0.390**	**3.016**	**0.003**
β_1	**-0.337**	**0.125**	**-2.709**	**0.007**	**-0.336**	**0.092**	**-3.664**	**0.000**
β_2	**0.053**	**0.026**	**2.024**	**0.043**	**0.047**	**0.013**	**3.517**	**0.000**
β_3	**-0.059**	**0.025**	**-2.349**	**0.019**	**-0.048**	**0.014**	**-3.477**	**0.001**
β_4	**0.050**	**0.056**	**0.896**	**0.370**	**0.045**	**0.032**	**1.398**	**0.162**
β_5	**-0.021**	**0.084**	**-0.254**	**0.799**	**-0.029**	**0.052**	**-0.561**	**0.575**
β_6	**-0.081**	**0.152**	**-0.531**	**0.596**	**-0.064**	**0.095**	**-0.680**	**0.496**
β_7	**-0.125**	**0.157**	**-0.792**	**0.428**	**-0.121**	**0.096**	**-1.261**	**0.207**
β_8	**-0.039**	**0.198**	**-0.195**	**0.846**	**-0.040**	**0.133**	**-0.304**	**0.761**
β_9	**0.139**	**0.171**	**0.812**	**0.417**	**0.135**	**0.114**	**1.184**	**0.237**
α					**-0.125**	**0.015**	**-8.210**	**0.000**
$LL_U(H_1)$	**-123.7314**				**-111.7175**			
$LL_R(H_0)$	**-140.5339**				**-139.2596**			
χ^2	**33.6049**				**55.0842**			
P_V	**0.0001**				**0.0000**			

From table (3), the estimated dispersion parameter from the CGPR model is negative, which is an indication of under-dispersion. The asymptotic '' Z '' statistics for testing the null hypothesis in (4.3) is approximately -8.21. Thus, the dispersion parameter α is significantly different from zero (1% level). The CPR model is not appropriate for this data since we reject the null hypothesis given in (4.3). The log-likelihood values for CPR and CGPR models are -123.73 and -111.72, respectively, which also indicate that modeling under-dispersed data using CGPR model is more appropriate than the CPR model.

The three independent variables (worker's wage x_1, worker's age x_2, experience years x_3) are significant under CPR model at 5% level but they significant under CGPR model at 1% level. The family size x_4, is significant under CGPR model at 20% level but this is not the case under the CPR model. The parameters estimates from both models are very similar. The standard errors from the CGPR model are smaller than the standard errors from the CPR model, so that the standard errors from the CGPR model are more appropriate in this case, because these model accounts for the under-dispersion exhibited by the data.

From CGPR model, at 1% level, we note that, the effect of each of worker's wage x_1 and experience years x_3 is statistically significant and is negatively associated with the number of absent days. This implies that the mean of absent days decreases with respect to each of large wages and experience years. At 1% level, the effect of worker's age x_2 is statistically significant and is positively associated with the number of absent days, this implies that the mean of absent days increases with respect to the larger ages. Family size x_4, has positive effect at 20% level, and means that, the mean of absent days increases with respect to the larger family size. The type of department which denoted by binary variable x_7 has negative effect at 25% level, and means that, the mean of worker's absent days in the textile department, $x_7 = 1$, is smaller than the mean of worker's absent days in the other departments , $x_7 = 0$. At 25% level, the reason for working in industry denoted by variable x_9 is positively associated with the number of absent days, this implies that the mean of absent days increases when the worker has got other source for income. The estimate of mean of days is given by

$$\hat{\mu}_i = \exp(1.169 - 0.337 x_1 + 0.053 x_2 - 0.059 x_3 + 0.05 x_4 - 0.021 x_5 - 0.081 x_6 - 0.125 x_7 - 0.039 x_8 + 0.139 x_9) \qquad (7.1)$$

For testing the goodness-of-fit of suggested (CGPR) model, we note from table 3 that the computed chi-square, $\chi^2 = 55.084$ degrees, df = 9, and $p-value < 0.01$. Thus, we can reject the null hypothesis, $H_0 : \beta_1 = \beta_2 = ... = \beta_9 = 0$, at 1% level, and this means, the suggested (CGPR) model which has estimated mean denoted by equation (7.1) is more appropriate for showing the relation between the worker's absent days as a dependent variable and the independent variables under studying ,(daily wage, worker's age, experience years, family

size, number of rooms in worker's house, motivations system, type of department, industry trips, and the reason for working in industry).

7. Conclusions

If we believe the Poisson regression (PR) model, then we have, $E(Y_i \mid \mathbf{x}_i)=Var(Y_i \mid \mathbf{x}_i)$ implying that the conditional mean function equals the conditional variance function. If $E(Y_i \mid \mathbf{x}_i) < Var(Y_i \mid \mathbf{x}_i)$;respectively $E(Y_i \mid \mathbf{x}_i) > Var(Y_i \mid \mathbf{x}_i)$; then we speak about overdispersion, respectively underdispersion. The Poisson model does not allow for over- or underdispersion. A richer model is obtained by using the generalized Poisson distribution instead of the Poisson distribution.

In statistics, *truncation* occurs when only those values which lie in a certain region are observed. This phenomenon is related to but differs from *censoring*, whereby particular sample values are known only to lie in a certain region. Thus, under censoring the number of unobserved values is known, whereas under truncation that number is unknown. Truncation and censoring may both be thought of as examples of non-ignorable non-response or more generally as examples of biased sampling. In the absent days data for workers in textile industry, the observed percentages of values ≥ 6 are, respectively, 3.75%. Also, the data is under-dispersed which indicates that the PR and CPR models are not appropriate either. To model underdispersion, the CGPR model discussed in section 2 is among the suitable model. We applied the CGPR model to the data and found that the parameters estimate more efficient than the similar to that of CPR model. Thus, we decided to exclude the parameter estimates of the CPR model to save space in the paper.

In summary, the estimated dispersion parameter from the data is negative and it is significantly different from zero. Based on the goodness-of-fit measure for the absent days data, the CGPR model seems to perform better than the CPR model in identifying daily wage, worker's age, experience years, family size, number of rooms in worker's house, motivations system, department, industry trips and the reason for working in industry use associated with the number of absent days for worker. Additional studies should be conducted in order to identify the important independent variables that may be accounted to plane the structural system and social factors for workers in industry. Worker's wage, age, experience years are important explanatory variables.

References

1- A.C. Cameron and P.K. Trivedi, Regression analysis of count data, Cambridge University Press, (1998).

2- C.Lange and J. C. Whittaker, Mapping quantitative trait loci using generalized estimating equations, Genetics, 159(2001), 1325–1337.

3- F. Famoye, Restricted generalized Poisson regression model, Communications in Statistics- Theory and Methods, 22(5) (1993), 1335–1354.

4- F. Famoye, J. T. Wulu and K. P. Singh, On the Generalized Poisson Regression Model with an Application to Accident Data, Journal of Data Science, 2(2004), 287-295.

5-F. Famoye and W.R. Wang, Censored generalized Poisson regression model, Computational Statistics and Data Analysis, 46 (2004), 547–560.

6- G. Q. Dlao, D. Y. Lin and F. Zou, Mapping quantitative trait loci with censored observations, Genetics 168(2004), 1689–1698.

7- J. Hinde, Compound Poisson regression models, Proceedings of the International Conference on Generalized Linear Models, Gilchrist, R., N.Y. GLIM 82(1982), 109-121.

8- P. Mccullagh and J. A. Nelder, Generalized Linear Models, Chapman & Hall, London (1989)

9- P.Thomson, A generalized estimating equations approach to quantitative trait locus detection of non-normal traits, Genet. Sel. Evol., 35 (2003), 257– 280.

10- W.H. Greene, Econometric analysis, third edition, Prentice-Hall International Inc, N.J. (1997)

11- W.R. Wang and F. Famoye, Modeling household fertility decisions with generalized Poisson regression, Journal of Population Economics,10(3) (1997), 273-283.