Eur. Actuar. J. (2012) 2:187–204
DOI 10.1007/s13385-012-0056-2

ORIGINAL RESEARCH PAPER

Poisson regression and Zero-inflated Poisson regression: application to private health insurance data

Younès Mouatassim · El Hadj Ezzahid

Received: 13 April 2012 / Revised: 16 August 2012 / Accepted: 14 September 2012 /
Published online: 17 October 2012

Abstract Modeling event counts is important in many fields. For this purpose, the Poisson regression model is often used. However, this model assumes the equidispersion of the data. Unfortunately, this assumption is often violated in the observed data. The source of overdispersion depends on many situations. When the source of overdispersion is the excess of zeroes, the Zero-inflated Poisson regression model fits better counts data. In this paper, we first review the theoretical framework of Poisson regression and Zero-inflated Poisson regression. The probability integral transform test and the Vuong's test are used to compare between the two models. Second, we fit these models to the number of claims in a private health insurance scheme. In our case, the number of claims is overdispersed because of the preponderance of zeroes in the data set. The results prove that Zero-inflated Poisson regression performs better the number of claims of the customers affiliated in the health insurance scheme in the Moroccan case.

Keywords Poisson model · Zero-inflated Poisson model · Excess of zeroes · Vuong's test · Probability integral transform · Goodness-of-fit · Health insurance

1 Introduction

The modeling of counts data is of a primary interest in many fields such as insurance, public health, epidemiology, psychology, and many other research areas.

Y. Mouatassim (✉)
Zurich Insurance, Casablanca, Morocco
e-mail: younes.mouatassim@gmail.com

E. H. Ezzahid
Faculty of Law and Economics, Mohammed V-Agdal University,
Rabat, Morocco
e-mail: ezzahidelhadj@yahoo.fr

Poisson distribution is widely assumed for modeling the distribution of the observed counts data [14], [22]. It assumes that the mean and variance are equal. However, this restriction is violated in many applications because data are often overdispersed. In this case, Poisson distribution underestimates the dispersion of the observed counts. The overdispersion occurs when the single parameter λ of Poisson distribution is unable to fully describe event counts. Generally, two sources of overdispersion are determined: heterogeneity of the population and excess of zeroes. The heterogeneity is observed when the population can be divided into many homogeneous subpopulations. The excess of zeroes is detected when the number of observed zeroes exceeds largely the number of zeroes reproduced by the fitted Poisson distribution. The existence of overdispersion is an incentive to seek alternative models that suit better the data.

Because of the restrictive nature of equidispersion assumption in standard Poisson model, researchers have developed techniques and tests that allow detecting the overdispersion (or under-dispersion) in the population. For more details, one can refer to [3], [4], [7], [9], [12–14], [16], [21], [25]. Most of the overdispersion tests do not care about its source. They cannot help to choose the alternative models.

Recently, Zero-inflated models have been developed to take into account the excess of zeroes in the data [15] has introduced the Zero-inflated Poisson regression in his paper entitled "Zero-inflated Poisson regression, with an application to defects in manufacturing". The Zero-inflated model can be seen as a finite mixture model where one distribution is considered as a degenerate process with a unit point mass at zero. In fact, the model divides population into two groups: the first one contains zero-outcomes and the second contains the nonzero outcomes. In case of Zero-inflated Poisson regression, the zero outcomes are modeled by including a proportion $1-p$ of extra zeroes and a proportion $p\exp(-\lambda_i)$ of zeroes generated from the Poisson distribution. The nonzero counts, in their turn, are modeled using zero-truncated Poisson model.

In this paper, we will fit Poisson regression and Zero-inflated Poisson regression to data set related to a private health insurance scheme. After estimating the models, we will calculate the probability integral transform and the Vuong's test in order to select the model which fits better the data. The remaining of this paper is organized as follows. In Sect. 2, we specify the Poisson regression, we present the estimation of its parameters by maximum likelihood and we introduce overdispersion tests. The zero-inflated Poisson regression is presented in Sect. 3. In the fourth and fifth sections, we present tests allowing the goodness of fit and the comparison between models, respectively. The results of applying Poisson regression and Zero-inflated Poisson regression to model the number of claims in private health insurance are exposed in Sect. 6. Finally, we provide some concluding remarks.

2 Poisson regression

Poisson regression is generally used to model counts data. It assumes that the response variable has a Poisson distribution and the logarithm of its expected value

can be modeled by a linear combination of unknown parameters. Let y_i be the response variable. We assume that y_i follows a Poisson distribution with mean λ_i, defined as a function of covariates x_i . Thus, the Poisson regression is given by the equation below:

$$P(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

where the conditional mean is specified by $\lambda_i = E(y_i|x_i) = exp(x'\beta)$. The vector $x_i' = x_{i,1}, x_{i,2}, \ldots, x_{i,P}$ contains the covariates and $\beta' = (\beta_1, \beta_2, \ldots, \beta_P)$ is the vector of unknown parameters. The number P defines the dimension of the vector of the covariates incorporated in the model.

Maximum likelihood techniques may be used to estimate the parameters of the Poisson regression. Given the assumption that the observations $(y_i|x_i)$ are independent, the log-likelihood function is given by:

$$\ln L(\beta) = \sum_{i=1}^{n} [y_i x'\beta - \exp(x_i'\beta) - \ln(y_i!)]$$

The likelihood equations are:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} (y_i - \lambda_i)x_i = 0$$

Therefore, the Hessian is:

$$\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \lambda_i x_i x_i'$$

The Hessian of the model is negative for all x and β. The log-likelihood function is, then, concave. Hence, Newton–Raphson iterative algorithm will converge rapidly and provide unique parameters estimate. The estimator of the asymptotic covariance matrix is given by:

$$\mathrm{Var}(\hat{\beta}) = \left(\sum_{i=1}^{n} \hat{\lambda}_i x_i x_i'\right)^{-1}$$

The hypothesis tests of the nullity of a single parameter or a set of parameters simultaneously can be carried using Wald test, Lagrange Multiplier test or Likelihood Ratio test. The Wald statistic is given by:

$$W = \hat{\beta}' \left(\sum_{i=1}^{n} \hat{\lambda}_i x_i x_i'\right)^{-1} \hat{\beta}$$

Under the null hypothesis, the Wald statistic follows a Chi-square with one degree of freedom. The Likelihood Ratio statistic is given by:

$$LR = 2\sum_{i=1}^{n} \ln\left(\frac{\hat{P}_i}{\hat{P}_{\mathrm{Restricted},i}}\right)$$

The Lagrange Multiplier statistic of Poisson model is given by:

$$\mathrm{LM} = \left[\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}(\mathrm{y_i} - \hat{\lambda}_i)\right]' \left[\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i x_i'}(\mathrm{y_i} - \hat{\lambda}_i)^2\right]^{-1} \left[\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}(\mathrm{y_i} - \hat{\lambda}_i)\right]$$

where λ_i' is computed using a restricted model. The LM statistic is to compare with a Chi-square with one degree of freedom.

The interpretation of a Poisson model differs according to the goals of the study. A researcher can be interested in the expected counts or in the distribution of counts. When the analysis of the expected value is the aim of the study, several measures namely the partial effects, the factor change and/or the percentage change can be computed to assess the change of the expected value for a change in an independent variable (i.e. a covariate) keeping other variables constant. If the interest is in the distribution of counts or just the probability of a specific count, the probability of a count for a given level of the independent variables can be computed [17].

The *partial effect* of $\mathrm{E(y|x)}$ with respect to $\mathrm{x_k}$ is given by:

$$\frac{\partial \mathrm{E(y|x)}}{\partial \mathrm{x_k}} = \beta_{\mathrm{k}} \exp(\mathrm{x}'\beta) = \mathrm{E(y|x)}\beta_{\mathrm{k}}$$

It is clear that the partial effects in Poisson models depend on both the coefficient of $\mathrm{x_k}$, that is β_{k}, and the value of the expected value of y given x. Therefore, partial effects of non linear models cannot be interpreted as a change of the expected value for a unit change in $\mathrm{x_k}$ as in linear models.

The *factor change* in $\mathrm{E(y|x)}$ for a change δ in $\mathrm{x_k}$holding all other factors constant is given by:

$$\frac{\mathrm{E(y|x, x_k} + \delta)}{\mathrm{E(y|x, x_k)}} = \exp(\beta_{\mathrm{k}}\delta)$$

Therefore, the expected value of y given x increases by the factor $\exp(\beta_{\mathrm{k}}\delta)$ following a change δ in $\mathrm{x_k}$ keeping other variables constant. When δ has the specific value one, the expected counts increases by the factor $\exp(\beta_{\mathrm{k}}\delta)$ following a unit change in $\mathrm{x_k}$.

The *percentage change* in the expected value of y given x following a δ change in $\mathrm{x_k}$, holding other variables constant [17] is given by:

$$100\frac{\mathrm{E(y|x, x_k} + \delta) - \mathrm{E(y|x, x_k)}}{\mathrm{E(y|x,x_k)}} = 100(\exp(\beta_{\mathrm{k}}\delta) - 1)$$

Another way to interpret count model is to compute with the *predicted probability*:

$$\hat{\Pr}(\mathrm{y} = \mathrm{m}|\mathrm{x}) = \frac{\exp(-\hat{\lambda})\hat{\lambda}^{\mathrm{m}}}{\mathrm{m!}} \text{ where } \hat{\lambda} = \exp(\mathrm{x}'\beta).$$

The *mean predicted probability* for each count m is:

$$\bar{\Pr}(\mathrm{y} = \mathrm{m}) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\exp(-\hat{\lambda}_{\mathrm{i}})\hat{\lambda}_{\mathrm{i}}^{\mathrm{m}}}{\mathrm{m!}}$$

This measure is to compare with the observed proportions of the sample at each count. Large differences between the mean probabilities and the observed proportions suggest that the model is inappropriate. However, small differences do not imply that the model is appropriate because an incorrect model can provide predictions close to observed proportions [17].

One important property of Poisson distribution is that its mean and variance are equal, $\mathrm{Var}(y_i|x_i) = \mathrm{E}(y_i|x_i) = \lambda_i$. In fact, the Poisson distribution is parameterized by a single scalar parameter (λ_i) so that all moments of y_i are a function of λ_i. In practice, the assumption of equidispersion may be violated for two main reasons. First, the frequency of zero counts is greater than the number of expected zeroes generated by the Poisson distribution. Second, the variance of observed counts data may exceed the mean due to unobserved heterogeneity. It is important to control overdispersion because, if it is large, it leads to grossly deflated standard errors and grossly inflated t-statistics in maximum likelihood estimation. For these reasons, many statistical tests are developed in order to detect overdispersion in data.

Cameron and Trivedi [4] set out a test for overdispersion based on a linear regression without the intercept. The test is designed so as to choose one of the following null and alternative hypotheses:

$$\mathrm{H}_0 : \mathrm{Var}(y_i) = \lambda_i$$

$$\mathrm{H}_1 : \mathrm{Var}(y_i) = \lambda_i + \alpha g(\lambda_i)$$

where α is an unknown parameter and $g(\cdot)$ is a definite function, most commonly $g(\lambda_i) = \lambda_i^2$ or $g(\lambda_i) = \lambda_i$. This test can be computed by estimating the Poisson model, constructing the fitted value of $\hat{\lambda}_i = \exp(x_i'\hat{\beta})$ and running the OLS regression without the intercept:

$$\frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i} = \alpha \frac{g(\hat{\lambda}_i)}{\hat{\lambda}_i} + e_i$$

where e_i is an error term.

The significance of the coefficient α in the OLS regression model implies the existence of overdispersion in data. Note that this test can also be used for detecting under-dispersion which is the situation where conditional variance is less than conditional mean.

Another test for overdispersion, introduced by Greene [8], is based on the Lagrange multiplier (LM) statistics. The LM statistic is defined by:

$$\mathrm{LM} = \left[\frac{\sum_{i=1}^{n} \hat{w}_i [(\hat{y}_i - \hat{\lambda}_i)^2 - y_i]}{\sqrt{2 \sum_{i=1}^{n} \hat{w}_i \hat{\lambda}_i^2}} \right]^2$$

where $\hat{w}_i$ define the weight of the alternative distribution. When the alternative is a Negative binomial negative, the weights equals to 1. In this case, the LM is given by:

$$\mathrm{LM} = \frac{(\mathrm{e}'\mathrm{e} - \mathrm{n}\bar{\mathrm{y}})^2}{2\hat{\lambda}'\hat{\lambda}}$$

Note that the limiting distribution of the LM statistic is Chi-Squared with one degree of freedom [8].

3 Zero-inflated Poisson model

As discussed above, the Poisson model is inadequate in the case of excess of zeroes in the sample because of violation of the equidispersion assumption. Lambert [15] has introduced Zero-inflated (ZIP) model as an alternative way to model counts data with excess of Zeroes. The idea of ZIP model is simple: it assumes that outcomes emanate from two processes [15]. One process models zero inflation, by including a proportion $1 - \mathrm{p}$ of extra zero and a proportion $\mathrm{p}\exp(-\lambda_\mathrm{i})$ of zeroes coming from the Poisson distribution; and the second models the nonzero counts using zero-truncated Poisson model. The ZIP model can, then, be formulated as follows:

$$\mathrm{P}(\mathrm{Y_i} = \mathrm{y_i}|\mathrm{x_i}, \mathrm{z_i}) = \begin{cases} \theta_\mathrm{i}(z_\mathrm{i}) + (1 - \theta_\mathrm{i}(z_\mathrm{i}))\mathrm{Pois}(\lambda_\mathrm{i}; 0|\mathrm{x_i}) & \text{if} \quad \mathrm{y_i} = 0 \\ (1 - \theta_\mathrm{i}(\mathrm{z_i}))\mathrm{Pois}(\lambda_\mathrm{i}; \mathrm{y_i}|\mathrm{x_i}) & \text{if} \quad \mathrm{y_i} > 0 \end{cases}$$

With $\mathrm{z_i}$ is a vector of covariates defining the probability θ_i, $\mathrm{Pois}(\lambda_\mathrm{i}; 0|\mathrm{x_i}) = \exp(-\lambda_\mathrm{i})$ and $\mathrm{Pois}(\lambda_\mathrm{i}; \mathrm{y_i}|\mathrm{x_i}) = \frac{e^{-\lambda_\mathrm{i}}\lambda_\mathrm{i}^{\mathrm{y_i}}}{\mathrm{y_i}!}$. The mean and the variance of ZIP are $\mathrm{E}(\mathrm{y_i}|\mathrm{x_i}, \mathrm{z_i}) = (1 - \theta_\mathrm{i})\lambda_\mathrm{i}$ and $\mathrm{Var}(\mathrm{y_i}|\mathrm{x_i}, \mathrm{z_i}) = (1 - \theta_\mathrm{i})(\lambda_\mathrm{i} + \theta_\mathrm{i}\lambda_\mathrm{i}^2)$.

One can observe clearly that the ZIP reduces to the classical Poisson model when θ_i= 0. Otherwise, the ZIP is over dispersed because the variance exceeds the mean. This overdispersion is not due to the heterogeneity of the data which can be handled using the negative binomial model. Instead, it arises from the splitting of the data into the two statistical processes because of the excess of zeroes. According to Lambert [15], we can model $\theta_\mathrm{i}(\mathrm{z_i})$ using a Logit model given by:

$$\theta_\mathrm{i}(\mathrm{z_i}) = \frac{\exp(\mathrm{z_i}'\gamma)}{[1 + \exp(\mathrm{z_i}'\gamma)]}$$

With $\mathrm{z_i}$ is a vector of covariates defining the probability θ_i and γ is a vector of its corresponding parameters. The vector of $\mathrm{z_i}$ can includes elements of $\mathrm{x_i}$ and Logit model can be substituted by Probit specification. The parameter θ_i can be also related to λ_i. In this paper, we assume that $\mathrm{y_1}...\mathrm{y_n}$ are independent and θ_i is not related to λ_i. We can define the likelihood function of $(\mathrm{Y_i})$ as follows:

$$\mathrm{L} = \prod_{\mathrm{y_i}=0} [\theta_\mathrm{i}(\mathrm{z_i}) + (1 - \theta_\mathrm{i}(\mathrm{z_i}))\exp(-\lambda_\mathrm{i})] \prod_{\mathrm{y_i}\neq 0} \left[(1 - \theta_\mathrm{i}(\mathrm{z_i}))\frac{\mathrm{e}^{-\lambda_\mathrm{i}}\lambda_\mathrm{i}^{\mathrm{y_i}}}{\mathrm{y_i}!}\right]$$

The log likelihood function is:

$$LL = \sum_{y_i=0} \log[\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta))] + \sum_{y_i\neq 0} [y_i x'\beta - \exp(x'\beta) - \log(y_i!)] - \sum_{i=1}^{n} \log([1+\exp(z_i'\gamma)])$$

Let D_i be an indicative function defined as follows:

$$D_i = \begin{cases} 1 & \text{if} \quad y_i = 0 \\ 0 & \text{Otherwise} \end{cases}$$

The log likelihood function becomes:

$$LL = \sum_{i=1}^{n} D_i \log[\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta))] + \sum_{i=1}^{n} (1-D_i)[y_i x'\beta - \exp(x'\beta) - \log(y_i!)] - n\log([1+\exp(z'\gamma)])$$

The log likelihood function (LL) can be maximized using EM algorithm [10], [15]. The EM algorithm estimates the expectations of the ith missing observation at each iteration and uses these expectations to estimate the parameters, and iterates until convergence. The missing data of this problem are defined as indicator variables $\delta = (\delta_1 \ldots \delta_n)'$ where $\delta_i = 1$ when y_i, is from the zero state and $\delta_i = 0$ when y_i is from the Poisson state. [Lambert [15]] defined the complete-data as follows:

$$LL_c(\beta, \gamma, y, \delta) = LL_c(\gamma, y, \delta) + LL_c(\beta, y, \delta) + \sum_{i=1}^{n} (1-\delta_i)\log(y_i!)$$

where

$$LL_c(\gamma, y, \delta) = f(\delta_i|\gamma) = \sum_{i=1}^{n} [\delta_i z_i \gamma - \log(1+\exp(z_i'\gamma))]$$

And

$$LL_c(\beta, y, \delta) = f(y_i|\delta_i, \beta) = \sum_{i=1}^{n} (1-\delta_i)[y_i x_i\prime\beta - \exp(x_i'\beta)]$$

This log likelihood can be easily maximized because $LL_c(\gamma, y, \delta)$ and $LL_c(\beta, y, \delta)$ can be maximized separately. The EM proceeds iteratively via three steps [11], [15]: E step, M step for β and M step for γ. For the (k + 1) iteration, the three steps are defined as follows:

E step: Estimate δ_i by its conditional expectation $\delta_i^{(k)}$given by the current estimates $\beta^{(k)}$ and $\gamma^{(k)}$

$$\delta_i^{(k)} = \begin{cases} [1+\exp(z_i\gamma^{(k)} - \exp(x_i^{(k)}))]^{-1} & \text{if} \quad y_i \succ 0 \\ 0 & \text{if} \quad y_i = 0 \end{cases}$$

M step for β: it consists on finding $\beta^{(k+1)}$ by maximizing $LL_c(\beta, y, \delta^{(k)})$. This can be accomplished by fitting a weighted log linear Poisson regression of y on the covariate matrix x with weights $1 - \delta^{(k)}$ [19].

M step for γ: it consists on finding $\gamma^{(k+1)}$ by maximizing $LL_c(\gamma, y, \delta^{(k)})$ as function of γ where

$$LL_c(\gamma, y, \delta^{(k)}) = \sum_{y_i=0} \delta_i^{(k)} z_i\gamma - \sum_{y_i=0} \delta_i^{(k)} \log(1 + \exp(z_i\gamma)) - \sum_{i=1}^{n} (1 - \delta_i^{(k)}) \log(1 + \exp(z_i\gamma))$$

Note that $\gamma^{(k+1)}$ can be found by fitting a weighted logistic regression of the response variable y on the covariate matrix z with weights $(1 - \delta^{(k)})$, [15].

Lambert [15] has developed the observed information matrix of the ZIP model. The expected information matrix is:

$$i_{\gamma,\beta} = \begin{bmatrix} z' & 0 \\ 0 & x' \end{bmatrix} \begin{bmatrix} d_{\gamma,\gamma} & d_{\gamma,\beta} \\ d_{\gamma,\beta} & d_{\beta,\beta} \end{bmatrix} \begin{bmatrix} z & 0 \\ 0 & x \end{bmatrix}$$

where $d_{\gamma,\gamma}$ is diagonal with elements $p(r - p)$, $d_{\gamma,\beta}$ is diagonal with elements $-\lambda p(1 - r)$, and $d_{\beta,\beta}$ is diagonal with elements $\lambda(1 - p) - \lambda^2 p(1 - p)$. If the quantity $n^{-1} i_{\gamma,\beta}$ has a positive definite limit [18], then the quantity $n^{1/2} \begin{bmatrix} \hat{\gamma} - \gamma \\ \hat{\beta} - \beta \end{bmatrix}$ is asymptotically normal distributed with mean 0 and variance $i_{\gamma,\beta}^{-1}$, [15].

The hypothesis tests can be carried using the Wald test. The Wald statistic is given by:

$$W = \hat{\beta}' i_{\gamma,\beta} \hat{\beta}$$

Under the null hypothesis, the statistic W is asymptotically Chi-squared distributed with one degree of freedom.

The difference of the likelihood ratio can be used also. If $(\hat{\gamma}_0, \hat{\beta}_0)$ maximizes the log likelihood of the ZIP model under a null hypothesis $\mathbf{H_0}$ of dimension q_0 and $(\hat{\gamma}, \hat{\beta})$ maximizes the log likelihood of the ZIP model under a nested alternative hypothesis $\mathbf{H_1}$ of dimension $q > q_0$, then the quantity $2[L(\hat{\gamma}, \hat{\beta}) - L(\hat{\gamma}_0, \hat{\beta}_0)]$ is asymptotically Chi-squared distributed with $q - q_0$ degrees of freedom, [15].

The interpretation of the parameters in the ZIP model λ is the same as in the Poisson regression, [17]. Whereas, the parameters γ can be interpreted through the odds which is the ratio of the probability that something is true divided by the probability that it is not true. The mathematical formulation is $\frac{\theta(z)}{1-\theta(Z)}$. The odds ratio is the ratio of two odds for different values of z_j, $\frac{\text{odds}(z_j+\delta)}{\text{odds}(z_j)} = \exp(\gamma_j\delta)$. If δ is enough small, we can easily prove that $\frac{\text{odds}(z_j+\delta)-\text{odds}(z_j)}{\text{odds}(Z_j)} \approx \gamma_j\delta$. Thus, the parameters γ_j can be interpreted as the relative change in the odds due to the small change δ in z_j.

4 Goodness test

The deviance information criterion is the statistic used to assess the goodness of fit of the Poisson regression. It is calculated as the difference in log likelihoods between the two models then multiplied by -2. For large samples the distribution of the deviance is approximately a Chi-squared with n-p degrees of freedom, where n is the number of observations and p the number of parameters [1]. A significant p value indicates that the deviance is greater than what would be expected under a null hypothesis of model equivalence; hence, the more complex model with an additional parameter or parameters is considered a significant improvement over the nested model [20].

5 Test for the comparison of the models

After fitting Poisson regression and ZIP regression to the data, we can ask the question: what is the best model for our analysis? As a response, many tests are developed. In this article we will focus on the vuong's test and the probability integral transforms.

5.1 Vuong's test

Vuong [24] has introduced a test which is a well suited method to compare ZIP regression to other non nested model for counts data.

Let $P_N(y_i|x_i)$ be the predicted probability of an observed count for case i from the modelN, we define m_i as follows:

$$m_i = \text{Log}\left(\frac{P_1(y_i|x_i)}{P_2(y_i|x_i)}\right)$$

Hence, the Vuong's test for the hypothesis $E(m_i) = 0$ is given by:

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}}$$

Under the null hypothesis, the Vuong's statistic is asymptotically normally distributed. At 5 % significance level, the first model is preferred if $V > 1.96$; If $V < -1.96$, then the second one is preferred and the two models are equivalent when $|V| < 1.96$.

5.2 Probability integral transforms

The probability integral transform (PIT) is the value attained by the predictive cumulative distribution function, denoted F, evaluated at the value that materializes. It has been used by [6] for the assessment of the model fit and predictive capabilities

of the model. If an observation is from a continuous predictive distribution, then the PIT has a standard uniform distribution. To check the assessment of the model, we plot the empirical CDF of a set of PIT values and we compare the graph to the identity function. The histogram of the PIT values and is more informative. If it is flat, the perfect prediction is shown. U-shaped and bump shaped histograms indicate underdispersed and overdispersed predictions, respectively.

In the case of count events, [Smith [23]] considered a randomized PIT value given by:

$$\begin{cases} \upsilon = P(y-1) + u[P(y) - P(y-1)] & \text{if} \quad y \geq 1 \,(1) \\ \upsilon = uP(0) & \text{if} \quad y = 0 \,(2) \end{cases}$$

where u is an independent observation of a uniform (0, 1). Under the perfect prediction, the randomized PIT is from uniform distribution. The histogram can be used to assess the predictive capabilities as is the case of the continuous predictive distribution.

Czado et al. [5] proposed a non-randomized yet uniform version of the PIT histogram. They replace the randomized PIT value in (1) and (2) by its conditional cumulative distribution function given the observed count y:

$$F(u) = \begin{cases} 0 & \text{if} \quad u \leq P(y-1) \\ [u - P(y-1)]/[P(y) - P(y-1)] & \text{if} \quad P(y-1) \leq u \leq P(y) \\ 1 & \text{if} \quad u > P(y) \end{cases}$$

If $x \geq 1$

$$F(u) = \begin{cases} u/P(0) & \text{if} \quad u \leq P(0) \\ 1 & \text{if} \quad u \geq P(0) \end{cases}$$

If $x = 0$. For calibrating the model, Czado et al. [5] used the aggregation over the predictions,

$$\bar{F}(u) = \frac{1}{n} \sum_{i=1}^{n} F^{(i)}(u) \text{ with } \quad 0 \leq u \leq 1$$

where $F^{(i)}$ is based on the predictive distribution $P^{(i)}$ and the observed count $x^{(i)}$. They compared the mean PIT to the identity function. This comparison can be performed by plotting a non-randomized PIT histogram. Under the perfect prediction, the histogram is flat. Underdispersion and overdispersion are detected in case of U-shaped and bump shaped histograms, respectively.

6 Application

The pure premium method is widely used by actuaries for pricing insurance products. The pure premium is the average premium that must be collected to pay for losses only. It is calculated by dividing the loss by the number of exposures. In other way, the pure premium (PP) is calculated as a multiplication of the frequency

and the severity of claims. Therefore, the formula of PP is given by: PP = Frequency × Severity

To define the PP, an actuary needs to model separately the frequency and the severity. In the past, the one way analysis was commonly used by actuaries. A one way analysis does not take into account the effect of explanatory variables neither on frequency nor on severity. Two major problems had been encountered. First, one way analysis can be distorted by correlations between rating factors. Second, it ignores the interdependence or interactions between factors [2].

Recently, the one way analysis has been replaced by generalized linear models (GLMs). This class of models relates the response variable to the factors. They take into account correlations and interactions between factors.

GLMs consist of a wide range of models. The use of each model depends essentially on the nature of the response variable. If this one is discrete, Poisson regression, negative binomial model and other discrete models can be used. If it is continuous, Gamma distribution and other continuous models can be fitted to the data. In insurance pricing, for example, the Gamma distribution and Poisson regression are widely used for modeling severity and frequency of a given risk respectively.

In this paper, we will fit Poisson regression and ZIP regression to a private health insurance counts data. The database contains information about claims and policyholders. The response variable is the number of claims per year received from the insured. The covariate matrix contains socioeconomic variables describing the insured people (Table 1).

Table 1 Variable description for the analyzed health insurance data set

Variable	Type	Description
Number_claims	Discrete	Number of claims received from the insured per year
Industrial_city	Binary	Takes the value 1 if insured lives in industrial city (Casablanca, Mohammedia, Kenitra or Tanger), 0 otherwise.
Gender_male	Binary	Takes the value 1 if the insured is male, 0 if the inured is female
Industrial_activity	Binary	Takes the value 1 if insured works in industrial firm, 0 otherwise
Services_activity	Binary	Takes the value 1 if the insured works in the services company (e.g insurance and bank), 0 otherwise
Age_30	Binary	Takes the value 1 if the insured have an age less than 30 years, 0 otherwise
Age_40	Binary	Takes the value 1 if the age of insured varies between 30 and 40 years, 0 otherwise
Age_60	Binary	Takes the value 1 if the age of insured varies between 40 and 60 years, 0 otherwise
Status_married	Binary	Takes the value 1 if insured is married, 0 otherwise
Status_single	Binary	Takes the value 1 if insured is single, 0 otherwise
Size_family	Discrete	Indicates the size of the family of the insured person
Exposure	Continuous	Indicates coverage period of the insured in the year. It varies between 0 and 1

6.1 Description of data

The number of observations in our data base is 84 331. Around 66 % of the insured people are male. In addition, 52 % are married, 45 % are single, and only 3 % are divorced or widowed. Moreover, 34 % of the insured people are less than 30 years, 66 % are between 31 years and 60 years and only 3 % exceeds 60 years. Geographically, we remark that 71 % of the insured people live and work in the Moroccan economic capital: Casablanca.

70 % of claims are notified being from male. The married insured persons have filed about 81 % of total claims in the database. The single insured have filed only 16 % of total claims even if their percentage in total insured persons is 45 %. 88 % of claims are received from insured that live in the industrial city. Persons aged between 40 and 60 years filed about 50 % of total claims, followed by the ones whose are between 30 and 40 years who filed around 30 % of total claims. The persons whose age is less than 30 years filed only 12 % of total claims even if they represent around 34 % of the total.

6.2 Results

The mean of the number of claims declared by one insured and in 1 year is about 5.16 and the variance is 102.29. It is clear that the mean is very small compared to the variance. This indicates the overdispersion of the data. Generally, a Poisson regression model is not an appropriate model to fit to the data in such a case.

Our objective is to fit Poisson regression and Zero-inflated Poisson regression to our data. Note that in ZIP regression model, we use the same covariate matrix for estimating λ and θ. The models specifications are indicated in Table 2.

First we fit the Poisson regression to the number of claims. Table 3 indicates the estimates of parameter. We remark that all explanatory variables are significant

Table 2 Specification of poisson regression model and ZIP regression model

Poisson regression model	ZIP regression model
$\text{Log}(\textit{Number\_Claims}) = \log(\text{Exposure}) + \text{Constant} + \beta_1 \times \text{Size\_Family} + \beta_2 \times \text{Industrial\_City} + \beta_3 \times \text{Gender\_Male} + \beta_4 \times \text{Industrial\_Activity} + \beta_5 \times \text{Services\_Activity} + \beta_6 \times \text{Status\_Married} + \beta_7 \times \text{Status\_Single} + \beta_8 \times \text{Age\_30} + \beta_9 \times \text{Age\_40} + \beta_{10} \times \text{Age\_60}$	*Poisson with log link* $\ln(\lambda) = \log(\text{Exposure}) + \text{Constant} + \beta_1 \times \text{Size\_Family} + \beta_2 \times \text{Industrial\_City} + \beta_3 \times \text{Gender\_Male} + \beta_4 \times \text{Industrial\_Activity} + \beta_5 \times \text{Services\_Activity} + \beta_6 \times \text{Status\_Married} + \beta_7 \times \text{Status\_Single} + \beta_8 \times \text{Age\_30} + \beta_9 \times \text{Age\_40} + \beta_{10} \times \text{Age\_60}$ *Binomial with logit link* $\log it(\theta) = \text{Constant} + \gamma_1 \times \text{Size\_Family} + \gamma_2 \times \text{Industrial\_City} + \gamma_3 \times \text{Gender\_Male} + \gamma_4 \times \text{Industrial\_Activity} + \gamma_5 \times \text{Services\_Activity} + \gamma_6 \times \text{Status\_Married} + \gamma_7 \times \text{Status\_Single} + \gamma_8 \times \text{Age\_30} + \gamma_9 \times \text{Age\_40} + \gamma_{10} \times \text{Age\_60}$

Table 3 Estimation of poisson regression

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.482479	0.011945	124.11	<2e-16
Log(Exposure)	1.545407	0.018485	83.60	<2e-16
Size_Family	0.102971	0.001255	82.05	<2e-16
Gender_Male	−0.233582	0.003662	−63.79	<2e-16
Status_Married	0.240690	0.009817	24.52	<2e-16
Status_Single	−0.564231	0.010451	−53.99	<2e-16
Industrial_City	0.349477	0.005225	66.89	<2e-16
Services_Activity	0.741885	0.006427	115.42	<2e-16
Industry_Activity	0.541030	0.003387	159.72	<2e-16
Age_30	−1.038485	0.008303	−125.07	<2e-16
Age_40	−0.557606	0.006800	−82.00	<2e-16
Age_60	−0.349429	0.006559	−53.27	<2e-16

because the associated p value to each factor is less than 5 %. The goodness test shows that the residual deviance for the model without covariates is very high (around 978967); and it fall down to 862317 when add the factor size of family. The minimum is obtained when all variables are added (see Table 4). The size_family, Industrial_city, Services_activity, status_married and Industry_activity have a positive sign. A positive change in these factors induces then an increase in the number of claims. The percentage change of the factor status_married is 27 %; this means that the number of claims filed by the married persons is 27 % more than the others. Whereas the percentages change of the factor status_single is around −43 %. Hence, single persons are more profitable for the insurance and should be of a prime interest of the underwriting strategies. For industrial cities, the percentage change is around 42 %. Thus, persons in great cities are more exposed to sickness that the small cities where the industrial activities are less preponderant. The

Table 4 Goodness-of-fit test of poisson regression: the variables are added one by one

	Df	Deviance	Resid.Df	Resid.Dev
NULL			83 678	978 967
Size_Family	1	116 650	83 677	862 317
Industrial_City	1	9 150	83 676	853 167
Male	1	4 225	83 675	848 942
Industry_Activity	1	23 526	83 674	825 417
Services_Activity	1	5 461	83 673	819 955
Married	1	32 041	83 672	787 914
Single	1	6 744	83 671	781 170
Age_30	1	12 402	83 670	768 769
Age_40	1	3 618	83 669	765 151
Age_60	1	2 427	83 668	762 724

percentages change of Industry_activity and Services_activity are 72 and 110 %, respectively. Persons in industrial and services activities are very exposed to sickness; they should be a bad target for the underwriters.

Second, we compute the test introduced by [4] for detecting overdispersion in the data (Table 5). The t-statistc of the parameter α of this test is $z = 83.510$ with p value $< 2.2e\text{-}16$. This indicates that the dependent count variable is overdispersed. A second measure allowing the dispersion test is the Probability integral transforms (PIT). The histogram of the PIT is bump shaped (Fig. 1). This indicates also an overdispersion of the response variable. Since the histogram (Fig. 2) is highly peaked at zero, we can state that the overdispersion is due to excess of zeroes.

Table 5 Test of overdispersion in poisson regression

Test	Statistic	Decisions
Cameron and Trivedi [4]	t-statistic = 83.5*10*	Parameter α of regression test is not null, then data are overdispersed
LM Statistic [8]	Obs.Var/ Theor.Var = 19.81 t-statistic = 1 670 932	p-value = 0, data are overdispersed

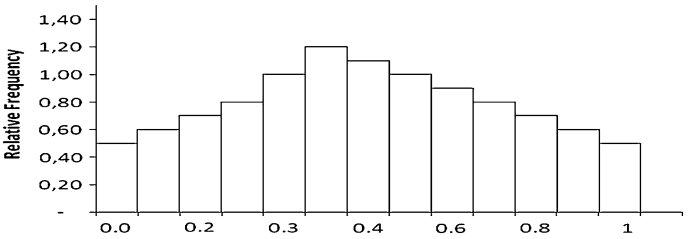

Fig. 1 Histogram of PIT of the Poisson regression

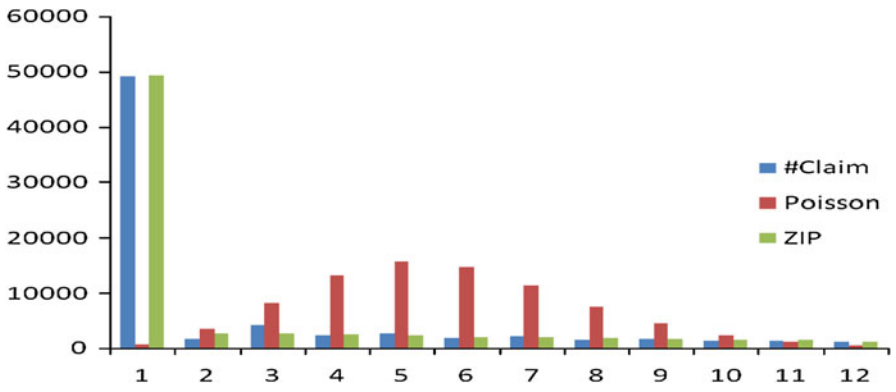

Fig. 2 Histogram of Number of claims

As discussed above, the overdispersion can be due to two facts: heterogeneity of data or excess of zero. The histogram of the number of claims (Fig. 2) is highly peaked at zero value because of the preponderance of zero in data. However, large numbers of claims are less frequently observed. This leads to right skewed distribution. An alternative way for modeling this type of data is the Zero-inflated Poisson regression which takes into account the excess of zeroes. Table 6 provides estimate parameters of Zero-inflated Poisson regression model. For nonzero outcomes, we can see that all factors are significant for λ since their p values are less than 5 %. However, the factor Industrial_city does not impact the parameter θ. The deviance test, measured as twice the difference between likelihood of the model without covariates and that of the full model ($2 \times (-233600 - (-265300)) = 63\ 400 \sim \chi_{18}$), proves that the full model is statistically significant. The sign of the parameters in the positive part of the ZIP model is the same in the Poisson model. However the percentages changes in the factors are largely changed; and are more realistic than that of the Poisson model. The percentages changes of the factors status_married, status_single, Industry_activity and services_activity are 22, −30, 12 and 31 %, respectively. The parameters of zero outcomes model can be interpreted in odds ratio. The change in odds ratio for the factor Size_family is −17 %; this means that the probability of notification of a claim can be increased by 17 % according to an increase by a unit in the size of family. One of important change in odds ratio is that of the factor services_activity. It is around −61 %. The probability of notification of a claim by persons working in services activity is 61 % greater than that of people working in the other sectors.

The histogram of the probability integral transform is flat (Fig. 3); this means that the zero inflated Poisson model gives a good prediction of the number of claims. This result can be shown also in the Fig. 2. In fact, the histogram (Fig. 2) shows that Zero-inflated Poisson distribution has the ability to reproduce the number of zeroes in the population better than standard Poisson distribution. A third measure used in this article is the vuong's test (Table 7). The computed statistic of this test is $V = 144.7145$. Under the null hypothesis this statistic is asymptotically normally distributed. If we consider the significance level of 5 %, we conclude easily that Zero-inflated Poisson regression fits the number of claims better than the standard Poisson regression.

7 Concluding remarks

In this paper, we have introduced two models usually fitted to counts data: Poisson regression and Zero-inflated Poisson regression. Maximum likelihood techniques are used to estimate the parameters of both models. Since the Hessian matrix associated to the Poisson regression is negative, the Newton–Raphson iterative algorithm converges rapidly and provides unique parameters estimates. The EM algorithm, used to maximize the likelihood of Zero-inflated Poisson regression, proceeds iteratively via three steps: E step, M step for truncated Poisson model and M step for binomial model. The EM algorithm estimates the expectations of missing data and iterates until convergence. Given a significance level (e.g. 5 %) Wald test,

Table 6 Estimation of Zero inflation poisson regression

	P(Y > 0)				*P(Y = 0)*			
	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	2.202180	0.011851	185.81	<2e-16	−0.086647	0.067913	−1.276	0.2020
Log(Exposure)	1.193137	0.022669	52.63	<2e-16				
Size_Family	0.038125	0.001365	27.93	<2e-16	−0.196234	0.008252	−23.781	<2e-16
Male	−0.071749	0.003771	−19.03	<2e-16	0.442767	0.017460	25.359	<2e-16
Married	0.203630	0.009921	20.53	<2e-16	−0.104544	0.050958	−2.052	0.0402
Single	−0.352513	0.010397	−33.90	<2e-16	0.240694	0.050410	4.775	1.8e-06
Industrial_City	0.346116	0.005228	66.21	<2e-16	0.014876	0.022927	0.649	0.5165
Services_Activity	0.271821	0.006466	42.04	<2e-16	−0.940671	0.030732	−30.609	<2e-16
Industry_Activity	0.110169	0.003389	32.50	<2e-16	−1.003600	0.017638	−56.900	<2e-16
Age_30	−0.456491	0.008203	−55.65	<2e-16	1.291134	0.049957	25.845	<2e-16
Age_40	−0.302112	0.006783	−44.54	<2e-16	0.763794	0.047903	15.944	<2e-16
Age_60	−0.186453	0.006676	−27.93	<2e-16	0.571890	0.048110	11.887	<2e-16

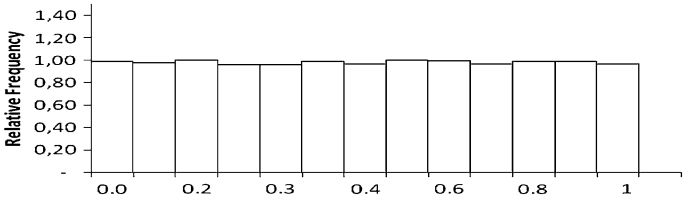

Fig. 3 The histogram of the PIT of the ZIP model

Table 7 Vuong's test

V statistic	Model 1: ZIP regression
Model 2: Poisson regression	**V** = 144.7145 Test-statistic is asymptotically distributed N (0, 1) under the null hypothesis. In this case: model 1 > model 2

Lagrange Multiplier test or Likelihood Ratio test are used to measure the significance of factors incorporated in each model. The equidispersion assumption of Poisson regression was tested using LM test introduced by [8] and regression test set by [4].

These tests have proved the overdispersion of the number of claims in a Moroccan private health insurance scheme. According to histogram, which is highly peaked at zero, we state that this overdispersion is due to the preponderance of zeroes in the population. In such case, we have shown that standard Poisson model is unable to reproduce the number of zeroes in the data and therefore, underestimates the dispersion of the population. Alternatively, we have fitted Zero-Inflated Poisson model. We have shown that this model simulates well the data and the number of zeroes reproduced by it is very close to the number of zeroes in the population. Finally, we have computed Vuong's test and the probability integral transforms for selecting the best model in the case of excess of zeroes. We can conclude that Zero-inflated Poisson regression fits excess of zeroes counts data better than standard Poisson regression.

References

1. Agresti A (1996) An introduction to categorical data analysis. Wiley, NewYork
2. Anderson D, Feldblum S, Modlin C, Schrimacher D, Schirmacher E, Thandi N (2004) Practitioner's guide to generalized linear models: a foundation for theory, interpretation and application. Watson Wyatt
3. Cameron AC, Trivedi PK (1986) Econometric models based on counts data: comparisons and applications of some estimators and tests. J Appl Econom 1(1):29–53
4. Cameron AC, Trivedi PK (1990) Regression-based tests for overdispersion in the Poisson model. J Econom 46(3):347–364
5. Czado C, Gneiting T, Held L (2009) Predictive model assessment for count data. Biometrics 65:1254–1261

6. Dawid AP (1984) Statistical Theorie: the prequential approach. J Royal Stat Soc Ser A Gen 147:278–292
7. Dean C, Lawless JF (1989) Tests for detecting overdispersion in poisson regression models. J Am Stat Assoc 84(406):467–472
8. Greene W (2002) Econometric analysis. Prentice Hall, USA
9. Gurmu S (1991) Tests for detecting overdispersion in the positive poisson regression model. J Bus Econ Stat Am Stat Assoc 9(2):215–222
10. Hall DB (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56:1030–1039
11. Hall DB, Shen J (2010) Robust estimation for zero-inflated Poisson regression. Scand J Stat 37:237–252
12. Hausman JA, Hall DB, Griliches Z (1984) Econometric models for counts data with an application to the patents-R&D relationship. Econometrica 52:909–938
13. King G (1989) A seemingly unrelated poisson regression model. Sociol Methods Res 17:235–255
14. King G (1989) Variance specification in event count models: from restrictive assumptions to a generalized estimator. Am J Political Sci 33:762–784
15. Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34:1–14
16. Lee L (1986) Specification tests for poisson regression models. Int Econ Rev 27:689–706
17. Long JS (1997) Regression models for categorical and limited dependent variables. Sage Publications, Thousand Oaks
18. McCullagh P (1983) Quasi-likelihood functions. Ann Stat 11:59–67
19. McCullagh P, Nelder J (1989) Generalized linear models. Chapman & Hall, London
20. Miller JM (2007) Comparing Poisson, Hurdle and ZIP model fit under varying degrees of skew and zero-inflation. PhD Thesis, University of Florida, Gainesville, Florida, USA
21. Mouatassim Y, Ezzahid E, Belasri Y (2012) Operational Value-at-Risk in Case of Zero-inflated Frequency. Int J Econ Finance 4(6):70–77
22. Nixon DC (1991) Event count models for supreme court dissents. Political Methodol 4:11–14
23. Smith JQ (1985) Diagnostic checks of non-standard time series models. J Forecast 4:283–291
24. Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 307–334
25. Zorn CJW (1996) Evaluating zero-inflated and hurdle Poisson specifications. Presented at Midwest Political Science Association, USA