

Estimation Techniques for Regression Model with Zero-inflated Poisson Data

Shakhawat Hossain¹ & Hatem A. Howlader¹

¹ Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, Canada

Correspondence: Shakhawat Hossain, Associate Professor, Department of Mathematics and Statistics, University of Winnipeg, 515 Portage Avenue, Winnipeg, MB, Canada R3B 2E9. Phone: 1-204-786-9492. E-mail: sh.hossain@uwinnipeg.ca

Received: August 10, 2015 Accepted: August 24, 2015 Online Published: September 23, 2015

doi:10.5539/ijsp.v4n4p64 URL: <http://dx.doi.org/10.5539/ijsp.v4n4p64>

Abstract

Researchers in many fields including biomedical often make statistical inferences involving the analysis of count data that exhibit a substantially large proportion of zeros. Subjects in such research are broadly categorized into low-risk group that produces only zero counts and high-risk group leading to counts that can be modeled by a standard Poisson regression model. The aim of this study is to estimate the model parameters in presence of covariates, some of which may not have significant effects on the magnitude of the counts in presence of a large proportion of zeros. The estimation procedures we propose for the study are the pretest, shrinkage, and penalty when some of the covariates may be subject to certain restrictions. Properties of the pretest and shrinkage estimators are discussed in terms of the asymptotic distributional biases and risks. We show that if the dimension of parameters exceeds two, the risk of the shrinkage estimator is strictly less than that of the maximum likelihood estimator, and the risk of the pretest estimator depends on the validity of the restrictions on parameters. A Monte Carlo simulation study shows that the mean squared errors (MSE) of shrinkage estimator are comparable to the MSE of the penalty estimators and in particular it performs better than the penalty estimators when the dimension of the restricted parameter space is large. For illustrative purposes, the methods are applied to a real life data set

Keywords: adaptive LASSO, asymptotic bias and risk, LASSO, likelihood ratio test, shrinkage and pretest estimators, zero-inflated data.

1. Introduction

Data with large number of zeros are often encountered in many studies including medical and public health. Failure to account for the extra zeros may result in biased parameter estimation and misleading inference. In recent years the regression method for modelling such data has become very popular. It is based on the assumption that the response is generated by a mixture of a degenerate distribution at zero and a standard Poisson distribution. For example, the National Medical Expenditure Survey data contain an excessive number of zeros in the hospital admission, and the probability mass at point zero exceeds that allowed under a standard parametric family of distributions. We will analyze this survey data in Section 5 where the number of hospital admission ranges from 0 to 8, and 80.4% of the respondents have no hospital admission, which presents possible zero-inflation. The zero-inflated Poisson regression (ZIPR) model can be used to analyze this data which assumes that the population of respondents can be divided into two subpopulations. The first of which generates counts (number of hospital admissions) that follow a Poisson distribution with the usual parameter, λ and the second subpopulation generates only the zeros (no hospital admission) with probability, p that follows binomial distribution with logit model.

The ZIPR models were first introduced by Mullahy (1986) in the econometric literature. Their use has become very broad since the publication of the pioneering work by Lambert (1992) in which manufacturing defects were considered. Ridout et al. (1998) cited examples of data with too many zeros from various disciplines including medicine, agriculture, and the use of recreational facilities. Hall (2000) considered ZIPR and zero-inflated binomial models with random intercept, and provided an EM algorithm for model estimation. Hinde & Demetrio (1998) reviewed the literature of ZIPR model and cited examples from agriculture, econometrics, manufacturing, road safety, species abundance, and medical sciences. Jansakul & Hinde (2002) extended the van den Broek (1995) score test to the more general situation where the zero probability is allowed to depend on covariates. Feng & Zhu (2011) studied a semiparametric ZIP mixture model, and a Monte Carlo EM algorithm was provided for model estimation. For other aspects of ZIPR models, see among others, Dietz & Bohning (2000) and Ridout et al. (2001).

The motivation of the present paper comes from the need to be able to identify the significant covariates of the ZIPR model in presence of a higher proportion of zeros in the response variable. With this motivation, we consider an unrestricted model that includes all covariates and possible extraneous variables and a restricted model that includes some covariates of concern while leaving out extraneous variables. One way to deal with this framework is to use the pretest procedures that test whether the coefficients of the extraneous variables are zero and then estimate the parameters in the model that include coefficients that are rejected by the test. Another approach is to use Stein type shrinkage estimators where the estimated regression coefficient vector is shrunk in the direction of linear restriction among the parameters. Sapra (2003) considered the pretest method to estimate the parameters of the Poisson regression model. For the properties of pretest and shrinkage estimation strategies for parametric and semiparametric linear models, see Nkurunziza (2013), Hossain et al. (2012), and Liang & Song (2009), among others.

This paper also studies the penalty estimators such as, least absolute shrinkage and selection operator (LASSO) and adaptive LASSO and compares their performances in terms of biases and mean squared errors with the shrinkage and pretest estimators through a simulation study. The LASSO was developed by Tibshirani (1996) for simultaneous variable selection and parameter estimation. Unlike LASSO, the adaptive LASSO proposed in Zou (2006) permits different weights for different parameters. The adaptive LASSO has been shown to have the oracle property, that is, consistency in variable selection and asymptotic normality. Hossain & Ahmed (2012) considered the shrinkage and penalty methods for estimating the Poisson regression model when some of the covariates may be inactive for the response. Zeng et al. (2014) proposed a variable selection approach for ZIPR models via adaptive LASSO.

The rest of this paper is organized as follows. The model and suggested estimators are introduced in Section 2. The asymptotic properties of the proposed estimators and their asymptotic distributional biases and risks are presented in Section 3. The results of a simulation study are reported in Section 4. Application to real life data and a comparison of our methods are described in Section 5. Finally, concluding remarks are given in Section 6.

2. Models and the Proposed Estimators

2.1 Models and Estimation Method

Suppose that the counts, Y_i , $i = 1, 2, \dots, n$ are generated independently according to a zero-inflated Poisson distribution. The zeros are assumed to arise in two distinct underlying processes. The first process occurs with probability p_i and produces only zeros, while the second process occurs with probability $1 - p_i$ and leads to a standard Poisson model with parameter λ_i and hence a chance of further zeros. In general, the inevitable zeros from the first process are called structural zeros and those from the Poisson process are called sampling zeros, see for example, Ridout et al. (1998) and Jansakul & Hinde (2002). These two processes give rise to a simple two-component mixture distribution with probability mass function

$$Pr(Y = y_i) = \begin{cases} p_i + (1 - p_i)\exp(-\lambda_i), & y_i = 0 \\ (1 - p_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, & y_i \geq 1, \quad 0 \leq p_i \leq 1 \end{cases} \quad (1)$$

and $E(Y_i) = (1 - p_i)\lambda_i$, $var(Y_i) = \lambda_i + \left(\frac{p_i}{1 - p_i}\right)\lambda_i^2$. The ZIPR model reduces to a Poisson regression model when $p_i = 0$, and exhibits overdispersion when $p_i > 0$. Lambert (1992) incorporated covariates using a log link for λ_i and logit link for p_i .

$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{and} \quad \log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{z}_i^\top \boldsymbol{\gamma}, \quad (2)$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^\top$ and $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iq})^\top$ are the covariate vectors, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the $p \times 1$ and $q \times 1$ vectors of unknown regression parameters. Let us assume $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$. The covariates that affect the Poisson parameter of the first process may or may not be the same that affect the probability of the second process. When the covariates are the same and λ_i and p_i are not functionally related, then $\mathbf{x}_i = \mathbf{z}_i$ and in that case the ZIPR involves twice as many parameters as the Poisson regression. In model (2), we are interested in testing the following hypothesis:

$$H_0 : \mathbf{H}\boldsymbol{\theta} = \mathbf{h} \quad \text{versus} \quad H_A : \mathbf{H}\boldsymbol{\theta} \neq \mathbf{h},$$

where \mathbf{H} is an $r \times (p + q)$ matrix of full rank and \mathbf{h} is a $r \times 1$ vector of constant terms.

For a random sample, $\mathbf{y} = (y_1, y_2, \dots, y_n)$, the log-likelihood function is given by

$$\begin{aligned} \ell = \ell(\boldsymbol{\lambda}, \boldsymbol{\beta}; \mathbf{y}) &= \sum_{i=1}^n \left\{ I_{(y_i=0)} \ln[p_i + (1-p_i)\exp(-\lambda_i)] + I_{(y_i>0)} [\ln(1-p_i) - \lambda_i + y_i \ln \lambda_i - \ln(y_i!)] \right\} \\ &= \sum_{i=1}^n \left\{ I_{(y_i=0)} \ln[\exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) + \exp(-\exp(\mathbf{x}_i^\top \boldsymbol{\beta}))] \right. \\ &\quad \left. + I_{(y_i>0)} [y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) - \ln(y_i!)] \right\} - \sum_{i=1}^n \ln(1 + \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})). \end{aligned} \quad (3)$$

where $I_{(\cdot)}$ is an indicator function, which is equal to 1 if the event is true and 0 otherwise.

The log-likelihood function (3) of the ZIPR model is quite complicated, especially as the first term involves both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Also, the responses are from a mixture distribution that includes both sets of the parameters p_i and λ_i . The computation thus becomes quite challenging in terms of the variance-covariance matrix and accuracy when using the Newton-Raphson algorithm. And it is due to the additional number of parameters to be estimated for the proposed model from the complicated nature of the likelihood function (3). To avoid this complication, we use the EM algorithm to maximize the log-likelihood function, see, Hall (2000) and Lambert (1992).

The EM algorithm is based on a latent variable U_i , where we observe U_i as 1, when Y_i is from the perfect, zero state (or first process) and U_i as 0, when Y_i is from the Poisson state (or second process). To formulate the log-likelihood for the complete data, we use the conditional probability:

$$\begin{aligned} &Pr(Y_i = y_i, U_i = u_i | \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\ &= Pr(Y_i = y_i | U_i = u_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) \times Pr(U_i = u_i | \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\ &= \left(\frac{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}{1 + \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} \right)^{u_i} \left(\frac{\exp(y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}))}{y_i! (1 + \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}))} \right)^{1-u_i} \end{aligned}$$

Thus, the complete log-likelihood based on (\mathbf{Y}, \mathbf{U}) is

$$\begin{aligned} L_c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{U}, \mathbf{x}_i, \mathbf{z}_i) &= \ln \left[\prod_{i=1}^n Pr(Y_i = y_i, U_i = u_i | \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) \right] \\ &= \sum_{i=1}^n \{ u_i \mathbf{z}_i^\top \boldsymbol{\gamma} - u_i \ln(1 + \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})) \} \\ &\quad + \sum_{i=1}^n (1 - u_i) \left[(y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) - \ln(y_i!) - \ln(1 + \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}))) \right] \\ &= L_{c1}(\boldsymbol{\gamma}; \mathbf{U}, \mathbf{x}_i, \mathbf{z}_i) + L_{c2}(\boldsymbol{\beta}; \mathbf{U}, \mathbf{x}_i, \mathbf{z}_i) - \sum_{i=1}^n (1 - u_i) \ln(y_i!), \end{aligned} \quad (4)$$

where $L_{c1} = \sum_{i=1}^n \{ u_i \mathbf{z}_i^\top \boldsymbol{\gamma} - \ln(1 + \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})) \}$, $L_{c2} = \sum_{i=1}^n (1 - u_i) (y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}))$, and $U = \{u_i; i = 1, 2, \dots, n\}$.

To implement the EM algorithm, we first initialize $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ by letting $U_i = I_{(y_i=0)}$. In the E-step, we use the initial values of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to calculate the expectation of U_i , and use it as the estimator of U_i . In the M-step, we use the estimate U_i to maximize $L_c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{U}, \mathbf{x}_i, \mathbf{z}_i)$, which gives the unrestricted maximum likelihood estimator for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The iteration $(k+1)$ of the EM algorithm requires the following steps.

E-Step: Estimate $U_i^{(l)}$ by using the means given $\boldsymbol{\gamma}^{(l)}$ and $\boldsymbol{\beta}^{(l)}$,

$$\begin{aligned} U_i^{(l)} &= E(U_i | y_i, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\beta}^{(l)}) \\ &= E(U_i = 1 | y_i, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\beta}^{(l)}) \\ &= \frac{Pr(Y_i = y_i | U_i = 1) Pr(U_i = 1)}{Pr(Y_i = y_i | U_i = 1) Pr(U_i = 1) + Pr(Y_i = y_i | U_i = 0) Pr(U_i = 0)} \\ &= \left[1 + \exp(-\exp(\mathbf{x}_i^\top \boldsymbol{\beta}^{(l)}) - \mathbf{z}_i^\top \boldsymbol{\gamma}^{(l)}) \right]^{-1}, \text{ if } y_i = 0, \text{ and } 0, \text{ if } y_i \geq 1. \end{aligned}$$

M-Step: Given $U_i = U_i^{(l)}$, maximize $L_{c1}(\boldsymbol{\gamma}; \mathbf{U}^{(l)}, \mathbf{x}_i, \mathbf{z}_i)$ and $L_{c2}(\boldsymbol{\beta}; \mathbf{U}^{(l)}, \mathbf{x}_i, \mathbf{z}_i)$ with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively:

$$\boldsymbol{\gamma}^{(l+1)} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\{ -L_{c1}(\boldsymbol{\gamma}; \mathbf{U}^{(l)}, y_i, \mathbf{x}_i, \mathbf{z}_i) \right\} \quad (5)$$

$$\boldsymbol{\beta}^{(l+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -L_{c2}(\boldsymbol{\beta}; \mathbf{U}^{(l)}, y_i, \mathbf{x}_i, \mathbf{z}_i) \right\} \quad (6)$$

The iteration stops when (5) and (6) converge simultaneously. The final estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\gamma}}^\top)^\top$ is known as the unrestricted maximum likelihood estimator (UMLE).

Let $I_{(\boldsymbol{\beta}, \boldsymbol{\gamma})}$ be the information matrix of the estimator $\hat{\boldsymbol{\theta}}$. If $\frac{1}{n}I_{(\boldsymbol{\beta}, \boldsymbol{\gamma})}$ has a positive definite limit satisfying some regularity conditions, as in the work of McCullagh (1983), the quantity $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean vector $\mathbf{0}$ and information matrix $I_{(\boldsymbol{\beta}, \boldsymbol{\gamma})}^{-1}$ (Lambert, 1992). The matrix $I_{(\boldsymbol{\beta}, \boldsymbol{\gamma})}$ can be partitioned as

$$\begin{pmatrix} I_{\boldsymbol{\beta}, \boldsymbol{\beta}} & I_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \\ I_{\boldsymbol{\gamma}, \boldsymbol{\beta}} & I_{\boldsymbol{\gamma}, \boldsymbol{\gamma}} \end{pmatrix},$$

where the elements $I_{\boldsymbol{\beta}, \boldsymbol{\beta}}, I_{\boldsymbol{\beta}, \boldsymbol{\gamma}} = I_{\boldsymbol{\gamma}, \boldsymbol{\beta}}^\top$, and $I_{\boldsymbol{\gamma}, \boldsymbol{\gamma}}$ are, respectively,

$$-\mathbb{E} \left[\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right], -\mathbb{E} \left[\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}} \right], \text{ and } -\mathbb{E} \left[\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top} \right]$$

with

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} &= \sum_{i=1}^n \left\{ I_{(y_i=0)} \left[\frac{-\exp(-\lambda_i)[(1-\lambda_i)p_i + (1-p_i)\exp(-\lambda_i)](1-p_i)\lambda_i}{[p_i + (1-p_i)\exp(-\lambda_i)]^2} \right] \right. \\ &\quad \left. + I_{(y_i>0)}(-\lambda_i) \right\} x_{ij}x_{jk}, \\ \frac{\partial^2 l}{\partial \beta_j \partial \gamma_r} &= \frac{\partial^2 l}{\partial \gamma_j \partial \beta_r} = \sum_{i=1}^n \left\{ I_{(y_i=0)} \left[\frac{\lambda_i \exp(-\lambda_i)}{[p_i + (1-p_i)\exp(-\lambda_i)]^2} \right] \right\} x_{ij}z_{ir}, \\ \frac{\partial^2 l}{\partial \gamma_r \partial \gamma_s} &= \sum_{i=1}^n \left\{ I_{(y_i=0)} \left[\frac{-(1-\exp(-\lambda_i))^2}{[p_i + (1-p_i)\exp(-\lambda_i)]^2} \right] + I_{(y_i>0)} \left[\frac{-1}{(1-p_i)^2} \right] \right\} z_{ir}z_{is}, \end{aligned}$$

where $j, k = 1, 2, \dots, p$ and $r, s = 1, 2, \dots, q$.

Suppose now that our interest centers in estimating the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from (4) under the linear restriction $\mathbf{H}\boldsymbol{\theta} = \mathbf{h}$. The steps of the EM-algorithm for estimating the parameters using log-likelihood (4) under the above restriction are the same. We simply replace $\boldsymbol{\theta}^{(l+1)} = (\boldsymbol{\beta}^{(l+1)}, \boldsymbol{\gamma}^{(l+1)})$ by $\boldsymbol{\theta}_c^{(l+1)} = (\boldsymbol{\beta}_c^{(l+1)}, \boldsymbol{\gamma}_c^{(l+1)})$ in the M-step. The resulting estimator, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}^\top, \tilde{\boldsymbol{\gamma}}^\top)^\top$ is known the restricted maximum likelihood estimator (RMLE).

The likelihood ratio test statistic can be used to test $H_0 : \mathbf{H}\boldsymbol{\theta} = \mathbf{h}$ vs. $H_a : \mathbf{H}\boldsymbol{\theta} \neq \mathbf{h}$. If $\tilde{\boldsymbol{\theta}}$ maximizes the log likelihood of the ZIPR model under H_0 of dimension $k-r$, where $k = p+q$ and $\hat{\boldsymbol{\theta}}$ maximizes the log likelihood of the ZIPR model under a nested alternative hypothesis H_A of dimension k , then the test statistic D_n is

$$\begin{aligned} D_n &= 2[l(\hat{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}})] \\ &= (\mathbf{H}\hat{\boldsymbol{\theta}} - \mathbf{h})^\top [\mathbf{H}\operatorname{var}(\hat{\boldsymbol{\theta}})\mathbf{H}^\top]^{-1} (\mathbf{H}\hat{\boldsymbol{\theta}} - \mathbf{h}) + o_p(1) \end{aligned} \quad (7)$$

Under H_0 and suitable regularity conditions, D_n is asymptotically distributed as chi-square with r degrees of freedom (Lambert, 1992). So that a likelihood ratio test can be performed using approximate critical values from the chi-square distribution.

2.2 The Pretest and Shrinkage Estimators

The pretest and shrinkage estimators are based on the test statistic D_n of (7) for testing $H_0 : \mathbf{H}\boldsymbol{\theta} = \mathbf{h}$. Specifically, the pretest estimator (PT) of $\boldsymbol{\theta}$ is defined as

$$\hat{\boldsymbol{\theta}}_{PT} = \hat{\boldsymbol{\theta}} - (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})I(D_n \leq \chi_{r,\alpha}^2),$$

where $I(A)$ is an indicator function of a set A , and $\chi_{r,\alpha}^2$ is the α -level critical value of the approximate distribution of D_n under H_0 . From the above definition, one can see that if the data yield $D_n < \chi_{r,\alpha}^2$, then $\hat{\boldsymbol{\theta}}_{PT} = \tilde{\boldsymbol{\theta}}$, otherwise

$\hat{\theta}_{PT} = \hat{\theta}$. So the PT is indeed a simple mixture of the UMLE and RMLE estimators. In an ordinary two-step procedure, one would test the hypothesis $H_0 : \mathbf{H}\boldsymbol{\theta} = \mathbf{h}$ first, then based on the test result decide which estimator should be adopted. The PT simply combines these two steps to form a single one. That is, testing and estimation are done simultaneously. For details, see Hossain et al. (2009) and Ahmed et al. (2006), and others.

Because of extreme choices for either the UMLE or RMLE, the pretest procedures are not admissible for some models, even though they may improve upon the UMLE, a well-documented fact in the literature (Judge & Bock, 1978). This motivates us to define a shrinkage estimator, which is a smoothed version of $\hat{\theta}_{PT}$:

$$\hat{\theta}_{SE} = \tilde{\theta} + \left(1 - (r-2)D_n^{-1}\right)(\hat{\theta} - \tilde{\theta}), \quad r \geq 3. \quad (8)$$

This estimator is a weighted average of UMLE $\hat{\theta}$ and RMLE $\tilde{\theta}$, where the weights are a function of the test statistic for testing $H_0 : \mathbf{H}\boldsymbol{\theta} = \mathbf{h}$.

We note that when the test statistic D_n is very small in comparison with $r-2$, i.e., when the ratio $(r-2)/D_n$ is greater than one in absolute value, the shrinkage estimator $\hat{\theta}_{SE}$ tends to shrink $\hat{\theta}$ overly towards $\tilde{\theta}$ and reversing the sign of $\hat{\theta}$. Replacing $(1 - (r-2)D_n^{-1})$ by $(1 - (r-2)D_n^{-1})_+$ in (8), where $(x)_+ = x1_{(x \geq 0)}$, the positive-part shrinkage estimator, $\hat{\theta}_{PSE}$ rectifies this problem. For details, see, Ahmed & Fallahpour (2012).

2.3 Penalty Estimators: LASSO and Adaptive LASSO

Let $\boldsymbol{\theta}$ be the parameter of interest and $\ell(\boldsymbol{\theta})$ be the log-likelihood function of the ZIPR, then the LASSO estimator (Tibshirani, 1996) is given by

$$\hat{\boldsymbol{\theta}}_{lasso}(\tau) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ -\ell(\boldsymbol{\theta}) + \tau \sum_{g=1}^{p+q} |\boldsymbol{\theta}_g| \right\},$$

where $\tau \geq 0$ is a penalization parameter controlling the amount of shrinkage on the regression coefficients. However, the LASSO applies same penalty to all the coefficients, which over-penalizes the important ones and accordingly results in biased estimators (Zou, 2006).

The adaptive LASSO (ALASSO) (Zou, 2006) offers an effective way to minimize this bias. It has been shown that the ALASSO enjoys the oracle property that is, the ALASSO is consistent in variable selection, and its estimators are asymptotically normal and unbiased. More explicitly, it assigns a higher weight to the small coefficients and lower weight to the large coefficients. The adaptive lasso estimator is defined as

$$\hat{\boldsymbol{\theta}}_{alasso}(\tau) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ -\ell(\boldsymbol{\theta}) + \tau \sum_{g=1}^{p+q} w_g |\boldsymbol{\theta}_g| \right\}, \quad (9)$$

where $\mathbf{w} = (w_1, w_2, \dots, w_{p+q})$ are adaptive weights, which are usually set to $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\theta}}|$, here $\hat{\boldsymbol{\theta}}$ are the maximum likelihood estimators. The ALASSO estimator can be obtained by minimizing penalized log-likelihood function. First, we expand the likelihood function based on a Taylor series expansion at $\hat{\boldsymbol{\theta}}$:

$$\ell(\boldsymbol{\theta}) \approx \ell(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \ell'(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \ell''(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \quad (10)$$

where ℓ' and ℓ'' are the first and second derivatives, respectively. Since $\ell(\hat{\boldsymbol{\theta}})$ is constant and $\ell'(\hat{\boldsymbol{\theta}}) = 0$, so the penalized log-likelihood function is equivalent to

$$\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top [-\ell''(\hat{\boldsymbol{\theta}})](\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \sum_{g=1}^{p+q} |\boldsymbol{\theta}_g|. \quad (11)$$

Let $\hat{\boldsymbol{\Sigma}}$ be the estimated variance-covariance matrix of $\hat{\boldsymbol{\theta}}$, which can be obtained from one of many statistical packages. And it provides an estimate of the Hessian matrix ℓ'' , given as $\ell'' = -I_{\hat{\boldsymbol{\theta}}} = -\hat{\boldsymbol{\Sigma}}^{-1}$. Here we use the *R* statistical package *pscl* (Jackman, 2012) in order to calculate $\hat{\boldsymbol{\Sigma}}$. Then we construct a pseudo-data set as

$$\mathbf{X}^* = \hat{\boldsymbol{\Sigma}}^{-1/2}, \mathbf{Y}^* = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\theta}}.$$

The order of the square matrix X^* is $p + q + 2$ and Y^* is a vector corresponding to X^* . The ALASSO estimator (9) can now be re-written as

$$\hat{\theta}_{alasso}(\tau) \approx \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2} (Y^* - X^* \theta)^\top (Y^* - X^* \theta) + \tau \sum_{g=1}^{p+q} w_g |\theta_g| \right\}, \quad (12)$$

The estimator in equation (12) is similar to the least squares estimator with adaptive LASSO penalization (Zou, 2006). Various efficient algorithms, such as the least angle regression algorithm (Efron et al., 2004), the predictor-corrector algorithm (Park & Hastie, 2007), and the coordinate descent algorithm (Friedman et al., 2010) can be used to conduct the minimization of Equation (12). Here we adopt the predictor-corrector algorithm which computes solutions along the entire penalization path of the coefficient estimates as τ varies. Starting at $\tau = \tau_{max}$, where τ_{max} is the largest τ that makes all the coefficients of $\hat{\theta}(\tau)$ nonzero. This algorithm computes a series of solutions, each time estimating the coefficients with a smaller τ . The penalization parameter τ is selected using k -fold cross validation. For each fold, we obtain a series of models based on Schwarz's Bayesian criterion, commonly referred to as Bayesian information criterion corresponding to the values of τ and compute the right side of (12) using the omitted fold. Then we choose the value of τ for which the average cross-validation in the right side of (12) is minimized.

In the case of pretest and shrinkage estimators, it is possible to determine theoretically when they have smaller asymptotic risks than the unrestricted MLE. Similar results are not available for LASSO and adaptive LASSO estimators.

3. Asymptotic Results

In this section we study the asymptotic behavior of the proposed estimators and develop the results in terms of $\hat{\theta}^*$, which could be any one of the estimators considered in this paper: $\hat{\theta}$, $\hat{\theta}_{PT}$, $\hat{\theta}_{SE}$, and $\hat{\theta}_{PSE}$. In order to proceed with the asymptotic results, we first partition $\theta = (\theta_1^\top, \theta_2^\top)^\top$, where θ_2 is the coefficient vector of the inactive covariates for the ZIPR model. The main focus here is to evaluate the performance of these estimators when θ_2 is close to the null vector. To derive any meaningful results we consider a sequence of local alternatives

$$K_n : H\theta = h + \frac{\omega}{\sqrt{n}}, \quad (13)$$

where $\omega = (\omega_1, \omega_2, \dots, \omega_r)^\top \in \mathfrak{R}^r$ is a given vector of real numbers. In this framework, $\theta = (\theta_1^\top, \mathbf{0}^\top)^\top$, and the quantity $\frac{\omega}{\sqrt{n}}$ is the magnitude of the distance between the true model and the restricted model. For any fixed ω , this distance shrinks as the sample size increases.

To study the asymptotic risks (ADR) of the estimators, we define a quadratic loss function by using a positive definite matrix W , namely

$$\mathcal{L}(\hat{\theta}^*; W) = \left[\sqrt{n}(\hat{\theta}^* - \theta) \right]^\top W \left[\sqrt{n}(\hat{\theta}^* - \theta) \right],$$

where $\hat{\theta}^*$ is any one of the estimators. The usual quadratic loss is defined when W is chosen as I , the identity matrix. A general W gives a loss function that weighs differently for different θ 's.

If V is the asymptotic variance-covariance matrix of $\hat{\theta}^*$, the ADR of $\sqrt{n}(\hat{\theta}^* - \theta)$ is given by $\operatorname{tr}(WV)$. We assume that the cumulative distribution function of $\hat{\theta}^*$ under K_n exists and can be denoted as

$$F(x) = \lim_{n \rightarrow \infty} P \left[\sqrt{n}(\hat{\theta}^* - \theta) \leq x | K_n \right],$$

where $F(x)$ is a nondegenerate distribution function. The ADR of $\hat{\theta}^*$ is then defined as

$$\begin{aligned} R(\hat{\theta}^*; W) &= \int \dots \int x' W x dF(x) \\ &= \operatorname{trace}(WV), \end{aligned} \quad (14)$$

where $V = \int \dots \int x x' dG(x)$ is the dispersion matrix for the distribution function $F(x)$.

The shrinkage estimators are, in general biased, the bias, however is accompanied by a reduction in variance. The asymptotic distributional bias (ADB) of an estimator $\hat{\theta}^*$ is defined as

$$\operatorname{ADB}(\hat{\theta}^*) = \lim_{n \rightarrow \infty} E \left\{ n^{\frac{1}{2}} (\hat{\theta}^* - \theta) \right\}.$$

Under the local alternatives (13), the following theorems help the derivation and numerical computation of the ADB and the ADR of the estimators.

Theorem 3.1 *If $I_{\beta,\gamma}$ is nonsingular and $\Delta = \omega^\top (\mathbf{H}I_{\beta,\gamma}^{-1}\mathbf{H}^\top)^{-1}\omega$, then under the local alternatives K_n in (13) and regularity conditions, we have as $n \rightarrow \infty$*

1. $\sqrt{n}(\mathbf{H}\hat{\boldsymbol{\theta}} - \mathbf{h}) \xrightarrow{L} N(\omega, \mathbf{H}I_{\beta,\gamma}^{-1}\mathbf{H}^\top)$.
2. The test statistic D_n in (7) converges to a non-central chi-squared distribution $\chi_r^2(\Delta)$ with r degrees of freedom and non-centrality parameter Δ .

For a concise statement of the results, we use the following notations:

$$\begin{aligned}\zeta &= \mathbf{I}^{-1}\mathbf{H}^\top(\mathbf{H}\mathbf{I}^{-1}\mathbf{H}^\top)^{-1} \\ g_{r+i}(\Delta) &= H_{r+i}(r-2, \Delta), \quad h_{j,r+i}(\Delta) = E(\chi_{r+i}^{-2j}(\Delta)) \\ k_{j,r+i}(\Delta) &= H_{r+i}(r-2, \Delta) + (r-2)^j E(\chi_{r+i}^{-2j}(\Delta)I(\chi_{r+i}^2(\Delta) < r-2)), \quad \text{and} \\ O_{j,r+i}(\Delta) &= E((1 - (r-2)E(\chi_{r+i}^{-2j}(\Delta)))^2 I(\chi_{r+i}^2(\Delta) < r-2))\end{aligned}$$

where $i = 2, 4$, $j = 1, 2$ and for simplicity we assume that $I_{\beta,\gamma} = \mathbf{I}$. Here $H_{r+2}(\cdot, \Delta)$ is the distribution function of the $\chi_r^2(\Delta)$ distribution.

Theorem 3.2 *Using the definition of ADB and Theorem 3.1, the ADBs of the estimators are,*

$$\begin{aligned}ADB(\hat{\boldsymbol{\theta}}) &= \mathbf{0}, \quad ADB(\tilde{\boldsymbol{\theta}}) = -\zeta\omega, \quad ADB(\hat{\boldsymbol{\theta}}_{PT}) = -\zeta\omega g_{r+2}(\Delta) \\ ADB(\hat{\boldsymbol{\theta}}_{SE}) &= -(r-2)\zeta\omega h_{1,r+2}(\Delta), \quad ADB(\hat{\boldsymbol{\theta}}_{PSE}) = ADB(\hat{\boldsymbol{\theta}}_{SE}) - \zeta\omega k_{1,r+2}(\Delta).\end{aligned}$$

The outline of the proof is similar to that of Theorem 3.1 of Ahmed & Fallahpour (2012).

Clearly, the bias of the estimators is a function of Δ . Therefore, for bias comparisons, it suffices to compare the scalar factor Δ only. It is clear that the ADB of RMLE is an unbounded function of Δ . The $ADB(\hat{\boldsymbol{\theta}}_{SE})$ and $ADB(\hat{\boldsymbol{\theta}}_{PSE})$ start from the origin, and as Δ increases, they increase to a maximum and then decrease to 0. Note that, $h_{1,r+2}(\Delta)$ is a decreasing log-convex function of Δ and the ADB of $\hat{\boldsymbol{\theta}}_{PT}$ is a function of Δ and α . For a fixed α , $ADB(\hat{\boldsymbol{\theta}}_{PT})$ starts at zero, increases to a point, then decreases gradually to zero.

Theorem 3.3 *Under the local alternatives K_n in (13) and the assumptions of Theorem 3.1, the ADRs of the estimator are*

$$\begin{aligned}ADR(\hat{\boldsymbol{\theta}}; \mathbf{W}) &= tr(\mathbf{W}\mathbf{I}^{-1}) \\ ADR(\tilde{\boldsymbol{\theta}}; \mathbf{W}) &= ADR(\hat{\boldsymbol{\theta}}; \mathbf{W}) - tr(\mathbf{W}\zeta\mathbf{H}\mathbf{I}^{-1}) + \omega^\top(\zeta^\top\mathbf{W}\zeta)\omega \\ ADR(\hat{\boldsymbol{\theta}}_{PT}; \mathbf{W}) &= ADR(\hat{\boldsymbol{\theta}}; \mathbf{W}) - g_{r+2}(\Delta)tr(\mathbf{W}\zeta\mathbf{H}\mathbf{I}^{-1}) + \omega^\top(\zeta^\top\mathbf{W}\zeta)\omega(2g_{r+2}(\Delta) - g_{r+4}(\Delta)) \\ ADR(\hat{\boldsymbol{\theta}}_{SE}; \mathbf{W}) &= ADR(\hat{\boldsymbol{\theta}}; \mathbf{W}) + ((r-2)^2 h_{2,r+2}(\Delta) - 2(r-2)h_{1,r+2}(\Delta))tr(\mathbf{W}\zeta\mathbf{H}\mathbf{I}^{-1}) \\ &+ \omega^\top(\zeta^\top\mathbf{W}\zeta)\omega((r-2)^2 h_{2,r+4}(\Delta) + 2(r-2)h_{1,r+2}(\Delta) - 2(r-2)h_{1,r+2}(\Delta)) \\ ADR(\hat{\boldsymbol{\theta}}_{PSE}; \mathbf{W}) &= ADR(\hat{\boldsymbol{\theta}}_{SE}; \mathbf{W}) - O_{1,r+2}(\Delta)tr(\mathbf{W}\zeta\mathbf{H}\mathbf{I}^{-1}) - O_{1,r+4}(\Delta)\omega^\top(\zeta^\top\mathbf{W}\zeta)\omega \\ &+ \frac{1}{2}(3g_{r+2}(\Delta) + k_{1,r+2}(\Delta))\omega^\top(\zeta^\top\mathbf{W}\zeta)\omega.\end{aligned}$$

The outline of the proof is similar to that of Theorem 3.2 of Ahmed & Fallahpour (2012).

By comparing the ADRs of the estimators, we see that, as Δ moves away from 0, the risk of $\tilde{\boldsymbol{\theta}}$ becomes unbounded. That is, the RMLE $\tilde{\boldsymbol{\theta}}$ dominates the unrestricted estimator at and near $\Delta = 0$. The risk of $\hat{\boldsymbol{\theta}}_{PSE}$ is asymptotically superior to $\hat{\boldsymbol{\theta}}_{SE}$ for all values of Δ , with strict inequality holds for some Δ . Thus, not only does $\hat{\boldsymbol{\theta}}_{PSE}$ confirm the inadmissibility of $\hat{\boldsymbol{\theta}}_{SE}$, but it also provides a simple superior estimator. Further, the largest risk improvement of $\hat{\boldsymbol{\theta}}_{PSE}$ over $\hat{\boldsymbol{\theta}}_{SE}$ is at and near the null hypothesis. Also, by comparing the ADRs of $\hat{\boldsymbol{\theta}}_{SE}$, $\hat{\boldsymbol{\theta}}_{PSE}$, and $\hat{\boldsymbol{\theta}}$, it can be easily shown that, under certain conditions $ADR(\hat{\boldsymbol{\theta}}_{PSE}, \mathbf{W}) \leq ADR(\hat{\boldsymbol{\theta}}_{SE}, \mathbf{W}) \leq ADR(\hat{\boldsymbol{\theta}}, \mathbf{W})$ for all $\Delta \geq 0$. For a given α , PT is not uniformly better than the unrestricted estimator near the null hypothesis. One may determine an α such that PT has a minimum guaranteed risk. For any given \mathbf{H} , \mathbf{I} , and \mathbf{h} , the relative efficiency (inverse of risk) of PT

to UMLE is a function of α and λ . If the minimum efficiency required is RE_0 , then we can choose α by solving the equation $\min_{\lambda} \{\text{Relative Efficiency}(\alpha, \Delta)\} = RE_0$. The exact solution may not be available, but one can use a numerical method to search for the minimum.

In order to explain and quantify the properties of the theoretical results, we conduct a simulation study to compare the performance of the suggested estimators.

4. Simulation Studies

The zero-inflated count responses are generated from the following mixture of two models:

$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{and} \quad \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{z}_i^\top \boldsymbol{\gamma},$$

where \mathbf{x}_i are generated from a multivariate normal distribution $N_p(0, \boldsymbol{\Sigma})$, with elements $\sigma_{ij} = \rho^{|i-j|}$ ($i, j = 1, 2, \dots, 20$) and \mathbf{z}_i are generated from uniform, and Poisson distributions. We consider the correlation, $\rho = 0.50$ among the covariates, and sample sizes $n = 150$ and 350 . For small samples, say $n = 50$, the likelihood function does not converge. The estimations of the variance-covariance matrix as well as the regression coefficients in such cases are unstable for the ZIPR model. We tried different sample sizes. In order for the likelihood function to converge, it needed a minimum sample of size 150. We set $\alpha = 0.05$ for the pretest method. Our objective is to test the hypothesis

$$H_0 : \boldsymbol{\theta}_2 = \mathbf{0} \quad \text{versus} \quad H_A : \boldsymbol{\theta}_2 \neq \mathbf{0}.$$

We set the true values of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top = \{(\boldsymbol{\beta}_1, \boldsymbol{\gamma}_1)^\top, (\boldsymbol{\beta}_2, \boldsymbol{\gamma}_2)^\top\}$ as:

$$\boldsymbol{\theta} = ((1.1, -0.36, -0.36, 1.08, 0.92)^\top, -0.2^\top), \mathbf{b}^\top),$$

where $\mathbf{b} = \boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2, \boldsymbol{\gamma}_2)$ is a zero vector with different lengths. The number of replications is set to 1,000 for all cases. We define the distance between the simulation model and the restricted model by

$$\Delta = \|\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\|^2,$$

where $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^\top, \mathbf{0}^\top)^\top$, is a 6×1 true parameter vector in the simulation model and $\|\cdot\|$ is the Euclidian norm. Samples are generated using Δ between 0 and 1.

Based on the simulated data, we estimate the mean squared errors (MSE) of all the estimators studied in this paper. We calculate the simulated mean squared error (SMSE) by using the the empirical formula

$$SMSE(\boldsymbol{\theta}^*) = \sum_{i=1}^{p+q} (\theta_i^* - \theta_i)^2,$$

where $\hat{\boldsymbol{\theta}}^*$ as before denotes any one of the estimators $\hat{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_{PT}, \hat{\boldsymbol{\theta}}_{SE}, \hat{\boldsymbol{\theta}}_{PSE}, \hat{\boldsymbol{\theta}}_{lasso}$, and $\hat{\boldsymbol{\theta}}_{alasso}$. The simulated relative efficiency (SRE) between two estimators is defined as $SRE(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^*) = SMSE(\hat{\boldsymbol{\theta}})/SMSE(\hat{\boldsymbol{\theta}}^*)$, where $\hat{\boldsymbol{\theta}}$ is the unrestricted estimator, treated as benchmark. A value of $SRE(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^*)$ greater than one indicates that $\hat{\boldsymbol{\theta}}^*$ performs better than $\hat{\boldsymbol{\theta}}$, and vice versa. Table 1 and Figure 1 give results for $\Delta \geq 0$ and $(k_1, r) = (6, 3), (6, 6), (6, 11), (6, 14)$, where $k = p + q = (k_1, r)$. All analyses are performed using the statistical software *R*.

We observe from Table 1 and Figure 1 that when $\Delta = 0$, the RMLE outperforms all other estimators for both the cases considered. As Δ moves away from 0, the RMLE becomes unbounded. That is, the SRE curve for RMLE with respect to UMLE sharply decays and that it goes below the horizontal line when $\Delta > 0$. The positive-shrinkage estimator converges to 1 at a slower rate as Δ moves away from 0. This indicates that in the event of imprecise auxiliary information about the regression parameters, the PSE has the smallest risk among all other estimators for a range of Δ irrespective of the value of ρ . Pretest estimator dominates the shrinkage estimators when the restriction is nearly correct. Otherwise, it becomes unbounded at a faster rate than the RMLE. However, with the increase in Δ , the SRE of PT approaches 1 at some point from below. This suggests that neither PT nor RMLE is uniformly better when $\Delta > 0$.

The simulated relative efficiency of the shrinkage estimator typically lies above that of the UMLE, and is nearly as good as the RMLE at and near the null hypothesis $\Delta = 0$. This improvement relative to the RMLE is substantial for the increasing number of inactive covariates. Some readers may be surprised by the strong performance of

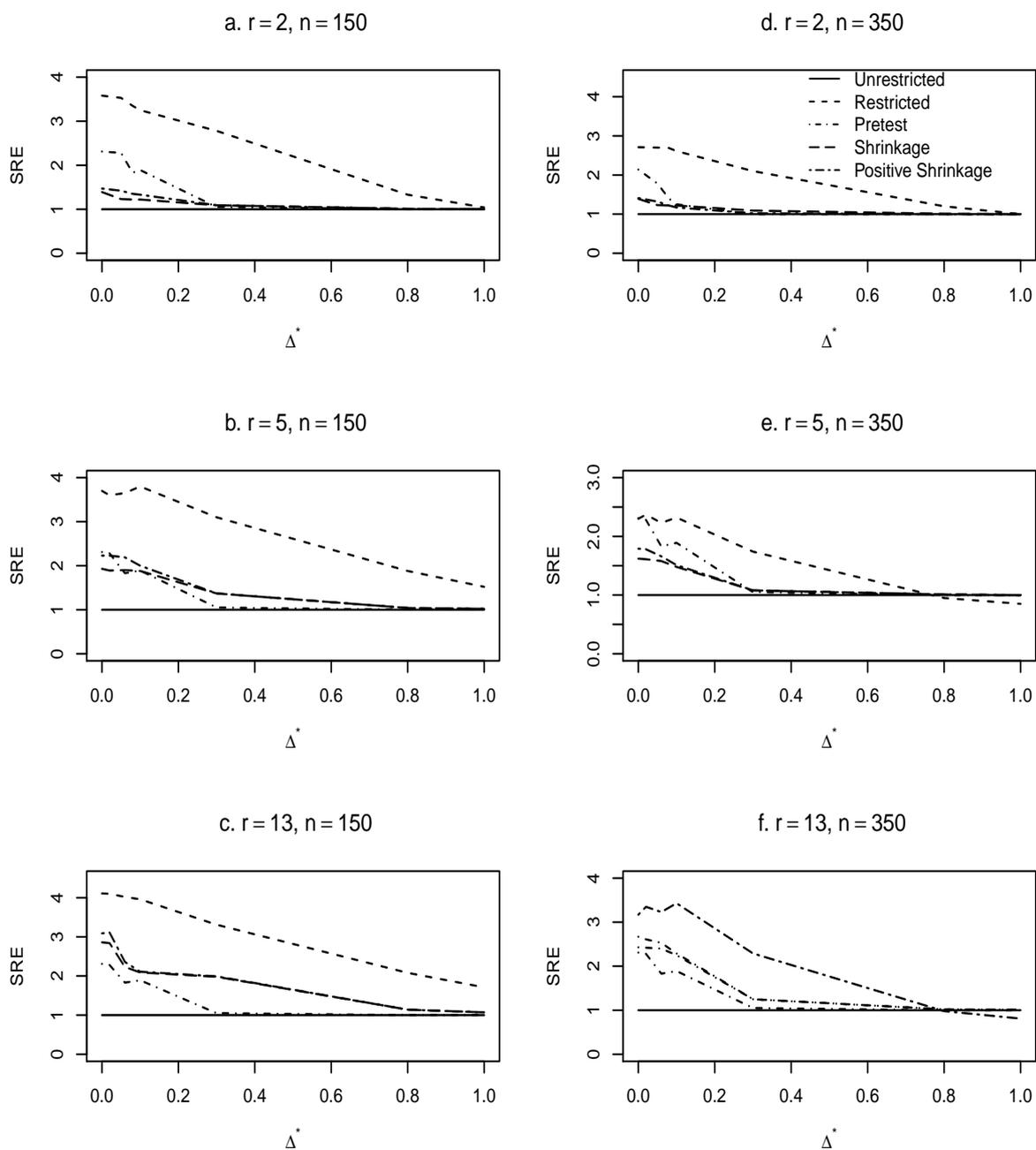


Figure 1. Simulated Relative efficiencies with respect to $\hat{\theta}$ of the estimators for $\Delta \geq 0$. Here $k_1 = 6, r = 2, 10, 18; n = 150$ for the first column (Figures a-c); $k_1 = 6, r = 2, 10, 18; n = 350$ for the second column (Figures d-e).

the shrinkage estimators relative to the UMLE. However, this is precisely the lesson of Theorems 3.2 and 3.3. Shrinkage strictly improves the asymptotic risk, and the improvement can be especially strong in high-dimensional cases. This holds true for other n and r values.

For the LASSO and adaptive LASSO estimators, we use the local quadratic approximation algorithm of Ulbricht (2010) for finding the entire solution path. We also use a k -fold cross validation procedure in order to choose the value of the penalization parameter τ that achieves the lowest BIC score. Table 1 shows that when $r < 6$, the LASSO and adaptive LASSO estimators are better than the PSE while it performs well for $r \geq 6$. The PSE is preferable in the presence of many inactive predictors in the model.

Table 1. Simulated relative efficiency of RE, SE, PSE, LASSO, and adaptive LASSO with respect to $\hat{\theta}$ when the restricted parameter space $H\theta = h$ is correct ($\Delta = 0$).

n=150				
Method	$p_2 = 3$	$p_2 = 6$	$p_2 = 11$	$p_2 = 14$
RE	3.58	3.70	4.08	4.11
PT	2.31	2.62	3.11	3.02
SE	1.39	1.93	2.76	2.87
PSE	1.47	2.23	2.87	3.07
LASSO	1.48	1.86	2.51	2.75
ALASSO	1.51	1.95	2.61	2.83
n=350				
RE	2.71	2.30	2.77	3.17
PT	2.14	2.04	2.11	2.59
SE	1.27	1.62	2.06	2.43
PSE	1.41	1.79	2.23	2.67
LASSO	1.43	1.53	1.63	1.67
ALASSO	1.49	1.60	1.72	1.81

5. Application to Real Life Data

The data are from the National Medical Expenditure Survey (NMES) conducted in 1987 and 1988 to provide a comprehensive picture of how Americans use and pay for health services. It includes 4406 respondents aged 66 or older and covered by Medicare program. Details of the data can be found in Deb & Trivedi (1997). In this example, the number of hospital stays (hosp) is used as the response variable and the covariates are self-perceived health status (poorh1th and exclh1th), number of chronic conditions (numchron), gender (male), age, number of years of education (school), family income (faminc), and private insurance indicator (privins). Approximately 80% of the responses are equal to zero, which corresponds to patients with zero day stay in the hospital. We first look at the count portion of the data, which refers to the respondents who have stayed at least one day in the hospital. On the other hand, if we look at the logistic portion (i.e. inflation part) of the data, which predict whether outcome is positive or zero. All covariates are included in both portions of the model to predict the number of hospital stays. To assess the effect of covariates on the hosp on each portion, ZIPR model is fitted. The AIC and BIC criteria for the mixture model (Wang et al., 1996) show that poorh1th, exclh1th and numchron are the important factors for the Poisson part and poorh1th, numchron, age, and male are the important factors for the logistic part of the ZIPR model. So the other six covariates in the Poisson part and five covariates in the logistic part are not significantly related to the response hosp. In this example, our hypothesis is $H\theta = \mathbf{0}$ where H is a 7×18 matrix and θ is a 18×1 vector of covariates.

We apply the bootstrap method (see, Jung et al., 2005) to examine the performance of the proposed estimation strategies for estimating the coefficients of the other seven covariates. We draw 1000 bootstrap samples of size $n = 2000$ by drawing 2000 rows with replacement from the data matrix (y_i, x_{ij}) and compute the point estimates, standard errors, and relative efficiencies of the proposed estimators. Here, the empirical distribution function \hat{F} based on original 4406 individuals is regarded as the true distribution, and the coefficients of the ZIPR closest to \hat{F} are regarded as true parameter values. The results given in Table 2 reveal that the restricted, pretest, shrinkage, positive shrinkage, LASSO, and adaptive LASSO estimators are superior to the maximum likelihood estimator, which is in agreement with our simulation results.

Table 2. Estimate (first row) and standard error (second row) for poorhlth (β_1), exclhlth (β_2), numchron (β_3) of the Poisson part and poorhlth (γ_1), numchron (γ_2), age (γ_3), and male (γ_4) of logistic part on the number of hospital stays. The SRE column gives the relative efficiency based on bootstrap simulation of the estimators with respect to UMLE.

Estimators	β_1	β_2	β_3	γ_1	γ_2	γ_3	γ_4	SRE
UMLE	-0.99	-0.313	0.123	-0.637	-0.302	-0.446	-0.381	1.000
	0.502	0.125	0.048	0.256	0.095	0.153	0.246	
RMLE	-0.715	-0.335	0.118	-0.555	-0.297	-0.375	-0.286	1.444
	0.186	0.127	0.042	0.230	0.079	0.086	0.107	
PT	-0.882	-0.321	0.122	-0.609	-0.301	-0.421	-0.354	1.194
	0.448	0.126	0.046	0.255	0.090	0.131	0.214	
SE	-0.861	-0.323	0.121	-0.601	-0.300	-0.414	-0.342	1.226
	0.334	0.124	0.045	0.243	0.086	0.117	0.182	
PSE	-0.863	-0.322	0.121	-0.602	-0.300	-0.415	-0.343	1.226
	0.332	0.124	0.044	0.244	0.086	0.117	0.182	
LASSO	-0.543	-0.190	0.121	-0.435	-0.254	-0.069	-0.417	1.178
	0.267	0.128	0.057	0.291	0.106	0.089	0.099	
ALASSO	-0.473	-0.197	0.121	-0.439	-0.257	-0.068	-0.445	1.221
	0.231	0.136	0.058	0.248	0.106	0.073	0.098	

6. Conclusion

In this paper, we have introduced the pretest, shrinkage, LASSO, and ALASSO estimators for zero-inflated Poisson regression model when it is suspected that some of the regression coefficients may be restricted to a subspace. We have presented ADB and ADR expressions of the pretest and shrinkage estimators. We conducted a Monte Carlo simulation experiment to examine the performance of the proposed estimators which showed that the RMLE outperforms the usual UMLE at or near the restriction. However, as we deviate from the restriction, risk of the RMLE becomes unbounded. And near the restriction, the risk of the pretest estimator is less than that of the UMLE. As Δ , the distance between the simulated and restricted models increases, the risk of the pretest estimator crosses the risk of the UMLE, reaching a maximum, and then decreasing monotonically to the risk of the UMLE. Furthermore, the shrinkage estimators with data based weights perform well if the restriction is true. In fact, the shrinkage estimators outperform the UMLE in the entire parameter space when the number of inactive parameter is greater than two. In addition, the performance of the shrinkage estimators improve relative to the UMLE when the number of inactive covariates increases. The penalty estimators are preferable when there are fewer inactive covariates. Thus in presence of many inactive covariates, the shrinkage estimators are attractive alternatives to penalty estimators.

Finally, we applied the proposed strategies to a real life data set to evaluate the relative performance of the proposed estimators. The results are consistent with our analytical and simulated findings. The theoretical and numerical results of zero-inflated Poisson can be extended to the entire class of generalized linear models. We are currently working on this project.

In terms of recommendations, the shrinkage estimators should be used instead of the UMLE when the restriction on regression coefficients agrees with the data. If we are uncertain of the quality of auxiliary information about the covariates, we can still use the shrinkage estimation because it offers a lower or at best an equal risk relative to the UMLE over the entire parameter space.

Acknowledgements

The authors thank the Editor and the referees for their helpful comments which has improved the earlier version of the manuscript. The research of Shakhawat Hossain was supported by the Natural Sciences and the Engineering Research Council of Canada grant and that of Hatem A. Howlader was supported by a grant from the University of Winnipeg.

References

- Ahmed, S. E., Hussein, A. A., & Sen, P. K. (2006). Risk comparison of some shrinkage M-estimators in linear models. *Journal of Nonparametric Statistics*, 18 (4-6), 401-415. <http://dx.doi.org/10.1080/10485250601046752>
- Ahmed, S. E., & Fallahpour, S. (2012). Shrinkage estimation strategy in quasi-likelihood models. *Statistics and Probability Letters*, 82, 2170-2179. <http://dx.doi.org/10.1016/j.spl.2012.08.001>
- Deb, P., & Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Economics*, 12(3), 313-336. [http://dx.doi.org/10.1002/\(SICI\)1099-1255\(199705\)12:3<313::AID-JAE440>3.0.CO;2-G](http://dx.doi.org/10.1002/(SICI)1099-1255(199705)12:3<313::AID-JAE440>3.0.CO;2-G)
- Dietz, E., & Bohning, D. (2000). On estimation of the Poisson parameter in zero-inflated Poisson models. *Computational Statistics and Data Analysis*, 34(4), 441-459. [http://dx.doi.org/10.1016/S0167-9473\(99\)00111-5](http://dx.doi.org/10.1016/S0167-9473(99)00111-5)
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2006). Least Angle Regression. *Annals of Statistics*, 32, 407-499.
- Feng, J., & Zhu, Z. (2012). Semiparametric analysis of longitudinal zero-inflated count data. *Journal of Multivariate Analysis*, 102, 61-72. <http://dx.doi.org/10.1016/j.jmva.2010.08.001>
- Friedman, J., Hastie, T., & Tibshirani, R. (2000). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1-22.
- Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56, 1030-1039. <http://dx.doi.org/10.1111/j.0006-341X.2000.01030.x>
- Hinde, J. P., & Demetrio, C. G. B. (1998). Over-dispersion: Models and estimation. *Computational Statistics and Data Analysis*, 27, 151-170. [http://dx.doi.org/10.1016/S0167-9473\(98\)00007-3](http://dx.doi.org/10.1016/S0167-9473(98)00007-3)
- Hossain, S., Doksum, K. A., & Ahmed, S. E. (2009). Positive-part shrinkage and Absolute Penalty Estimators in Partially Linear Models. *Linear Algebra and its Applications*, 430(3), 2749-2761. <http://dx.doi.org/10.1016/j.laa.2008.12.015>
- Hossain, S., Ahmed, S. E., & Howlader, H. A. (2012). Model selection and parameter estimation of a multinomial logistic regression model. *Journal of Statistical Computation and Simulation*, 84(7), 1412-1426. <http://dx.doi.org/10.1080/00949655.2012.746347>
- Hossain, S., & Ahmed, S. E. (2012). Shrinkage and Penalty Estimators of a Poisson Regression Model. *Australian and New Zealand Journal of Statistics*, 54(3), 359-373. <http://dx.doi.org/10.1111/j.1467-842X.2012.00679.x>
- Jackman, S. (2012). pscl: Political Science Computational Laboratory, Stanford University. *R package version 1.4.9*. <http://CRAN.R-project.org/package=pscl>
- Jansakul, N., & Hinde, J. P. (2002). Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis*, 40, 75-96. [http://dx.doi.org/10.1016/S0167-9473\(01\)00104-9](http://dx.doi.org/10.1016/S0167-9473(01)00104-9)
- Judge, G.G., & Bock, M. E. (1978). *The Statistical Implications of pretest and Stein-Rule Estimators in Econometrics*. Amsterdam: North Holland Publishing Company.
- Jung, B. C., Jhun, M., & Lee, J. W. (2005). Bootstrap Tests for Overdispersion in a Zero-Inflated Poisson Regression Model. *Biometrics*, 61, 626-629. <http://dx.doi.org/10.1111/j.1541-0420.2005.00368.x>
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1-14. <http://dx.doi.org/10.2307/1269547>
- Liang, H., & Song, W. (2009). Improved estimation in multiple linear regression models with measurement error and general constraint. *Journal of Multivariate Analysis*, 100, 726-741. <http://dx.doi.org/10.1016/j.jmva.2008.08.003>
- McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, 11, 59-67. <http://dx.doi.org/10.1214/aos/1176346056>
- Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3), 341-365. [http://dx.doi.org/10.1016/0304-4076\(86\)90002-3](http://dx.doi.org/10.1016/0304-4076(86)90002-3)
- Nkurunziza, S. (2013). Extension of some important identities in shrinkage-pretest strategies. *Metrika*, 76(7), 937-947. <http://dx.doi.org/10.1007/s00184-012-0425-5>

- Park, M. Y., & Hastie, T. (2007). An L_1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B*, 69, 659-677. <http://dx.doi.org/10.1111/j.1467-9868.2007.00607.x>
- Ridout, M. S., Demetrio, C. G. B., & Hinde, J. P. (1998). *Models for count data with many zeros*. In: International Biometric Conference. Cape Town, pp. 179-190.
- Ridout, M. S., Hind, J. P., & Demetrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57, 219-223. <http://dx.doi.org/10.1111/j.0006-341X.2001.00219.x>
- Sapra, S. K. (2003). Pre-test estimation in Poisson regression model. *Applied Economics Letters*, 10, 541-543. <http://dx.doi.org/10.1080/1350485032000100215>
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58(7), 267-288.
- van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, 51, 738-743. <http://dx.doi.org/10.2307/2532959>
- Wang, P., Puterman, M. L., Cockburn, I., & Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics*, 52, 381-400. <http://dx.doi.org/10.2307/2532881>
- Zeng, H., Wei, Y., Zhao, Y., Liu, L., Zhang, R., Gou, J., Huang, S., & Chen, F. (2014). Variable selection approach for zero-inflated count data via adaptive LASSO. *Journal of Applied Statistics*, 41(4), 879-894. <http://dx.doi.org/10.1080/02664763.2013.858672>
- Zou, H. (2006). The Adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429. <http://dx.doi.org/10.1198/016214506000000735>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).