REGRESSION ANALYSIS OF BIG COUNT DATA VIA A-OPTIMAL SUBSAMPLING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Xiaofeng Zhao

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2018

Purdue University

Indianapolis, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Fei Tan, Co-Chair

Department of Mathematical Sciences

Dr. Hanxiang Peng, Co-Chair

Department of Mathematical Sciences

Dr. Fang Li

Department of Mathematical Sciences

Dr. Zuofeng Shang

Department of Mathematical Sciences

Dr. Honglang Wang

Department of Mathematical Sciences

Approved by:

Dr. Evgeny Mukhin

Head of the Graduate Program

This thesis is dedicated to my parents.

ACKNOWLEDGMENTS

I want to take this chance to express my sincere gratitude for all the efforts from Professor Fei Tan and Professor Hanxiang Peng, who became to be my advisor and co-advisor four years ago and guide me to the area of big data analysis. During these years, their contributions of time, energy, and patience have made my Ph.D pursuit productive and effective.

Besides my advisors, I would like to thank Professor Benzion Boukai, Professor Fang Li, Professor Jyoti Sarkar, who helped me in their classes when I began to study in the math department. I would also like to acknowledge the effort from Professor Zhongmin Shen, who encouraged me to become part of the Ph.D program. My sincere thanks also goes to Professor Zuofeng Shang, Professor Honglang Wang, who generously became my thesis committee.

At last, I would like to appreciate those who have supported my study and research in Indiana University Purdue University Indianapolis.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

Table Page

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

Zhao, Xiaofeng Ph.D., Purdue University, August 2018. Regression Analysis of Big Count Data Via A-Optimal Subsampling. Major Professors: Fei Tan and Hanxiang Peng.

There are two computational bottlenecks for Big Data analysis: (1) the data is too large for a desktop to store, and (2) the computing task takes too long waiting time to finish. While the Divide-and-Conquer approach easily breaks the first bottleneck, the Subsampling approach simultaneously beat both of them. The uniform sampling and the nonuniform sampling–the Leverage Scores sampling– are frequently used in the recent development of fast randomized algorithms. However, both approaches, as Peng and Tan (2018) have demonstrated, are not effective in extracting important information from data. In this thesis, we conduct regression analysis for big count data via A-optimal subsampling. We derive A-optimal sampling distributions by minimizing the trace of certain dispersion matrices in general estimating equations (GEE). We point out that the A-optimal distributions have the same running times as the full data M-estimator. To fast compute the distributions, we propose the A-optimal Scoring Algorithm, which is implementable by parallel computing and sequentially updatable for stream data, and has faster running time than that of the full data M-estimator. We present asymptotic normality for the estimates in GEE's and in generalized count regression. A data truncation method is introduced. We conduct extensive simulations to evaluate the numerical performance of the proposed sampling distributions. We apply the proposed A-optimal subsampling method to analyze two real count data sets, the Bike Sharing data and the Blog Feedback data. Our results in both simulations and real data sets indicated that the A-optimal distributions substantially outperformed the uniform distribution, and have faster running times than the full data M-estimators.

1. INTRODUCTION

This dissertation is concerned with fast regression methods for big count data with nonnegative integer response. Our approach is *optimal subsampling*.

1.1 Review of Regression Analysis of Count Data

Count data are observations of the number of occurrences of a behavior in a fixed period of time. Count data are common, for example, hospital visits, blog comments, car/bike renters, and questionnaire respondents.

Analysis of count data is an important task in social sciences and economics. Since linear regression does not take into account the restricted number of count response values it is not an appropriate technique for count data. Standard regression methods include Poisson, overdispersed Poisson, negative binomial, and zero-inflated Poisson regressions, as well as truncated methods and quasi-likelihood approach.

The Poisson regression and Negative binomial approach are often used in count data analysis. It is motivated by the usual consideration for regression analysis, meanwhile, seek to protect and exploit the nonnegative and integer-valued characteristic of the outcome as much as possible. The scope of count data is very wide, including sociology, marketing, demographic economics, crime victimology, political science, doctor visits, credit reports, recreational trips, bank failures, accident insurance, doctoral publications, and manufacturing defects. Count data analysis has drawn a lot of attention and been a influential part in statistic modeling.

The most frequently used regression approach for count variables is probably Poisson regression. However, Poisson regression requires distributional assumptions. It is often of limited use in real data because real count data usually exhibit overdispersion, an inflated number of zeros, an absence of certain counts, censoring counts, and missing counts.

Overdisperson can be addressed by generalizing the Poisson model to, for instance, quasi-Poisson models. Another useful approach is the negative binomial regression. These models are related to the generalized linear models family see, e.g., Nelder and Wedderburn 1972; McCullagh and Nelder 1989; Dobson (2002).

The above models can deal with over-dispersion rather well, but are not enough for modeling excess zeros. To address this, researchers have developed methods for zero-inflated data by including another model component to capture zero counts. This is done by a mixture model that unifies a count component and a point mass at zero, see Cameron and Trivedi (2005).

To deal with truncated data and censored counts, Hurdle models were proposed in (Mullahy 1986). These models combine a count component that is left-truncated with a hurdle component that is right-censored.

1.2 Big Data Analysis

Big Data are on a massive scale with regard to volume, velocity, variety, and veracity that exceed both the capacity of the conventional software tools and operating systems and the physical spaces of computers, see e.g. Wang, *et al.* (2015); Fan, *et al.* (2013). Massive data pose two computational bottlenecks: (1) the data exceed a computer's memory, and (2) the computing task requires too long waiting time to finish. The two bottlenecks can be simultaneously addressed by *judiciously* choosing a sub-data as a surrogate for the full data and completing the data analysis. This is the goal that this dissertation will pursue.

While the often used Divide-and-Conquer approach readily breaks the memory limit, the proposed subsampling approach not only breaks the limit but speed up computing as well as possesses other useful statistical properties. Due to its mathematical simplicity and computational ease, the *uniform sampling* is often used in subsampling for intensive computing and for development of fast randomized algorithms and in re-sampling for Monte Carlo and bootstrap. The uniform sampling, however, is not effective in extracting information, see a simulation study in in Peng an Tan (2018). In this dissertation, non-uniform sampling distributions on data points by the criterion of A-optimality will be sought, that is, by minimizing the trace of certain variance-covariance matrix. Equivalently, Distributions will be sought to minimizing the sum of the component variances of certain subsampling estimate. Mathematicians, computer scientists and statisticians have already made important progress in this area. Drineas, *et al.* (2006a) constructed fast Monte Carlo algorithms to approximate matrix multiplication. Drineas, *et al.* (2006b) presented a sampling algorithm for the least squares fit problem and studied its algorithmic properties. A key feature of the above algorithms is the non-uniform sampling. Ma and Sun (2014) and Ma, *et al.* (2015) used the leverage scores as non-uniform importance sampling distributions for big data linear regression. Zhu, *et al.* (2015) obtained optimal subsampling distributions for large sample linear regression. Wang, *et al.* (2015) constructed optimal subsampling for large logistic regression. Xu, *et al.* (2016) studied subsampled newton methods with non-uniform sampling. Wang, *et al.* (2017) developed information-based subdata selection for large linear regression. Peng and Tan (2018a, 2018b) investigated A-optimal subsampling for Big Data linear regression and constructed fast algorithms. Liang, *et al.* (2013) proposed a resampling-based stochastic approximation for large geostatistical data. Kleiner, *et al.* (2014) gave a scalable bootstrap for massive data. Avron, *et al.* (2010) used random-sampling and random-mixing techniques to describe a fast LS solver for dense highly overdetermined systems. Drineas, *et al.* (2010) constructed randomized algorithms for faster least squares approximation. See also the monograph by Mahoney (2011) on nonuniform random subsampling for matrix based machine learning.

Fan *et al.* (2014) proposed salient features of big data such as heterogeneity, noise accumulation, spurious correlation and incidental endogeneity. Two very commonly used method to handle big data issue are Divided and Conquer and the Subsampling.

The uniform subsmapling is simple in mathematics and easy in computation, so it is frequently used, such as Monte Carlo and bootstap method. Unfortunately, the uniform sampling can not detect important observations. Ma, *et al.* (2014) conduct the leverage score based non-uniform subsampling method, this method used the estimate from a subsample taken randomly from the full sample to approximate the full sample ordinary least square estimate, they proposed BLEV,SLEV,and LEVUNW method to perform subsampling. Drineas,*et al.* (2004) proposed the non-uniform distribution to develop fast algorithms to approximate the product of two matrices, the idea is to minimize the expected squared Frobenius distance of the product and its approximate. Ma *et al.* (2015) proposed the OPT and PL subsampling method in linear regression model, they discussed the sampling probability by minimizing the trace of the intermittent part of the variance-covariance matrix of the subsampling estimator, derived asymptotic normality and performed simulations and real data analysis. Peng and Tan (2018) derived asymptotic expansions for the subsampling estimator and the asymptotic normality under appropriate conditions in linear regression model, proposed A-optimal probability distribution to estimate a smooth function of the regression coefficient, proposed data truncation for fast computing. Wang, *et al.* (2017) proposed non-uniform subsampling probabilities that minimize the asymptotic mean squared error of subsampling estimator in logistic regression, established consistency and asymptotic normality of the estimator.

This dissertation will develop the A-optimal subsampling theory for arbitrary data structure and general estimating procedures. These results are parallel to those obtained in the linear regression model in Peng and Tan (2018a). Since we are concerned with a resampling procedure, the data structure can be *arbitrary.* That is, data can be random or deterministic, dependent or independent, complete or incomplete (missing/censored/truncated), time-series data, longitudinal data, spatial correlated data, etc. We shall pursue both the algorithmic properties (i.e. how long it takes to compute the approximating subsampling estimator), and the statistical inference (i.e. under what conditions the approximating subsampling estimator is valid). We shall focus on fast algorithms, parallel computing, sequential updating and subsample size determination for the former, and on deriving A-optimal distributions, asymptotic normality, and dimension asymptotics (how growing dimensions affect the subsampling estimates) for the latter.

The rest of the thesis is organized as follows. In Chapter 2, we introduce the Count data regression and demonstrate several examples. In Chapter 3, we study the A-optimal subsampling distributions and establish the consistency and asymptotic normality theorem. We report the large simulation results in Chapter 4. In chapter 5, 6, we report the real count data analysis of the Bike Sharing data. In chapter 7, we report the real count data analysis of the Blog Feedback data.

2. COUNT DATA REGRESSION

In a count data regression model, the mean of a count response Y_i and covariate vector $\mathbf{x}_i$ satisfy

$$E(Y_i) = \mu_i(\beta) = h(\mathbf{x}_i^\top \beta), \quad i = 1, \ldots, n, \tag{2.0.1}$$

where $\beta \in \mathbb{R}^p$ is a regression parameter and h is an inverse link function. Typically, $h(t) = \exp(t)$ (the inverse log link).

2.1 Poisson Regression, Overdisperson and Negative Binomial Regression

The Poisson distribution is commonly used for modeling count data.

Example 1 Let Y has Poisson distribution with mean parameter μ, Poi(μ). Then the probability mass function of Y is given by

$$f_{\text{poi}}(y; \mu) = \exp(-\mu)\frac{\mu^y}{y!}, \quad y = 0, 1, 2, \ldots \tag{2.1.1}$$

For a Poisson random variable Y, the mean and variance are equal, $\text{Var}(Y) = \mu = E(Y)$. In real data, the equality of mean and variance is usually not met. This is termed as *overdisperson.*

When overdispersion occurs in a real data, the SE of estimates in Poisson regression model are deflated, leading to exaggerated test statistic for parameters and hence false significant findings. Overdispersion can often be tested by the usual goodness of fit statistic. In our real data analysis, we should perform such tests.

The negative binomial distribution is an option to handle overdisperson.

Example 2 Let Y have a negative binomial distribution with mean μ and overdisperson parameter $\alpha > 0$, $\mathrm{Nb}(\mu, \alpha)$. Then the probability mass function of y is given by

$$f_{\mathrm{nb}}(y;\mu,\alpha) = \frac{\Gamma(y+1/\alpha)}{\Gamma(1/\alpha)y!}(1+\alpha\mu)^{-1/\alpha}(\mu/(\mu+1/\alpha))^{-y},\ y = 0, 1, 2, \ldots \tag{2.1.2}$$

For a negative binomial random variable Y, the mean $E(Y) = \mu$ and variance $\mathrm{Var}(Y) = \mu + \alpha\mu^2$ satisfy $\mathrm{Var}(Y) \geq E(Y)$, and $\mathrm{Var}(Y) = E(Y)$ if and only if $\alpha = 0$.

Another powerful option to handle overdisperson is the *quasi-likelihood model*. This has the advantage of requiring only to specify the mean and variance but not a distribution for the response Y. Specifically, the statistical inference is based on the quasi-likelihood equation,

$$\sum_{i=1}^{n} \frac{y_i - \mu_i(\beta)}{V_i(\beta,\phi)} h'(\mathbf{x}_i^\top \beta)\mathbf{x}_i = 0, \tag{2.1.3}$$

where $\mu_i(\beta) = E(Y_i|\mathbf{x}_i)$ and $V_i(\beta,\phi) = \mathrm{Var}(Y_i|\mathbf{x}_i)$ are the mean function and variance function which are to be specified. Here ϕ is an overdisperson parameter.

The quasi-likelihood model has great flexibility and unifies several models in the sense that the maximum likelihood estimate (MLE) of the models are special cases. Setting $V_i = \mu_i$, equation (2.1.3) gives the MLE of the Poisson model. If $V_i = \mu_i(1 + \alpha\mu_i)$ with $\phi = \alpha$, then equation (2.1.3) is the estimating equation for the MLE of the negative binomial model. Another frequent choice of the variance for overdisperson is $V_i = \phi\mu_i$ with $\phi > 0$. All the three cases can be unified with $V_i = \mu_i + \alpha\mu_i^p$ for $p = 1, 2$.

2.2 Zero-inflated Poisson Regression

In many real count data, there is an excess of zero counts for which the Poisson distribution can not account. Consider a mixture model combining a degenerate distribution at 0 and a Poisson distribution defined by

$$f_{\mathrm{zip}}(y;\mu,\rho) = \rho f_0(y) + (1-\rho) f_{\mathrm{poi}}(y;\mu), \quad y = 0, 1, 2, \ldots, \tag{2.2.1}$$

where $f_0(y) = \mathbf{1}[y = 0]$ is the point mass at zero (the degenerate distribution at zereo) to account for structural zeros. Since

$$f_{\mathrm{zip}}(0; \mu, \rho) = \rho + (1 - \rho) \exp(-\mu),$$

it thus follows from $0 \leq f_{\mathrm{zip}}(0; \mu, \rho) \leq 1$ that $1/(1 - \exp(\mu)) \leq \rho \leq 1$. This shows that ρ can be negative. A positive ρ represents the probability of structural zeros above the amount of zeros expected under the Poisson distribution f_{poi}. A negative ρ means that the amount of zeros is below the expected under Poisson, and this does not occur very often. The MLE $\hat{\beta}_n$ can be obtained by solving the score equation

$$\sum_{i=1}^{n} \frac{f_{\mathrm{poi}}(y_i; \mu_i)}{f_{\mathrm{zip}}(y_i; \mu_i, \rho)} \frac{y_i - \mu_i(\beta)}{\mu_i(\beta)} h'(\mathbf{x}_i^\top \beta)\mathbf{x}_i = 0. \tag{2.2.2}$$

To estimate ρ, one can obtain another equation differentiating the log likelihood with respect to ρ. For simplicity, we shall estimate ρ by the sample percentage $\hat{\rho}$ of structural zeros. Substituting $\hat{\rho}$ in (2.2.2), we solve for $\hat{\beta}_n$.

2.3 Truncated Models

Suppose realizations of a count random variable Y less than a positive integer l are omitted. Then the resulting distribution is called *left-truncated*. For simplicity, only left-truncation is considered and right-truncation is similar. Let Y has pmf $g(y; \theta)$ and cdf $G(y; \theta)$ with parameter θ. Then left-truncated count distribution is given by

$$f(y; \theta | y \geq l) = \frac{g(y; \theta)}{\bar{G}(l - 1; \theta)}, \quad y = l, l + 1, \ldots, \tag{2.3.1}$$

where $\bar{G} = 1 - G$ is the survival function.

Choosing g to be the pmf of the negative binomial, the left-truncated negative binomial can be obtained. As a limiting case of this, the left truncated distribution of Poisson $\mathrm{Poi}(\mu)$ can be obtained as follows:

$$f(y; \mu | y \geq l) = \frac{\mu^y}{(\exp(\mu) - \sum_{i=1}^{l-1} \mu^i / i!) y!}, \quad y = l, l + 1, \ldots. \tag{2.3.2}$$

This has the mean $E(Y) = \mu + \delta$ and variance $\mathrm{Var}(Y) = \mu - \delta(\mu - l)$, where

$$\delta = \frac{f_{\text{poi}}(l-1;\mu)}{\bar{F}_{\text{poi}}(l-1;\mu)}\mu. \tag{2.3.3}$$

These exhibit that the mean of the left-truncated random variable is bigger than the corresponding mean of the un-truncated distribution, whereas the truncated variance is smaller. The MLE of $\hat{\beta}_n$ can be obtained as the solution of the following equation

$$\sum_{i=1}^{n} \frac{y_i - \mu_i(\beta) - \delta_i(\beta)}{\mu_i(\beta)} h'(\mathbf{x}_i^\top \beta)\mathbf{x}_i = 0. \tag{2.3.4}$$

2.4 Censored Counts

Censoring of count observations may arise from aggregation or from the resulting samples in which high counts are not observed. Health data and social media data are examples. Consider a latent count variable Z that is censored from above at point c (right censoring) and covariate variable $\mathbf{x}$. Let $Y = Z$ if $Z \leq c$. Suppose Z satisfies the regression model

$$Z = \mu(\mathbf{x};\beta) + \varepsilon, \tag{2.4.1}$$

where ε is a random error with mean $E(\varepsilon) = 0$. Suppose there are available independent observations $(Y_i, d_i, \mathbf{x}_i), i = 1, \ldots, n$, where $d_i = \mathbf{1}[Z_i \leq c]$ is the censoring indicator, and $Y_i = Z_i$ if $\delta_i = 1$. Suppose Y has pmf $g(y;\theta)$ and cdf $G(y;\theta)$. The log-likelihood function for the independent observations are

$$\ell_n(\beta) = \sum_{i=1}^{n} d_i \log g(Y_i;\theta(\beta)) + (1-d_i)\log \bar{G}(c-1;\theta(\beta)). \tag{2.4.2}$$

For the right censored Poisson model, the maximum likelihood estimating equation is given by

$$\sum_{i=1}^{n} \frac{d_i(Y_i - \mu_i(\beta)) + (1-d_i)\delta_i(\beta)}{\mu_i(\beta)} h'(\mathbf{x}_i^\top \beta)\mathbf{x}_i = 0. \tag{2.4.3}$$

where $\delta_i(\beta)$ is the adjustment factor associated with the left-truncated Poisson model, given by

$$\delta_i(\beta) = \frac{f_{\text{poi}}(c-1;\mu_i(\beta))}{\bar{F}_{\text{poi}}(c-1;\mu_i(\beta))}\mu_i(\beta). \tag{2.4.4}$$

3. A-OPTIMAL SAMPLING DISTRIBUTIONS AND ASYMPTOTIC THEORY

Let $\{Z_{ni} : 1 \leq i \leq n,\, n \geq 1\}$ be a sequence of random variables defined on some probability space $(\Omega, \mathbb{P})$ and $\beta \in \mathcal{B} \subset \mathbb{R}^p$ be a parameter vector. Consider a triangular array of smooth functions $\{\psi_{ni}(Z_{ni}; \beta)\colon\, 1 \leq i \leq n,\, n \geq 1\}$ taking values in $\mathbb{R}^p$ with each $\mathbb{E}(\psi_{ni}(Z_{ni}; \beta_0)) = 0$ for a unique $\beta_0 \in \mathcal{B}$. We estimate β_0 by $\hat{\beta}_n$ which solves the estimating equations,

$$\Psi_n(\beta) = \sum_{i=1}^{n} \psi_{ni}(Z_{ni}; \beta) = 0. \tag{3.0.1}$$

Following Chatterjee and Bose (2005), we assume $\{(\psi_{ni}(Z_{ni}; \beta_0), \mathscr{F}_i), i = 1, \ldots, n, n \geq 1\}$ forms a martingale difference, i.e., $\mathbb{E}(\psi_{ni}(Z_{ni}; \beta_0)|\mathscr{F}_{i-1}) = 0$, where $\{\mathscr{F}_i, i = 1, 2, \ldots,\}$ is an increasing sequence of sigma-algebras, see Chapter 5 of Borovskikh and Korolyuk (1974).

We consider the case that the sample size n is *extremely large* and the estimate $\hat{\beta}_n$ is not available or time-consuming to obtain it. Our approach to tackling this big data estimation problem is *A-optimal subsampling*, that is, we seek the A-optimal sampling distribution on the data points and use it take a subsample as a surrogate of the whole sample.

Let $\pi_n = (\pi_{ni}, i = 1, \ldots, n)$ be a sampling distribution on the n data points Z_{ni}. We use it to take a subsample $Z^* = \{Z_j^* : j = 1, \ldots, r\}$ with the subsample size $r << n$. Let $\pi^* = (\pi_j^* : j = 1, \ldots, r)$ be the corresponding sampling probabilities. We now approximate the estimate $\hat{\beta}_n$ by the subsampling generalized bootstrap estimate $\hat{\beta}^*_{r_n}$ which solves the estimating equations

$$\Psi_r^*(\beta) =: \sum_{j=1}^{r} \frac{\psi_{nj}(Z_{nj}^*; \beta)}{\pi_j^*} = 0. \tag{3.0.2}$$

The theory of weighted (generalized) bootstrap has been extensively studied in the literature, see e.g. Mammen (1993) and Chatterjee and Bose (2002). However, the choices in existing weights are limited; most of them are exchangeable non-negative random variables that are independent of data; and only some of them can improve Efron's bootstrap using tedious Edgeworth expansions. See Chapter II of the monograph by Barbe and Bertail (1995) and the references therein. Unlike existing weights, we shall allow the weights to depend on the data. In fact, we shall derive numerous weights by minimizing the trace of certain variance-covariance. They are referred to as the A-optimal weights which are different from existing weights: they are data driven so dependent of the data and not exchangeable.

3.1 A Theorem from Chung, Tan and Peng (2018)

NOTATION. Abbreviate $\psi_{ni}(\beta) = \psi_{ni}(Z_{ni};\beta)$, its d-th component $\psi_{ni,d}(\beta)$, and $\psi_{ni} = \psi_{ni}(\beta_0)$. Let $\dot\psi_{ni}(\beta) = \partial/\partial\beta\psi_{ni}(\beta) \in \mathbb{R}^p$ and $\ddot\psi_{ni}(\beta) = \partial/\partial\beta^\top\dot\psi_{ni}(\beta)$ ($p \times p$ matrix) be the first and second partial derivatives with respect to parameter β. For matrix A, denote $A^\top$ the transpose of A, $A^{\otimes 2} = AA^\top$, $A^{-\top} = (A^{-1})^\top$, $\mathbb{E}^{-1}(A) = (\mathbb{E}(A))^{-1}$, and $A^{(s)} = 1/2(A + A^\top)$. Write $\|A\|$ the euclidean norm, $\|A\|_o$ the spectral norm, $\lambda_{\max}(A)$ ($\lambda_{\mathrm{amin}}(A)$) the maximum (minimum absolute) eigenvalue of A, etc.

To quote a theorem from Chung, Tan and Peng (2018), we introduce the following assumptions. Let

$$J_n(\beta) = \sum_{i=1}^n \pi_{ni}^{-1}\psi_{ni}(\beta)^{\otimes 2},\ \lambda_n = \lambda_{\max}^{1/2}(J_n(\hat\beta_n)),\ \Sigma_n = \dot\Psi_n^{-1}J_n\dot\Psi_n^{-\top}\big|_{\hat\beta_n}. \tag{3.1.1}$$

Let $\delta_n > 0$ be an arbitrary sequence. Typically, $\delta_n = \min(\pi_{ni}, i = 1, \ldots, n)$.

(R1) $$\delta_n\lambda_n^2 \xrightarrow{p} \infty, \quad \mathbb{P}(\delta_n^{-1}\lambda_n^{-2}\lambda_{\mathrm{amin}}(\dot\Psi_n^{(s)}(\hat\beta_n)) > 0) \to 1.$$

(R2) Each component $\psi_{ni,d}(\beta)$ admits the second order expansion

$$\psi_{ni,d}(\beta_0 + t) = \psi_{ni,d}(\beta_0) + \dot\psi_{ni,d}^\top(\beta_0)t + 1/2t^\top\ddot\psi_{ni,d}(\tilde\beta_{ni,d})t, \quad d = 1, \ldots, p,$$

for $\|t\| \le t_0$ with some $t_0 > 0$, where $\tilde\beta_{ni,d}$ lies in between β_0 and $\beta_0 + t$.

(R3) The sampling probabilities π_{ni} and subsample size r_n satisfy

$$\sum_{i=1}^{n} \pi_{ni}^{-1} \|\dot{\psi}_{ni}(\hat{\beta}_n)\|^2 = o_p(p_n^{-1} r_n \delta_n^2 \lambda_n^4).$$

(R4) There exists a neighborhood $\mathbb{N}_0$ of β_0 such that $\ddot{\Psi}_{n,d}(\beta)$ is either positive or negative definite in $\mathbb{N}_0$ and that there is a rv $\eta_{ni,d}$

$$\sup_{\beta \in \mathbb{N}_0} \lambda_{\mathrm{amax}}(\ddot{\Psi}_{n,d}(\beta)) \leq \eta_{ni,d}, \quad d = 1, \ldots, p,$$

where the random vector $\eta_{ni} = (\eta_{ni,1}, \ldots, \eta_{ni,p})^\top$ satisfies

$$\sum_{i=1}^{n} (n + (r_n \pi_{ni})^{-1}) \|\eta_{ni}\|^2 = o_p(p_n^{-2} r_n \delta_n^4 \lambda_n^6).$$

(R5) $$\lambda_{\max}(J_n(\hat{\beta}_n))/\lambda_{\min}(J_n(\hat{\beta}_n)) = O_p(1).$$

(R6) Fix $u \in \mathbb{R}^{p_n}$ with $\|u\| = 1$. The double array $z_{nj}^* = s_n^{-1} u^\top \dot{\Psi}_n^{-\top}(\hat{\beta}_n) \psi_{nj}^*(\hat{\beta}_n)/\pi_{nj}^*$, $j = 1, 2, \ldots, r,\ r \geq 1$ satisfies the Lindeberg condition: for every $t > 0$,

$$\sum_{i=1}^{n} \pi_{ni} \|z_{n,i}\|^2 \mathbf{1}[\|z_{ni}\| \geq \sqrt{r} t] = o_p(1), \quad as \quad r \to \infty,$$

where $s_n^2 = u^\top \Sigma_n u$.

We quote the following theorem from Chung, Tan and Peng (2018).

Theorem 3.1.1 *Suppose (R1)–(R5) hold. Assume $\hat{\beta}_n$ is a solution of (3.0.1) such that $\hat{\beta}_n = \beta_0 + o_p(1)$. Assume*

$$\sum_{i=1}^{n} \pi_{ni}^{-1} \|\psi_{ni}(\hat{\beta}_n)\|^2 = O_p(p_n \lambda_n^2). \tag{3.1.2}$$

Then these exists a sequence of solutions $\hat{\beta}_{r_n}^*$ of (3.0.2) such that if $p_n/(r_n \delta_n^2 \lambda_n^2) = o_p(1)$, then

$$\dot{\Psi}_n(\hat{\beta}_n) \sqrt{r_n} (\hat{\beta}_{r_n}^* - \hat{\beta}_n) = -\frac{1}{\sqrt{r_n}} \sum_{j=1}^{r_n} \frac{\psi_{nj}^*(\hat{\beta}_n)}{\pi_j^*} + o_p(\lambda_n). \tag{3.1.3}$$

If, further, (R5)–(R6) are satisfied for $u \in \mathbb{R}^{r_n}$ with $\|u\| = 1$, then

$$s_n^{-1} \sqrt{r_n} u^\top (\hat{\beta}_{r_n}^* - \hat{\beta}_n) \Rightarrow \mathscr{N}(0, 1), \quad in\ probability, \quad r \to \infty. \tag{3.1.4}$$

3.2 A Second Theorem from Chung, Tan and Peng (2018)

Let

$$J_{1n}(\beta) = \sum_{i=1}^{n} \mathbb{E}(\psi_{ni}(\beta)^{\otimes 2}), \quad \lambda_{1n} = \lambda_{\max}^{1/2}(J_{1n}).$$

(R1') $\lambda_{1n} \to \infty, \quad \inf_{n \geq n_0}\{\lambda_{1n}^{-2}\lambda_{\text{amin}}(\mathbb{E}(\dot{\Psi}_n^{(s)}))\} > 0.$

(R3') $\sum_{i=1}^{n} \mathbb{E}\big(\|\dot{\psi}_{ni} - \mathbb{E}(\dot{\psi}_{ni})\|^2\big) = o(p_n^{-1}\lambda_{1n}^4).$

(R4') Same as (R4) except that η_{ni} are replaced with η_{1ni} which satisfy

$$\sum_{i=1}^{n} \|\eta_{1ni}\|^2 = o_p(n^{-1}p_n^{-2}\lambda_n^6).$$

(R5') $\lambda_{\max}(J_{1n})/\lambda_{\min}(J_{1n}) = O(1).$

(R6') Fix $u \in \mathbb{R}^{p_n}$ with $\|u\| = 1$. Let $s_{1n}^2 = u^\top \mathbb{E}^{-1}(\dot{\Psi}_n) \sum_{i=1}^{n} \psi_{ni}^{\otimes 2} \mathbb{E}^{-\top}(\dot{\Psi}_n)u$. The double array $z_{1ni} = s_{1n}^{-1}u^\top \mathbb{E}^{-1}(\dot{\Psi}_n)\psi_{ni}$, $i = 1, 2, \ldots, n$, $n \geq 1$ satisfies

$$\sum_{i=1}^{n} \|z_{1ni}\|^2 = o_p(1), \quad \mathbb{E}(\max_i \|z_{1ni}\|) = o(1).$$

for every $t > 0$.

We quote the following theorem from Chung, Tan ane Peng (2018), which describes the asymptotic behaviors of the M-estimator for both fixed and growing parameter dimension.

Theorem 3.2.1 *Suppose (R1'), (R2), (R3')–(R5') hold. Then these exists a sequence of solutions $\hat{\beta}_n$ of (3.0.1) such that if $p_n/\lambda_{1n}^2 = o(1)$, then*

$$p_n^{-1/2}\lambda_{1n}(\hat{\beta}_n - \beta_0) = O_p(1), \tag{3.2.1}$$

$$\lambda_{1n}^{-1}\mathbb{E}(\dot{\Psi}_n)(\hat{\beta}_n - \beta_0) = -\lambda_{1n}^{-1}\sum_{i=1}^{n}\psi_{ni} + o_p(1). \tag{3.2.2}$$

If, further, (R5')–(R6') are satisfied for $u \in \mathbb{R}^p$ with $\|u\| = 1$, then

$$s_{1n}^{-1}u^\top(\hat{\beta}_n - \beta_0) \Rightarrow \mathscr{N}(0, 1), \quad \textit{in probability.} \tag{3.2.3}$$

3.3 The A-optimal Sampling Distribution

In view of Theorem 3.1.1 and (3.1.1), we have

$$\mathrm{Var}^*(\hat{\beta}^*_{r_n}) = \frac{1}{r}\Sigma_n + o_p(1) = \frac{1}{r}\sum_{i=1}^{n}\frac{1}{\pi_i}\dot{\Psi}_n^{-1}\psi_{ni}\psi_{ni}^\top\dot{\Psi}_n^{-\top}|_{\hat{\beta}_n} + o_p(1). \tag{3.3.1}$$

As Σ_n is a function of the sampling distribution $\pi = (\pi_1, \ldots, \pi_n)$ on the data points, we seek a sampling distribution which minimizes the trace of the matrix Σ_n. Following Peng and Tan (2018), we write

$$\tau(\pi) =: \mathrm{Tr}\big(\Sigma_n\big) = \sum_{i=1}^{n}\frac{\|a_{ni}\|^2}{\pi_i}, \quad \pi \in \mathscr{P}_n,$$

where $a_{ni} = \dot{\Psi}_n^{-1}\psi_{ni}|_{\hat{\beta}_n}$, and $\mathscr{P}_n$ is the probability simplex $\mathscr{P}_n = \{\pi : \pi_i \geq 0, \sum_i \pi_i = 1\}$ in $\mathbb{R}^n$. Using Lagrange multipliers, we readily derive the minimizer which is stated in the following theorem. As usual, the minimizer is referred to as *A-optimal*. Equivalently, an A-optimal distribution minimizes the sum of the component variances of the subsampling estimator $\hat{\beta}^*_{r_n}$. Let

$$\hat{H}_k = A_n(\dot{\Psi}_n^\top\dot{\Psi}_n)^{-k/2}A_n^\top|_{\hat{\beta}_n}, \quad k = 0, 1, 2. \tag{3.3.2}$$

where $A_n(\beta) = (\psi_{n1}(\beta), \ldots, \psi_{nn}(\beta))^\top$. The following theorem is quoted from Chung, Tan and Peng (2018).

Theorem 3.3.1 *Suppose $\dot{\Psi}_n(\hat{\beta}_n)$ is invertible. Then the square roots of the diagonal entries of $\hat{H}_2$ gives an (asymptotically) A-optimal distribution $\hat{\pi}$ on the data points for $\hat{\beta}^*_{r_n}$ to approximate $\hat{\beta}_n$. Suppose, further, $\psi_{ni}(\hat{\beta}_n) \neq 0$ for $i = 1, \ldots, n$. Then $\hat{\pi}$ is unique.*

Specifically, the sampling probabilities are given by

$$\hat{\pi}_i \propto \|a_{ni}\| = (\psi_{ni}^\top(\dot{\Psi}_n^\top\dot{\Psi}_n)^{-1}\psi_{ni})^{1/2}|_{\hat{\beta}_n}, \quad i = 1, \ldots, n, \tag{3.3.3}$$

where $\pi_i \propto a_i$ denotes $\pi_i = a_i / \sum_{i=1}^n a_i$ for $a_i \geq 0, i = 1, \ldots, n$.

Remark 3.3.1 For conditions (R3)–(R4) and (R6) to hold, the sampling probabilities must be bounded away from zero, which is not required for the other conditions. As a result, our discussion below shall involve truncation only in (R3)-(R4) and (R6).

3.4 Asymptotic Behaviors under A-optimal Sampling for Fixed p

Let $l_{ni}, i = 1, 2, \dots, n, n \geq 1$ be a double array of positive numbers. Like in Peng and Tan (2018), we truncate $\hat{\pi}$ from below by $l_n = (l_{ni}/n)$ as follows:

$$\hat{\pi}_{ni}^{(l_n)} \propto \hat{\pi}_{ni}\mathbf{1}[\hat{\pi}_{ni} \geq l_{ni}/n] + l_{ni}/n\mathbf{1}[\hat{\pi}_{ni} < l_{ni}/n], \quad i = 1, \dots, n. \tag{3.4.1}$$

Though, typically, we require $l_{ni} \geq l_0 > 0$ for some l_0, we shall investigate conditions to allow for $l_{ni} \to 0$ as n tends to infinity.

(R11) There is some constant $c_0 > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \|\psi_{ni}(\beta_0)\| = c_0 + o_p(1).$$

(R12) There is a constant matrix $\dot{\Psi}_0$ with $\lambda_{\text{amin}}(\dot{\Psi}_0) > 0$ such that

$$\frac{1}{n}\dot{\Psi}_n = \frac{1}{n}\sum_{i=1}^{n} \dot{\psi}_{ni}(\beta_0) = \dot{\Psi}_0 + o_p(1).$$

(R13) There is a positive definite matrix A_0 such that

$$\delta_n \sum_{i=1}^{n} \frac{\psi_{ni}^{\otimes 2}}{\|\psi_{ni}\|} = A_0 + o_p(1).$$

(R31) There is a positive sequence of $l_n = (l_{ni} : i = 1, \dots, n)$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\|\dot{\psi}_{ni}\|^2}{\|\psi_{ni}\|}\mathbf{1}[\|\psi_{ni}\| \geq l_{ni}] = o_p(r_n).$$

(R41) There exists a neighborhood $\mathbb{N}_0$ of β_0 such that $\ddot{\Psi}_{n,d}(\beta)$ is either positive or negative definite in $\mathbb{N}_0$ and that there is a rv $\eta_{ni,d}$

$$\sup_{\beta \in \mathbb{N}_0} \lambda_{\text{amax}}(\ddot{\Psi}_{n,d}(\beta)) \leq \eta_{ni,d}, \quad d = 1, \dots, p,$$

where the random vector $\eta_{ni} = (\eta_{ni,1}, \dots, \eta_{ni,p})^\top$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \big(1 + \frac{\mathbf{1}[\|\psi_{ni}\| \geq l_{ni}]}{r_n\|\psi_{ni}\|}\big)\|\eta_{ni}\|^2 = o_p(r_n n \delta_n).$$

(R61) The double array $z_{nj}^{(l_n)*} = \Sigma_n^{-1/2}\dot{\Psi}_n^{-\top}(\hat{\beta}_n)\psi_{nj}^*(\hat{\beta}_n)/\hat{\pi}_{nj}^{(l_n)*}$, $j = 1, 2, \ldots, r$, $r \geq 1$ satisfies the Lindeberg condition: for every $t > 0$,

$$\sum_{i=1}^{n} \hat{\pi}_{ni}^{(l_n)} \|z_{n,i}^{(l_n)}\|^2 \mathbf{1}[\|z_{ni}\| \geq \sqrt{r}t] = o_p(1), \quad as \quad r \to \infty.$$

Theorem 3.4.1 *Suppose (R11)-(R13), (R2), (R31)-(R41) and (R4') hold. Assume $\hat{\beta}_n$ is a solution of (3.0.1) such that $\hat{\beta}_n = \beta_0 + o_p(1)$. Then these exists a sequence of solutions $\hat{\beta}_{r_n}^*$ of (3.0.2) such that*

$$\dot{\Psi}_n(\hat{\beta}_n)\sqrt{r_n}(\hat{\beta}_{r_n}^* - \hat{\beta}_n) = -\frac{1}{\sqrt{r_n}} \sum_{j=1}^{r_n} \frac{\psi_{nj}^*(\hat{\beta}_n)}{\pi_j^*} + o_p(\hat{\lambda}_n). \tag{3.4.2}$$

If, further, (R61) hold for the truncated sampling distribution in (3.4.1), then

$$V_n^{-1/2}\sqrt{r_n}(\hat{\beta}_{r_n}^* - \hat{\beta}_n) \Rightarrow \mathscr{N}(0, 1), \quad \textit{in probability}, \quad r_n \to \infty. \tag{3.4.3}$$

where V_n equals Σ_n in (3.1.1) under the truncated sampling distribution (3.4.1).

Proof of Theorem 3.4.1. We shall verify the conditions of Theorem 3.1.1 for the case of fixed dimenion $p_n = p$. In this case, (R31)-(R41) and (R61) imply (R3)-(R4) and (R6), respectively. Let $\hat{\psi}_{ni} = \psi_{ni}(\hat{\beta}_n)$. By (R2), (R12), (R41) and (R4'),

$$\frac{1}{n}\sum_{i=1}^{n}(\|\hat{\psi}_{ni}\| - \|\psi_{ni}\|) = o_p(1). \tag{3.4.4}$$

This and (R11) yield

$$\frac{1}{n}\sum_{i=1}^{n}\|\hat{\psi}_{ni}\| = c_0 + o_p(1). \tag{3.4.5}$$

By (R4') again,

$$\frac{1}{n}\dot{\Psi}_n(\hat{\beta}_n) - \frac{1}{n}\dot{\Psi}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\dot{\psi}_{ni}(\hat{\beta}_n) - \dot{\psi}_{ni}\right) = o_p(1). \tag{3.4.6}$$

This and (R12) give

$$\frac{1}{n^2}\left(\dot{\Psi}_n^\top \dot{\Psi}_n|_{\hat{\beta}_n}\right)^{-1} = \left(\dot{\Psi}_0^\top \dot{\Psi}_0\right)^{-1} + o_p(1). \tag{3.4.7}$$

Thus there exist constants $0 < b_0 \leq B_0 < \infty$ such that

$$b_0\|\hat{\psi}_{ni}\|/n \leq \hat{\pi}_i \leq B_0\|\hat{\psi}_{ni}\|/n, \quad i = 1, \ldots, n. \tag{3.4.8}$$

Let us write $J_n(\beta) = J_n(\beta, \pi)$ and $\hat{J}_n = J_n(\hat{\beta}_n, \hat{\pi})$. Then by (R13) and (3.4.7),

$$\delta_n \hat{J}_n = \sum_{i=1}^{n} \|\hat{\psi}_{ni}\| \delta_n \sum_{i=1}^{n} \frac{\hat{\psi}_{ni}^{\otimes 2}}{\|\hat{\psi}_{ni}\|} = n(A_0 c_0 + o_p(1)). \tag{3.4.9}$$

Thus $\delta_n \hat{\lambda}_n^2 = \delta_n \lambda_{\max}(\hat{J}_n) = c_1(n + o_p(1))$ for some constant $c_1 > 0$. Consequently, (R5) holds; (R12) and (3.4.6) imply (R1); (3.4.5) yields (3.1.2). We now apply Theorem 3.1.1 to finish the proof. ■

3.4.1 Asymptotics for Generalized Count Regression

Consider the estimating equation (2.1.3),

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i(\beta)}{V_i(\beta, \phi)} h'(X_i^\top \beta) X_i = 0, \tag{3.4.10}$$

where $\mu_i(\beta) = E(Y_i|X_i)$, $V_i(\beta, \phi) = \mathrm{Var}(Y_i|X_i)$, and ϕ is an overdisperson parameter. Typically, $V_i(\beta, \phi) = V(\mu_i, \phi)$ for some positive variance function V. Here we assume $V(\mu, \alpha) = \mu + \alpha\mu^p$ for $\alpha \geq 0$ and $p = 1, 2$. This cover Poisson, overdispersed Poisson and Negative Binomial distributions. We also consider the log link link, so $h(t) = \exp(t)$. The parameter estimator $\hat{\beta}_n$ is the solution to the above equation. This $\hat{\beta}_n$ can be approximated by the subsampling estimator $\hat{\beta}^*_{r_n}$, which solves the equation

$$\sum_{j=1}^{r} \frac{y_j^* - \mu_j^*(\beta)}{\pi_j^* V_j^*(\beta, \phi)} h'(X_j^{*\top} \beta) X_j^* = 0, \tag{3.4.11}$$

where $\mu_j^*(\beta) = h(X_j^* \beta)$ and $V_j^*(\beta, \phi) = V_j(\mu_j^*(\beta), \phi)$. For the canonical link, $\dot{\Psi}_n(\beta) = -\sum_{i=1}^n \exp(X_i^\top \beta) X_i^{\otimes 2}$. In this case, an approximation to the sampling probabilities is given by

$$\bar{\pi}_i \propto \left(X_i^\top \dot{\Psi}_n^{-2}(\hat{\beta}_n) X_i)\right)^{1/2} \exp(1/2 X_i^\top \hat{\beta}_n), \quad i = 1, \ldots, n. \tag{3.4.12}$$

Our simulation results show that the $\bar{A}$-optimal sampling distribution $\bar{\pi}$ can substantially improve the uniform and the leverage sampling.

One calculates $\lambda_{1n} = \lambda_{\max}^{1/2}(J_{1n})$ and $V_{1n}^2 = \mathbb{E}^{-1}(\dot{\Psi}_n)J_{1n}\mathbb{E}^{-\top}(\dot{\Psi}_n)$, where

$$J_{1n} = \sum_{i=1}^n \mathbb{E}(\psi_{ni}^{\otimes 2}) = \sum_{i=1}^n \frac{h'(X_i^\top \beta_0)^2}{V_i(\beta_0, \alpha)} X_i^{\otimes 2}, \quad \mathbb{E}(\dot{\Psi}_n) = -\sum_{i=1}^n \frac{\mu_i}{1+\alpha\mu_i^{p-1}} X_i^{\otimes 2}.$$

Also $\hat{\lambda}_n = \lambda_{\max}^{1/2}(\hat{J}_n)$ and $\Sigma_n = \dot{\Psi}_n^{-1} J_n \dot{\Psi}_n^{-\top}\big|_{\hat{\beta}_n}$, where

$$\hat{J}_n = J_n(\hat{\beta}_n, \hat{\pi}) = \sum_{i=1}^n \hat{\pi}_{ni}^{-1}\psi_{ni}(\hat{\beta}_n)^{\otimes 2} = \sum_{i=1}^n \|\dot{\Psi}_n^{-1}\psi_{ni}\| \sum_{i=1}^n \frac{\psi_{ni}^{\otimes 2}}{\|\psi_{ni}\|}\Big|_{\hat{\beta}_n}. \tag{3.4.13}$$

Suppose there exist $c_0 > 0$ and positive definite matrices $\dot{\Psi}_0, A_0$ such that

$$\frac{1}{n}\sum_{i=1}^n \frac{|Y_i - \mu_i|}{1+\alpha\mu_i^{p-1}}\|X_i\| = c_0 + o_p(1), \tag{3.4.14}$$

$$\frac{1}{n}\sum_{i=1}^n \frac{\mu_i}{1+\alpha\mu_i^{p-1}} X_i^{\otimes 2} = -\dot{\Psi}_0 + o_p(1), \tag{3.4.15}$$

$$\delta_n \sum_{i=1}^n \frac{|Y_i - \mu_i|}{1+\alpha\mu_i^{p-1}} \frac{X_i^{\otimes 2}}{\|X_i\|} = A_0 + o_p(1). \tag{3.4.16}$$

Theorem 3.4.2 *Suppose the $n \times p$ matrix $(X_1, \ldots, X_n)^\top$ has full rank. Suppose $n^{-1}J_{1n} = J_{10} + o(1)$ for some positive definite matrix J_{10}. Assume* $\inf_n \lambda_{\text{amin}}\big(n^{-1}\mathbb{E}(\dot{\Psi}_n)\big) > 0$. *If (3.4.15) and (R4') hold, then these exists a sequence of solutions $\hat{\beta}_n$ of (3.0.1) such that*

$$\hat{\beta}_n = \beta_0 - \mathbb{E}^{-1}(\dot{\Psi}_n)\sum_{i=1}^n \psi_{ni} + o_p(n^{-1/2}). \tag{3.4.17}$$

Assume, further, the double array $z_{1ni} = s_{1n}^{-1}\mathbb{E}^{-1}(\dot{\Psi}_n)\psi_{ni}$, $i = 1, 2, \ldots, n$, $n \geq 1$ satisfies

$$\sum_{i=1}^n \|z_{1ni}\|^2 = o_p(1), \quad \mathbb{E}(\max_i \|z_{1ni}\|) = o(1). \tag{3.4.18}$$

Then

$$V_{1n}^{-1}(\hat{\beta}_n - \beta_0) \Rightarrow \mathcal{N}(0,1), \quad \textit{in probability.} \tag{3.4.19}$$

Furthermore, suppose (3.4.14), (3.4.16) and (R31)-(R41) hold. Then there exists a sequence of solutions $\hat{\beta}_{r_n}^*$ of (3.0.2) to approximate $\hat{\beta}_n$ such that

$$\dot{\Psi}_n(\hat{\beta}_n)\sqrt{r_n}(\hat{\beta}_{r_n}^* - \hat{\beta}_n) = -\frac{1}{\sqrt{r_n}}\sum_{j=1}^{r_n} \frac{\psi_{nj}^*(\hat{\beta}_n)}{\pi_j^*} + o_p(\hat{\lambda}_n). \tag{3.4.20}$$

If, additionally, (R61) holds, then

$$V_n^{-1}\sqrt{r_n}(\hat{\beta}^*_{r_n} - \hat{\beta}_n) \Rightarrow \mathscr{N}(0,1), \quad \textit{in probability}, \quad r_n \to \infty. \tag{3.4.21}$$

where V_n is given in Theorem 3.4.1.

Proof of Theorem 3.4.2. We apply Theorem 3.2.1 to prove (3.4.17). In fact, by assumptions, $\lambda_{1n} = O(n^{1/2})$ and hence (R1') and (R5') hold, whereas (3.4.15) implies (R3'). Applying (3.2.1) proves (3.4.17), while (3.4.19) follows from (3.2.3). We now apply Theorem 3.4.1 to finish the proof. ∎

4. SIMULATION STUDY

In this chapter, we use simulation studies to evaluate the A-optimal subsampling approach proposed in previous sections. The design matrix $\mathbf{X}$ is generated from one of the four following multivariate distributions. (1) Gaussian distribution $N(0,\Sigma)$, $\Sigma_{i,j} = 0.3^{|i-j|}$. (2) Mixture Gaussian distribution with $\frac{1}{2}N(0,\Sigma)+\frac{1}{2}N(0,3\Sigma)$. (3) Log-normal distribution $LN(0,\frac{1}{2}\Sigma)$. (4) The t distribution with 5 degree of freedom $\mathbf{T}_5(0,\frac{1}{2}\Sigma)$. We choose $n = 50,000$, $p = 50$, $\boldsymbol{\beta} = (0.1, -0.1\times\mathbf{1}^\top_{(p/2)}, 0.1\times\mathbf{1}^\top_{(p/2)})$. We consider the response y_i from Poisson distribution and Negative Binomial distribution with variance structure $V(y_i) = \mu_i+5\mu_i^2$. We use logarithm link in the above two situations: $\log(\mu_i) = \mathbf{x}_i^\top\boldsymbol{\beta}, \quad i = 1,\cdots,n$. The subsampling probabilities are calculated from the following formulas:

$$\begin{aligned}
\hat{\pi}_i^{(2)} &= \frac{\|(\mathbf{X}^\top\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n\|(\mathbf{X}^\top\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|},\\
\hat{\pi}_i^{(1)} &= \frac{\|(\mathbf{X}^\top\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n\|(\mathbf{X}^\top\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\||\hat{e}_i|},\\
\hat{\pi}_i^{(0)} &= \frac{\|\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n\|\mathbf{x}_i\||\hat{e}_i|},\\
\bar{\pi}_i^{(2)} &= \frac{\|(\mathbf{X}^\top\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n\|(\mathbf{X}^\top\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i},\\
\bar{\pi}_i^{(1)} &= \frac{\|(\mathbf{X}^\top\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n\|(\mathbf{X}^\top\mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\|\hat{g}_i},\\
\bar{\pi}_i^{(0)} &= \frac{\|\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n\|\mathbf{x}_i\|\hat{g}_i}, \quad i = 1,\cdots,n.
\end{aligned}$$

For Poisson regression,

$$W(\hat{\boldsymbol{\beta}}) = \mathrm{Diag}(\hat{\mu}_i), \quad \hat{\mu}_i = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}),$$
$$\hat{e}_i = y_i - \hat{\mu}_i, \quad \hat{g}_i = \sqrt{\hat{\mu}_i} \quad i = 1,\cdots,n.$$

For Negative Binomial distribution,

$$W(\hat{\boldsymbol{\beta}}) = \text{Diag}(\frac{\hat{\mu}_i}{1+\alpha\hat{\mu}_i}), \quad \hat{\mu}_i = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}}),$$

$$\hat{e}_i = \frac{y_i - \hat{\mu}_i}{1+\alpha\hat{\mu}_i}, \quad \hat{g}_i = \sqrt{\frac{\hat{\mu}_i}{1+\alpha\hat{\mu}_i}} \quad i = 1, \cdots, n.$$

We take subsamples of size r according to the sampling distributions calculated from the above step, and obtain the estimate $\hat{\boldsymbol{\beta}}^*_r$. We calculate the empirical mean square errors at different subsample sizes r for each of $B = 1000$ subsamples using the following formula:

$$MSE = \frac{1}{B}\sum_{b=1}^{B} \|\hat{\boldsymbol{\beta}}^*_{r,b} - \hat{\boldsymbol{\beta}}\|^2,$$

where $\hat{\boldsymbol{\beta}}^*_{r,b}$ is the estimate from the b^{th} subsample with subsample size r.

For the purpose of demonstrating the effects of different design matrix $\mathbf{X}$ on the subsampling probabilities, we plot the boxplots of the subsampling probabilities based on different design matrix $\mathbf{X}$: shown in Figure 4.1 for Poisson regression model and Figure 4.2 for Negative Binomial regression model. A close examination of the table values reveals that among all the six subsampling methods, GA data have the most homogeneous subsampling probabilities. For each plot, the $\hat{\boldsymbol{\pi}}^{(k)}$ are more spread out than $\bar{\boldsymbol{\pi}}^{(k)}$, but the median values of $\bar{\boldsymbol{\pi}}^{(k)}$ are a bit bigger than those of $\hat{\boldsymbol{\pi}}^{(k)}$, the variances of $\hat{\boldsymbol{\pi}}^{(k)}$ are larger than those of $\bar{\boldsymbol{\pi}}^{(k)}$, $k = 0, 1, 2$.

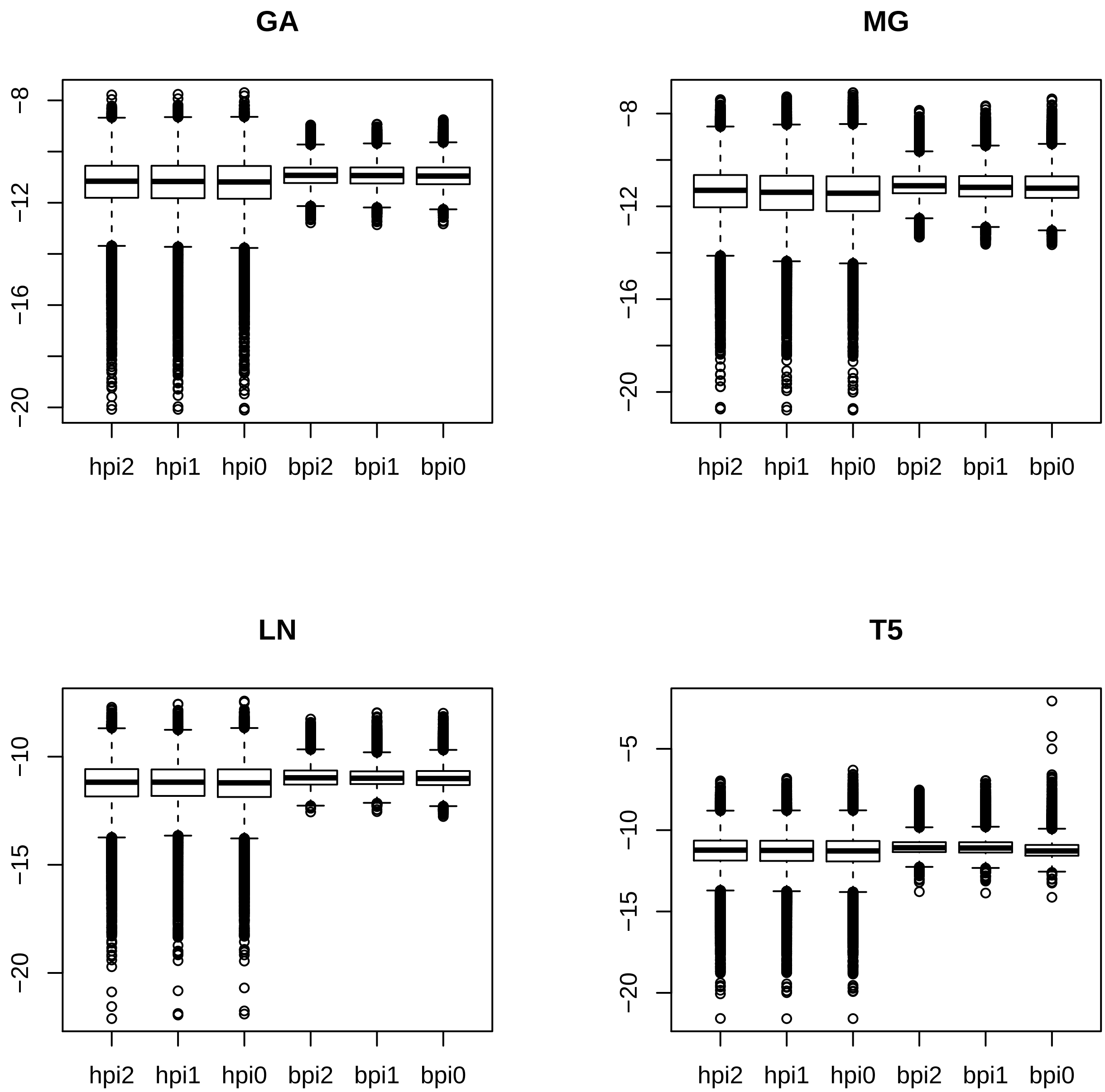


Figure 4.1. Boxplots of the logarithm of subsampling probabilities of different data sets for Poisson regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$.

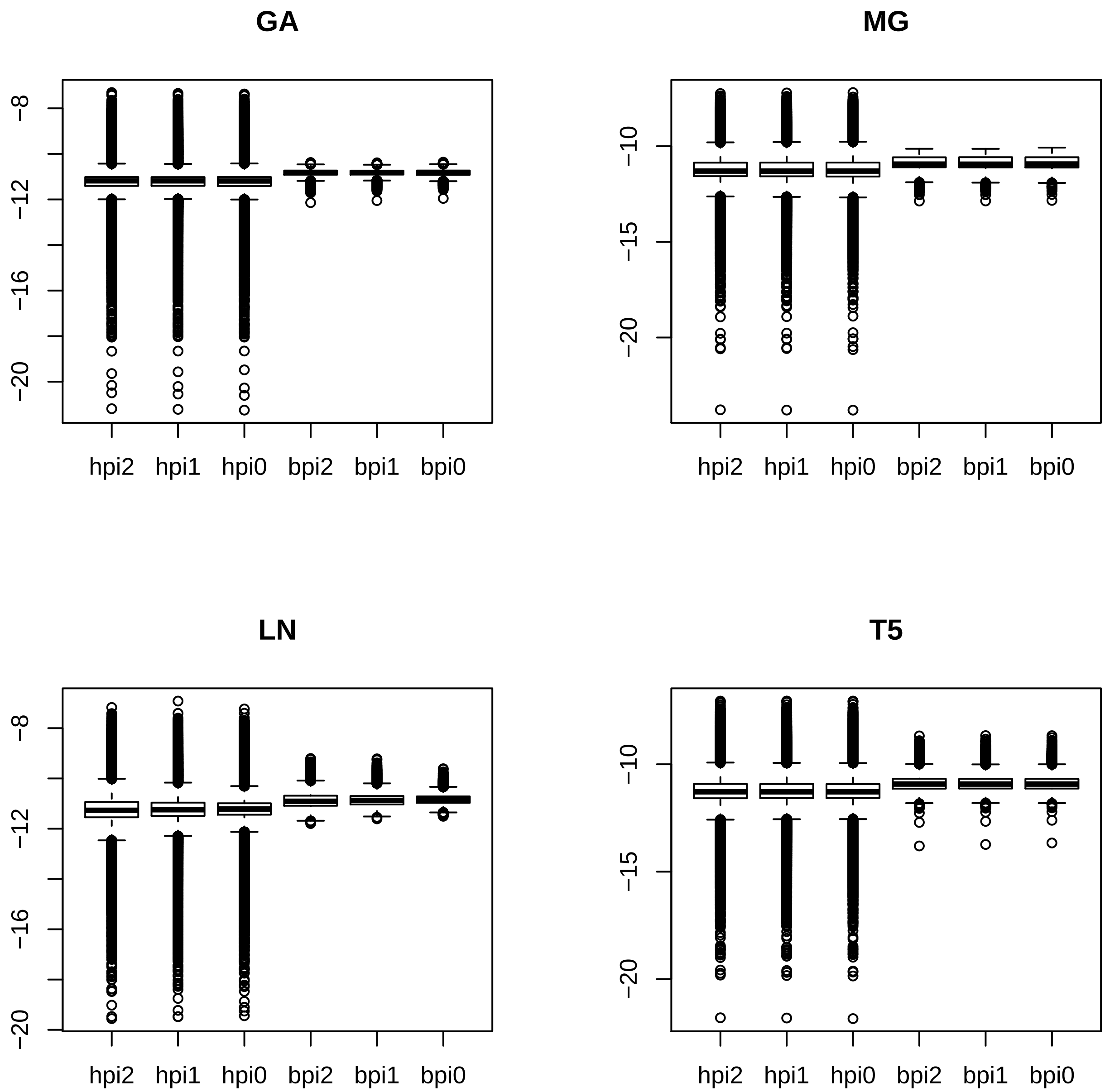
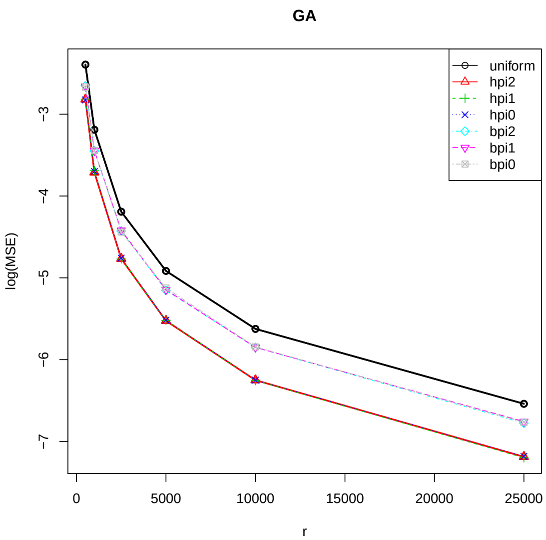
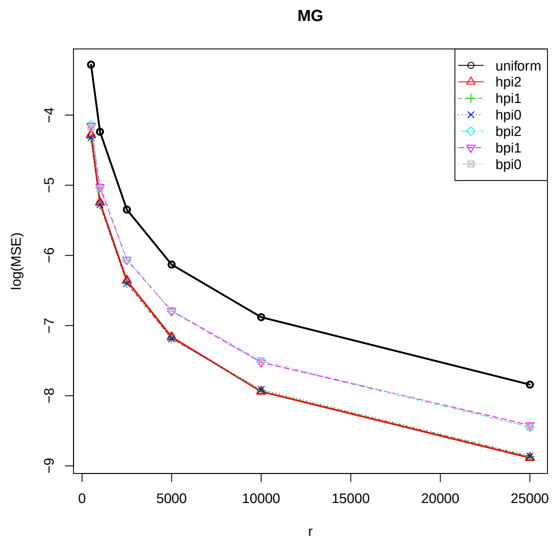


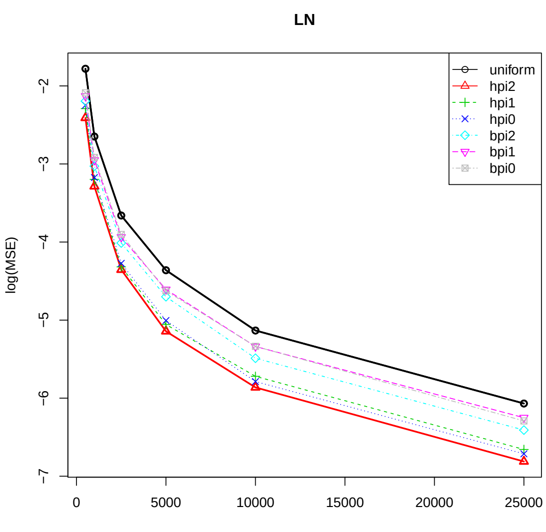
Figure 4.2. Boxplots of the logarithm of subsampling probabilities of different data sets for Negative Binomial regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$.



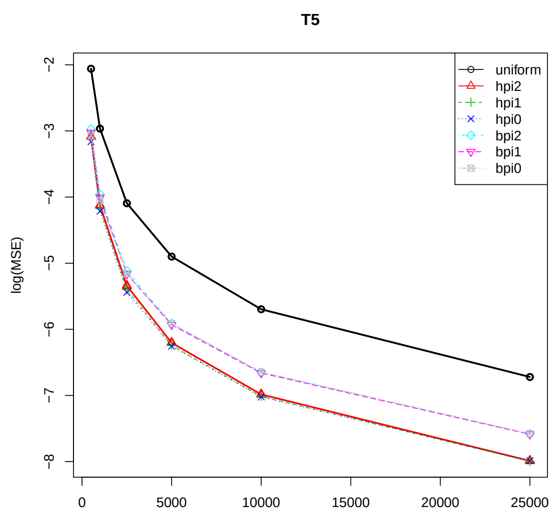
(a) GA



(b) MG

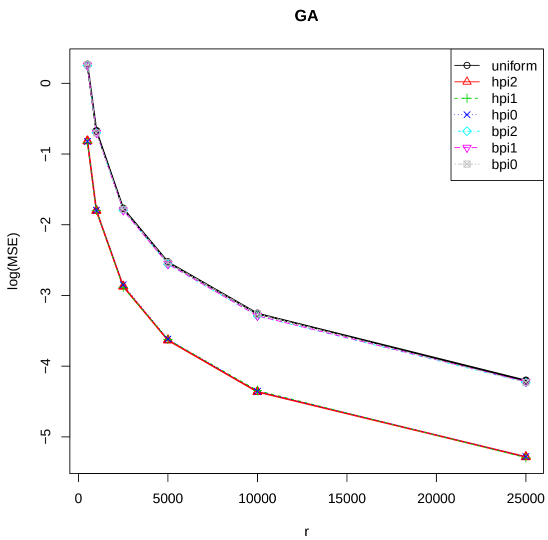


(c) LN

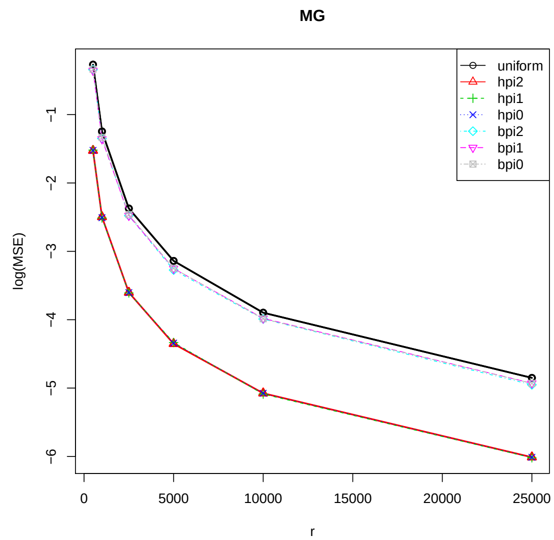


(d) T5

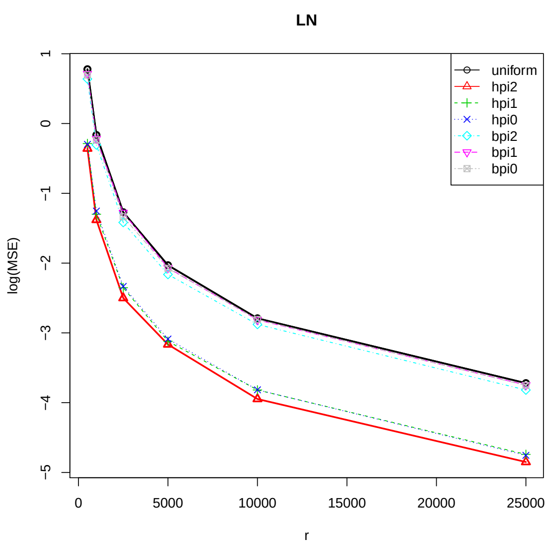
Figure 4.3. Log of the MSEs of subsampling estimator against different subsample sizes r in Poisson regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$.



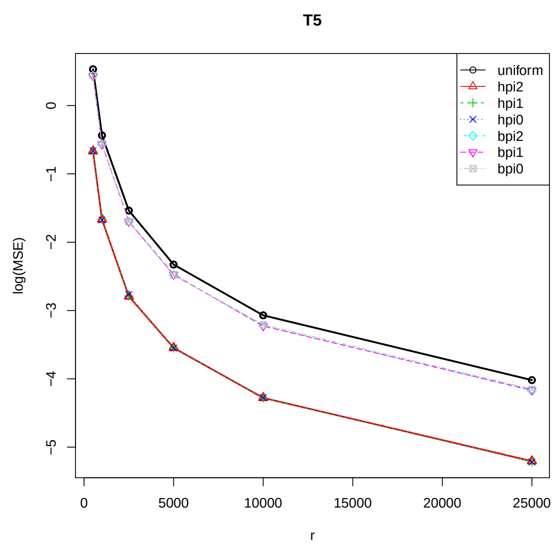
(a) GA



(b) MG



(c) LN

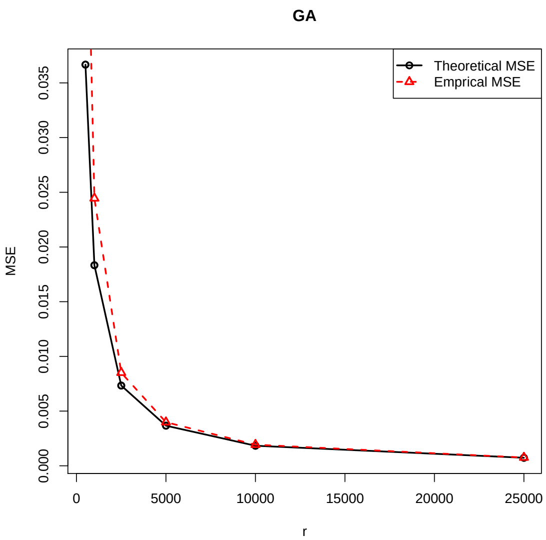


(d) T5

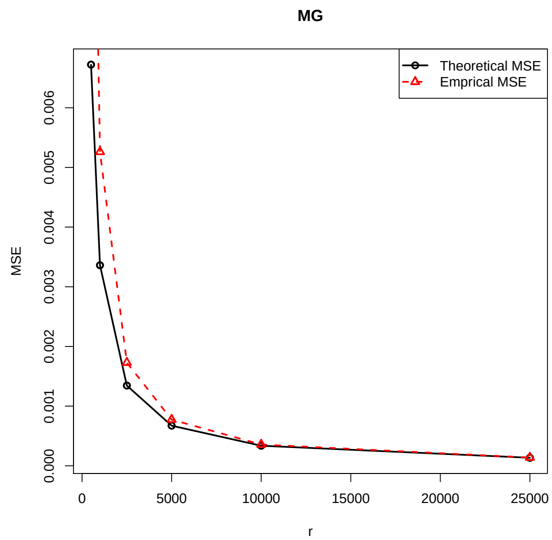
Figure 4.4. Log of the MSEs of subsampling estimator against different subsample sizes r in Negative Binomial regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$.

In Figure(4.3), we plot the logarithm of MSE of $\hat{\boldsymbol{\beta}}_r^*$ for GA, MG, LN, T5 data using uniform and proposed six subsampling methods in Poisson regression. Figure(4.4) is for Negative Binomial regression. For the four data sets, the MSE values decrease as the subsample size r increases. For all the four data sets, first of all, $\hat{\boldsymbol{\pi}}^{(k)}$, $\bar{\boldsymbol{\pi}}^{(k)}$ produce smaller MSE than the uniform subsampling; second, between $\hat{A}$-sampling and $\bar{A}$-sampling, $\hat{A}$-sampling perform the best; third, $\hat{\boldsymbol{\pi}}^{(2)}$ is the best among $\hat{\boldsymbol{\pi}}^{(k)}$, $\bar{\boldsymbol{\pi}}^{(2)}$ is the best among $\bar{\boldsymbol{\pi}}^{(k)}, k = 0, 1, 2$.

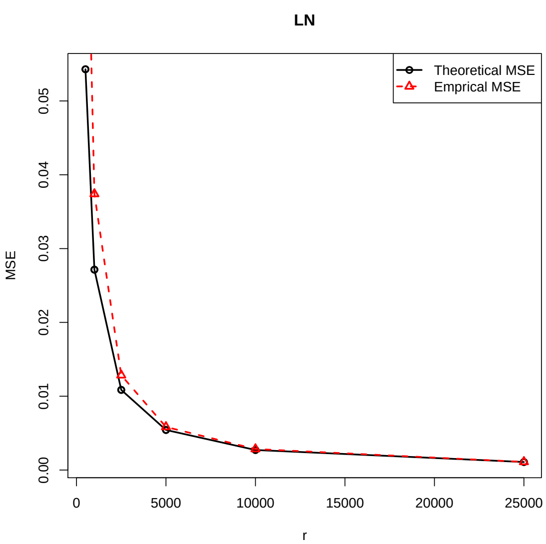
We calculate the theoretical MSE using the formula $\mathrm{tr}(\hat{\mathbf{V}})$, where $\mathrm{tr}(\hat{\mathbf{V}})$ is the trace of variance-covariance matrix of subsampling estimator $\hat{\boldsymbol{\beta}}_r^*$, and compare it with the empirical MSE in Figure(4.5-4.6). For small subsample sizes, there are some differences, but as the subsample size r increases, the differences gradually diminish.



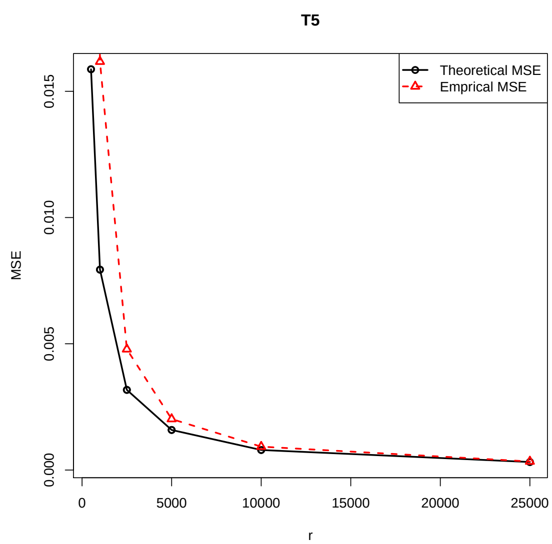
(a) GA



(b) MG

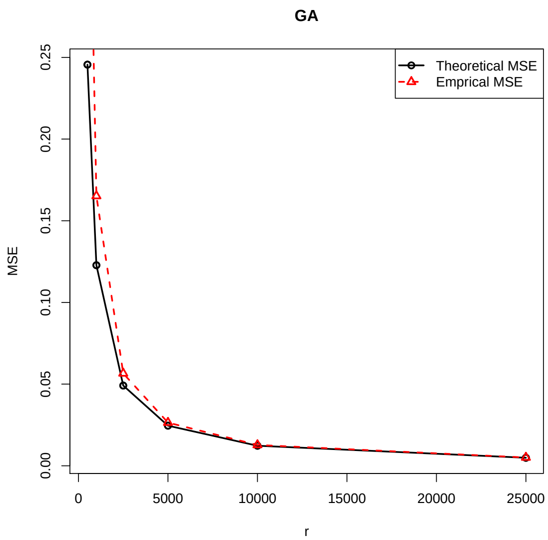


(c) LN

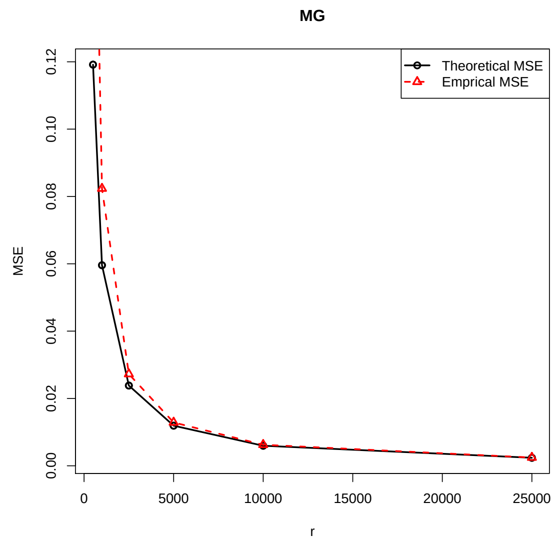


(d) T5

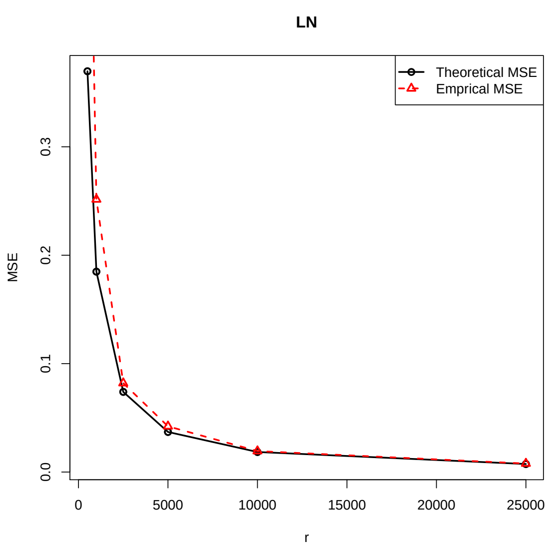
Figure 4.5. Theoretical and Empirical MSEs under $\hat{\boldsymbol{\pi}}^{(2)}$ subsampling for Poisson regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$.



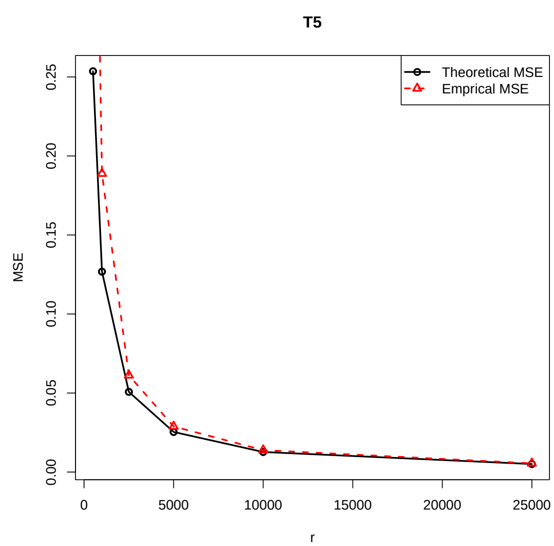
(a) GA



(b) MG



(c) LN

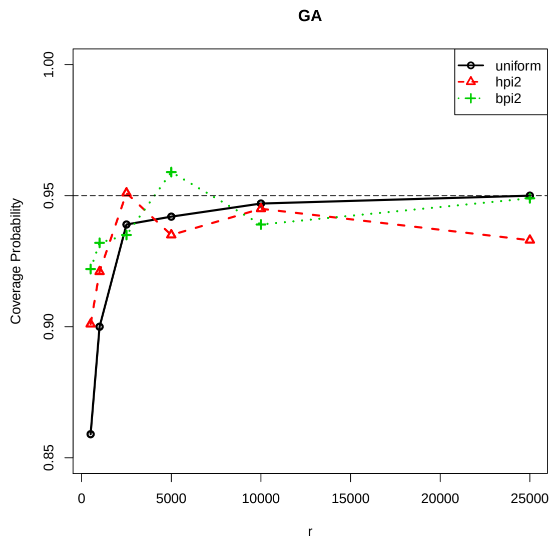


(d) T5

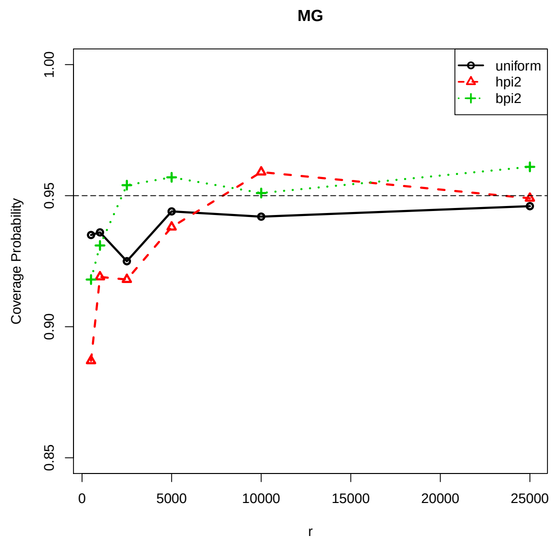
Figure 4.6. Theoretical and Empirical MSEs under $\hat{\boldsymbol{\pi}}^{(2)}$ for Negative Binomial regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$.

To exhibit the performance of the proposed method, we build confidence intervals based on asymptotic normality. We choose the second component of parameter, β_2, to demonstrate. The 95% confidence interval is calculated as $\hat{\beta}^*_{2,r} \pm Z_{0.975} SE(\hat{\beta}^*_{2,r})$, where $SE(\hat{\beta}^*_{2,r}) = \sqrt{\hat{\mathbf{V}}_{22}}$. We repeat the simulation 2,000 times and compute the percentage that the confidence intervals catch the true β_2. We report the results in Figure(4.7) for Poisson regression. The results for Negative Binomial regression are similar, and are included in Figure (4.8).

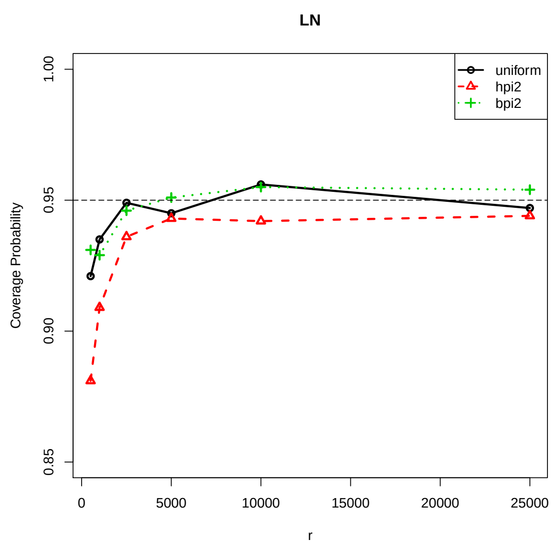
Figure(4.7) shows that when subsample size r is small, the coverage probabilities are lower than the nominal level, as the subsample size r increase, the coverage probabilities are close to the nominal level. Except for GA and LN data, the coverage probabilities under $\hat{\boldsymbol{\pi}}^{(2)}$ and $\bar{\boldsymbol{\pi}}^{(2)}$ were close to the nominal 95% than the uniform subsampling.



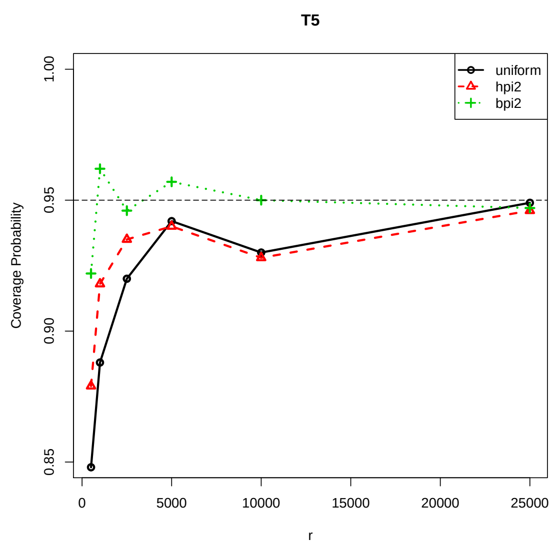
(a)



(b)



(c)



(d)

Figure 4.7. Simulated percentages of the 95% confidence intervals which caught the true parameter β_2 for different subsample sizes r, pre-subsample size $r_0 = 500$ with $n = 50,000$, $p = 50$ under $\hat{\boldsymbol{\pi}}^{(2)}$, $\bar{\boldsymbol{\pi}}^{(2)}$ and uniform subsampling in Poisson regression

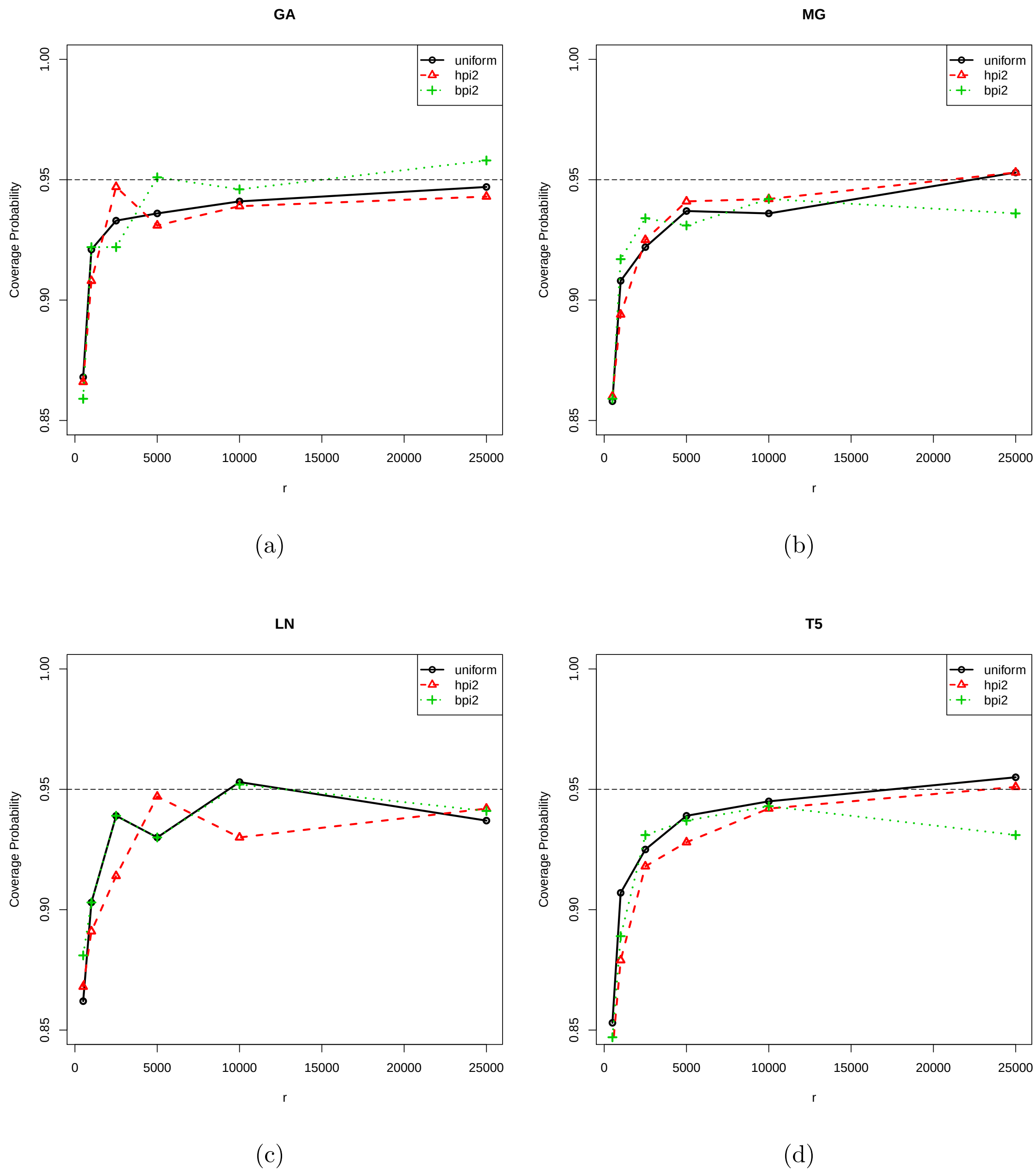


Figure 4.8. Simulated percentages of the 95% confidence intervals which caught the true parameter β_2 for different subsample sizes r, pre-subsample size $r_0 = 500$ with $n = 50,000$, $p = 50$ under $\hat{\boldsymbol{\pi}}^{(2)}$, $\bar{\boldsymbol{\pi}}^{(2)}$ and uniform subsampling in Negative Binomial regression

Next, we report the relative MSE ratios of the proposed sampling methods to uniform sampling in Tables 4.1-4.6. First, all the values in the tables are less than one, indicating all the proposed subsampling methods are better than the uniform subsampling method; $\hat{\boldsymbol{\pi}}^{(k)}$ are better than $\bar{\boldsymbol{\pi}}^{(k)}$; $\hat{\boldsymbol{\pi}}^{(2)}$ is the best. Sometimes $\hat{\boldsymbol{\pi}}^{(0)}$ and $\hat{\boldsymbol{\pi}}^{(1)}$ are very close to $\hat{\boldsymbol{\pi}}^{(2)}$. The simulated MSE ratios under the truncated $\hat{\boldsymbol{\pi}}^{(k)}$ and $\bar{\boldsymbol{\pi}}^{(k)}$ are close to the untruncated ones, $k = 0, 1, 2$. We also report the A-optimal Scoring method in Tables 4.7-4.8. We first choose a uniform pre-subsample of size $r_0 = 500$; obtain an initial estimate $\hat{\boldsymbol{\beta}}^*_{r_0}$ to approximated $\hat{\boldsymbol{\beta}}$; then approximate the proposed subsampling probabilities and use them to draw subsamples; and calculate the subsampling estimator $\hat{\boldsymbol{\beta}}^*_r$.

Table 4.1.
Simulated ratios of the MSE of the proposed subsampling estimator to the MSE of the uniform subsampling estimator for Poisson regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$.

r	500	1000	2500	5000	10000	25000
r/n	1%	2%	5%	10%	20%	50%
			GA			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.6533	0.5937	0.5627	0.5434	0.5343	0.5231
$\hat{\boldsymbol{\pi}}^{(1)}$	0.6554	0.6064	0.5613	0.5461	0.5322	0.5170
$\hat{\boldsymbol{\pi}}^{(0)}$	0.6494	0.6003	0.5672	0.5480	0.5346	0.5262
$\bar{\boldsymbol{\pi}}^{(2)}$	0.7715	0.7705	0.7898	0.7888	0.7972	0.7897
$\bar{\boldsymbol{\pi}}^{(1)}$	0.7665	0.7743	0.7960	0.7939	0.7975	0.8046
$\bar{\boldsymbol{\pi}}^{(0)}$	0.7592	0.7753	0.7794	0.8120	0.8020	0.7979
			MG			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3678	0.3642	0.3629	0.3558	0.3467	0.3524
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3502	0.3502	0.3492	0.3478	0.3504	0.3588
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3529	0.3536	0.3489	0.3465	0.3565	0.3609
$\bar{\boldsymbol{\pi}}^{(2)}$	0.4230	0.4521	0.4856	0.5137	0.5268	0.5451
$\bar{\boldsymbol{\pi}}^{(1)}$	0.4186	0.4567	0.4905	0.5166	0.5251	0.5608
$\bar{\boldsymbol{\pi}}^{(0)}$	0.4098	0.4436	0.4877	0.5193	0.5393	0.5466
			LN			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.5328	0.5285	0.4992	0.4573	0.4823	0.4756
$\hat{\boldsymbol{\pi}}^{(1)}$	0.6002	0.5776	0.5177	0.4989	0.5560	0.5549
$\hat{\boldsymbol{\pi}}^{(0)}$	0.6267	0.5914	0.5418	0.5250	0.5200	0.5248
$\bar{\boldsymbol{\pi}}^{(2)}$	0.6602	0.6842	0.7031	0.7120	0.7010	0.7114
$\bar{\boldsymbol{\pi}}^{(1)}$	0.7049	0.7390	0.7586	0.7811	0.8152	0.8336
$\bar{\boldsymbol{\pi}}^{(0)}$	0.7348	0.7644	0.7840	0.7679	0.8163	0.7998
			T5			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3587	0.3137	0.2867	0.2714	0.2760	0.2810
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3469	0.2987	0.2709	0.2608	0.2678	0.2784
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3318	0.2872	0.2598	0.2578	0.2657	0.2822
$\bar{\boldsymbol{\pi}}^{(2)}$	0.4013	0.3695	0.3596	0.3636	0.3861	0.4229
$\bar{\boldsymbol{\pi}}^{(1)}$	0.3807	0.3527	0.3426	0.3562	0.3812	0.4207
$\bar{\boldsymbol{\pi}}^{(0)}$	0.3622	0.3351	0.3445	0.3629	0.3867	0.4240

Table 4.2.
Simulated ratios of the MSE of the proposed subsampling estimator to the MSE of the uniform subsampling estimator for Poisson regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$ and truncation 10%.

r	500	1000	2500	5000	10000	25000
r/n	1%	2%	5%	10%	20%	50%
			GA			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.6434	0.5718	0.5499	0.5325	0.5310	0.5159
$\hat{\boldsymbol{\pi}}^{(1)}$	0.6271	0.5811	0.5450	0.5389	0.5265	0.5199
$\hat{\boldsymbol{\pi}}^{(0)}$	0.6299	0.5823	0.5481	0.5410	0.5302	0.5163
$\bar{\boldsymbol{\pi}}^{(2)}$	0.7730	0.7662	0.7898	0.7965	0.7980	0.7960
$\bar{\boldsymbol{\pi}}^{(1)}$	0.7688	0.7658	0.8021	0.8032	0.7968	0.8079
$\bar{\boldsymbol{\pi}}^{(0)}$	0.7701	0.7726	0.7866	0.8121	0.8122	0.7935
			MG			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3671	0.3571	0.3534	0.3444	0.3483	0.3540
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3525	0.3527	0.3404	0.3464	0.3534	0.3633
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3488	0.3410	0.3441	0.3527	0.3524	0.3629
$\bar{\boldsymbol{\pi}}^{(2)}$	0.4236	0.4483	0.4925	0.5070	0.5306	0.5486
$\bar{\boldsymbol{\pi}}^{(1)}$	0.4201	0.4555	0.4938	0.5070	0.5340	0.5497
$\bar{\boldsymbol{\pi}}^{(0)}$	0.4111	0.4473	0.4915	0.5202	0.5354	0.5510
			LN			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.5230	0.5275	0.4854	0.4466	0.4847	0.4764
$\hat{\boldsymbol{\pi}}^{(1)}$	0.5865	0.5411	0.5404	0.4924	0.5453	0.5439
$\hat{\boldsymbol{\pi}}^{(0)}$	0.5853	0.5853	0.5359	0.4973	0.5124	0.5395
$\bar{\boldsymbol{\pi}}^{(2)}$	0.6571	0.6894	0.6773	0.6833	0.7002	0.7404
$\bar{\boldsymbol{\pi}}^{(1)}$	0.6965	0.7325	0.7799	0.7791	0.8176	0.8318
$\bar{\boldsymbol{\pi}}^{(0)}$	0.7126	0.7565	0.8055	0.7710	0.8076	0.8029
			T5			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3538	0.3060	0.2815	0.2722	0.2753	0.2823
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3394	0.2900	0.2678	0.2595	0.2650	0.2817
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3233	0.2793	0.2604	0.2587	0.2659	0.2824
$\bar{\boldsymbol{\pi}}^{(2)}$	0.4081	0.3721	0.3595	0.3680	0.3872	0.4241
$\bar{\boldsymbol{\pi}}^{(1)}$	0.3844	0.3565	0.3451	0.3600	0.3812	0.4232
$\bar{\boldsymbol{\pi}}^{(0)}$	0.3613	0.3356	0.3453	0.3667	0.3885	0.4258

Table 4.3.
Simulated ratios of the MSE of the proposed subsampling estimator to the MSE of the uniform subsampling estimator for Poisson regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$ and truncation 30%.

r	500	1000	2500	5000	10000	25000
r/n	1%	2%	5%	10%	20%	50%
			GA			
$\hat{\pi}^{(2)}$	0.6196	0.5769	0.5551	0.5372	0.5371	0.5317
$\hat{\pi}^{(1)}$	0.6198	0.5752	0.5480	0.5435	0.5381	0.5373
$\hat{\pi}^{(0)}$	0.6185	0.5723	0.5465	0.5486	0.5377	0.5345
$\bar{\pi}^{(2)}$	0.7816	0.7805	0.8033	0.8041	0.8077	0.8126
$\bar{\pi}^{(1)}$	0.7832	0.7811	0.8103	0.8055	0.8137	0.8168
$\bar{\pi}^{(0)}$	0.7774	0.7823	0.7997	0.8125	0.8140	0.8043
			MG			
$\hat{\pi}^{(2)}$	0.3667	0.3625	0.3568	0.3573	0.3536	0.3633
$\hat{\pi}^{(1)}$	0.3515	0.3556	0.3491	0.3544	0.3660	0.3674
$\hat{\pi}^{(0)}$	0.3502	0.3493	0.3488	0.3590	0.3515	0.3611
$\bar{\pi}^{(2)}$	0.4309	0.4629	0.4868	0.5226	0.5351	0.5524
$\bar{\pi}^{(1)}$	0.4250	0.4539	0.4985	0.5176	0.5354	0.5630
$\bar{\pi}^{(0)}$	0.4102	0.4484	0.4977	0.5193	0.5393	0.5617
			LN			
$\hat{\pi}^{(2)}$	0.5193	0.5118	0.4905	0.4791	0.4721	0.5021
$\hat{\pi}^{(1)}$	0.5619	0.5496	0.5325	0.5132	0.5637	0.5466
$\hat{\pi}^{(0)}$	0.5596	0.5675	0.5274	0.5120	0.5204	0.5371
$\bar{\pi}^{(2)}$	0.6654	0.6930	0.7116	0.7232	0.7366	0.7309
$\bar{\pi}^{(1)}$	0.6989	0.7329	0.7832	0.7546	0.8173	0.8252
$\bar{\pi}^{(0)}$	0.7181	0.7604	0.7909	0.7819	0.8316	0.8255
			T5			
$\hat{\pi}^{(2)}$	0.3608	0.3160	0.2893	0.2763	0.2826	0.2880
$\hat{\pi}^{(1)}$	0.3460	0.2966	0.2785	0.2700	0.2724	0.2864
$\hat{\pi}^{(0)}$	0.3295	0.2843	0.2629	0.2658	0.2763	0.2954
$\bar{\pi}^{(2)}$	0.4121	0.3778	0.3704	0.3645	0.3875	0.4239
$\bar{\pi}^{(1)}$	0.3882	0.3588	0.3493	0.3602	0.3860	0.4232
$\bar{\pi}^{(0)}$	0.3676	0.3374	0.3424	0.3639	0.3911	0.4198

Table 4.4.
Simulated ratios of the MSE of the proposed subsampling estimator to the MSE of the uniform subsampling estimator for Negative Binomial regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$.

r	500	1000	2500	5000	10000	25000
r/n	1%	2%	5%	10%	20%	50%
			GA			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3390	0.3243	0.3328	0.3319	0.3310	0.3398
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3315	0.3259	0.3281	0.3359	0.3372	0.3358
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3374	0.3265	0.3408	0.3374	0.3333	0.3416
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9850	0.9747	0.9765	0.9699	0.9668	0.9738
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9881	0.9752	0.9775	0.9747	0.9695	0.9845
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9992	0.9739	0.9913	0.9955	0.9978	0.9757
			MG			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.2843	0.2863	0.2924	0.2974	0.3078	0.3132
$\hat{\boldsymbol{\pi}}^{(1)}$	0.2863	0.2819	0.2908	0.3030	0.3040	0.3107
$\hat{\boldsymbol{\pi}}^{(0)}$	0.2854	0.2831	0.2922	0.3004	0.3076	0.3129
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9295	0.9020	0.9000	0.8748	0.9118	0.9040
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9164	0.8945	0.9006	0.8936	0.9203	0.9243
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9347	0.9163	0.9142	0.8970	0.9152	0.9229
			LN			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3208	0.2963	0.2923	0.3214	0.3148	0.3229
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3447	0.3214	0.3389	0.3364	0.3584	0.3603
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3409	0.3361	0.3454	0.3474	0.3590	0.3554
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8698	0.8666	0.8634	0.8762	0.9167	0.9062
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9364	0.9482	0.9942	0.9643	0.9789	0.9733
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9197	0.9289	0.9370	0.9564	0.9849	0.9673
			T5			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3013	0.2923	0.2844	0.2955	0.2986	0.3053
$\hat{\boldsymbol{\pi}}^{(1)}$	0.2979	0.2933	0.2863	0.2956	0.2983	0.3027
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3034	0.2898	0.2924	0.2944	0.2998	0.3014
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9115	0.8764	0.8493	0.8565	0.8599	0.8543
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9087	0.8787	0.8516	0.8658	0.8545	0.8632
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9107	0.8861	0.8461	0.8546	0.8730	0.8752

Table 4.5.
Simulated ratios of the MSE of the proposed subsampling estimator to the MSE of the uniform subsampling estimator for Negative Binomial regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$ and truncation 10%.

r	500	1000	2500	5000	10000	25000
r/n	1%	2%	5%	10%	20%	50%
			GA			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3158	0.3146	0.3273	0.3301	0.3375	0.3390
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3158	0.3184	0.3284	0.3253	0.3370	0.3366
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3171	0.3162	0.3269	0.3308	0.3362	0.3391
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9836	0.9777	0.9828	0.9807	0.9761	0.9685
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9797	0.9771	0.9901	0.9670	0.9732	0.9783
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9693	0.9804	0.9931	0.9763	0.9792	0.9720
			MG			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.2793	0.2801	0.2901	0.2987	0.3014	0.3108
$\hat{\boldsymbol{\pi}}^{(1)}$	0.2756	0.2722	0.2888	0.3009	0.3048	0.3099
$\hat{\boldsymbol{\pi}}^{(0)}$	0.2793	0.2762	0.2959	0.2989	0.3078	0.3116
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9420	0.9153	0.9214	0.9208	0.8974	0.9136
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9524	0.9160	0.9236	0.9062	0.8968	0.9199
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9404	0.9213	0.9039	0.9301	0.9096	0.9147
			LN			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.2936	0.2887	0.2768	0.3003	0.3024	0.3175
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3125	0.3169	0.3067	0.3230	0.3348	0.3719
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3233	0.3062	0.3069	0.3233	0.3294	0.3652
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8520	0.8418	0.8104	0.8698	0.8743	0.8878
$\bar{\boldsymbol{\pi}}^{(1)}$	0.8721	0.9179	0.8642	0.9182	0.9409	0.9457
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9088	0.9586	0.8802	0.8937	0.9499	0.9804
			T5			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.2855	0.2843	0.2843	0.2881	0.2902	0.3015
$\hat{\boldsymbol{\pi}}^{(1)}$	0.2871	0.2817	0.2812	0.2910	0.2969	0.3014
$\hat{\boldsymbol{\pi}}^{(0)}$	0.2875	0.2819	0.2842	0.2903	0.2991	0.2960
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8808	0.8615	0.8441	0.8464	0.8579	0.8484
$\bar{\boldsymbol{\pi}}^{(1)}$	0.8945	0.8723	0.8583	0.8475	0.8497	0.8476
$\bar{\boldsymbol{\pi}}^{(0)}$	0.8965	0.8792	0.8621	0.8601	0.8516	0.8470

Table 4.6.
Simulated ratios of the MSE of the proposed subsampling estimator to the MSE of the uniform subsampling estimator for Negative Binomial regression based on the full sample estimator $\hat{\boldsymbol{\beta}}$ with $n = 50,000$, $p = 50$ and truncation 30%.

r	500	1000	2500	5000	10000	25000
r/n	1%	2%	5%	10%	20%	50%
			GA			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3163	0.3154	0.3307	0.3333	0.3416	0.3471
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3091	0.3200	0.3286	0.3334	0.3417	0.3435
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3163	0.3199	0.3363	0.3349	0.3400	0.3476
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9869	0.9854	0.9928	0.9835	0.9831	0.9716
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9797	0.9859	0.9910	0.9656	0.9860	0.9828
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9671	0.9795	0.9934	0.9739	0.9882	0.9734
			MG			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.2735	0.2780	0.2944	0.3023	0.3077	0.3155
$\hat{\boldsymbol{\pi}}^{(1)}$	0.2715	0.2762	0.2930	0.3069	0.3116	0.3143
$\hat{\boldsymbol{\pi}}^{(0)}$	0.2796	0.2809	0.2962	0.3068	0.3150	0.3187
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9551	0.9141	0.9206	0.9370	0.9004	0.9148
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9483	0.9256	0.9297	0.9127	0.9088	0.9279
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9340	0.9295	0.9061	0.9305	0.9129	0.9191
			LN			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.2909	0.2874	0.2925	0.3050	0.2907	0.3126
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3255	0.3129	0.3126	0.3418	0.3258	0.3435
$\hat{\boldsymbol{\pi}}^{(0)}$	0.3119	0.3235	0.3249	0.3390	0.3134	0.3446
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8524	0.8349	0.8412	0.8808	0.8313	0.8860
$\bar{\boldsymbol{\pi}}^{(1)}$	0.8938	0.8668	0.9118	0.9429	0.8637	0.9391
$\bar{\boldsymbol{\pi}}^{(0)}$	0.8918	0.8835	0.9237	0.9518	0.9301	0.9241
			T5			
$\hat{\boldsymbol{\pi}}^{(2)}$	0.2921	0.2842	0.2888	0.2932	0.2981	0.3083
$\hat{\boldsymbol{\pi}}^{(1)}$	0.2880	0.2847	0.2876	0.2957	0.3039	0.3047
$\hat{\boldsymbol{\pi}}^{(0)}$	0.2867	0.2885	0.2919	0.2911	0.2998	0.3048
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8796	0.8935	0.8612	0.8459	0.8582	0.8555
$\bar{\boldsymbol{\pi}}^{(1)}$	0.8767	0.8819	0.8668	0.8484	0.8623	0.8532
$\bar{\boldsymbol{\pi}}^{(0)}$	0.8964	0.8898	0.8797	0.8615	0.8484	0.8537

Table 4.7.
Simulated ratios of the MSE of the proposed subsampling estimator to the MSE of the uniform subsampling estimator for Poisson regression using A-optimal Scoring method with pre-subsample size $r_0 = 500$, $n = 50,000$, $p = 50$.

r	500	1000	2500	5000	10000	25000
r/n	1%	2%	5%	10%	20%	50%
			GA			
$\hat{\pi}^{(2)}$	0.7778	0.7375	0.7749	0.8050	0.8276	0.8499
$\hat{\pi}^{(1)}$	0.7794	0.7594	0.7781	0.7898	0.8259	0.8778
$\hat{\pi}^{(0)}$	0.7792	0.7657	0.7750	0.8096	0.8413	0.8725
$\bar{\pi}^{(2)}$	0.7805	0.7879	0.8036	0.8237	0.8300	0.8205
$\bar{\pi}^{(1)}$	0.7930	0.7888	0.8188	0.8341	0.8271	0.8174
$\bar{\pi}^{(0)}$	0.7911	0.7967	0.8293	0.8419	0.8494	0.8416
			MG			
$\hat{\pi}^{(2)}$	0.4192	0.4869	0.5671	0.6089	0.7003	0.7533
$\hat{\pi}^{(1)}$	0.4339	0.5021	0.5856	0.6567	0.7313	0.7869
$\hat{\pi}^{(0)}$	0.4486	0.5219	0.5941	0.6723	0.7247	0.7884
$\bar{\pi}^{(2)}$	0.4270	0.4712	0.4905	0.5279	0.5555	0.5557
$\bar{\pi}^{(1)}$	0.4195	0.4579	0.5144	0.5157	0.5618	0.5620
$\bar{\pi}^{(0)}$	0.4254	0.4603	0.4854	0.5371	0.5735	0.5805
			LN			
$\hat{\pi}^{(2)}$	0.6271	0.6467	0.6639	0.6623	0.7056	0.7639
$\hat{\pi}^{(1)}$	0.6990	0.7057	0.6935	0.7226	0.8034	0.8218
$\hat{\pi}^{(0)}$	0.7114	0.7384	0.7262	0.7301	0.8335	0.8643
$\bar{\pi}^{(2)}$	0.6606	0.6884	0.7185	0.7238	0.7160	0.7500
$\bar{\pi}^{(1)}$	0.6960	0.7362	0.7549	0.7833	0.8286	0.8412
$\bar{\pi}^{(0)}$	0.7329	0.7824	0.8193	0.7992	0.8546	0.8145
			T5			
$\hat{\pi}^{(2)}$	0.3184	0.3077	0.2828	0.2969	0.3139	0.3291
$\hat{\pi}^{(1)}$	0.3079	0.2933	0.2964	0.2957	0.3111	0.3295
$\hat{\pi}^{(0)}$	0.3260	0.3087	0.3022	0.3084	0.3240	0.3419
$\bar{\pi}^{(2)}$	0.3956	0.3808	0.3626	0.3719	0.3927	0.4156
$\bar{\pi}^{(1)}$	0.3744	0.3483	0.3500	0.3596	0.3853	0.4209
$\bar{\pi}^{(0)}$	0.3419	0.3425	0.3521	0.3628	0.3967	0.4285

Table 4.8.
Simulated ratios of the MSE of the proposed subsampling estimator to the MSE of the uniform subsampling estimator for Negative Binomial regression using A-optimal Scoring method with presubsample size $r_0 = 500$, $n = 50,000$, $p = 50$.

r	500	1000	2500	5000	10000	25000
r/n	1%	2%	5%	10%	20%	50%
			GA			
$\hat{\pi}^{(2)}$	0.3814	0.3811	0.3822	0.3847	0.3840	0.3920
$\hat{\pi}^{(1)}$	0.3837	0.3790	0.3835	0.3858	0.3867	0.4050
$\hat{\pi}^{(0)}$	0.3788	0.3841	0.3851	0.3821	0.3868	0.3954
$\bar{\pi}^{(2)}$	1.0045	0.9018	0.9891	0.9718	0.9738	0.9896
$\bar{\pi}^{(1)}$	0.9895	0.9757	0.9905	0.9849	0.9701	0.9860
$\bar{\pi}^{(0)}$	0.9831	0.9543	0.9872	0.9744	0.9925	0.9855
			MG			
$\hat{\pi}^{(2)}$	0.3140	0.3386	0.3478	0.3578	0.3852	0.3800
$\hat{\pi}^{(1)}$	0.3232	0.3385	0.3438	0.3672	0.3859	0.3866
$\hat{\pi}^{(0)}$	0.3300	0.3405	0.3480	0.3670	0.3803	0.3801
$\bar{\pi}^{(2)}$	0.9098	0.9207	0.8999	0.9189	0.9346	0.8895
$\bar{\pi}^{(1)}$	0.9286	0.9233	0.9198	0.9253	0.9341	0.9117
$\bar{\pi}^{(0)}$	0.9521	0.9209	0.9021	0.9141	0.9454	0.9161
			LN			
$\hat{\pi}^{(2)}$	0.3759	0.3577	0.3380	0.3625	0.3892	0.3796
$\hat{\pi}^{(1)}$	0.4049	0.3750	0.3696	0.3977	0.4197	0.4378
$\hat{\pi}^{(0)}$	0.3976	0.3793	0.3573	0.3858	0.4499	0.4148
$\bar{\pi}^{(2)}$	0.8391	0.8651	0.8383	0.8846	0.9278	0.9738
$\bar{\pi}^{(1)}$	0.9403	0.9732	0.8511	0.9292	0.9367	0.9631
$\bar{\pi}^{(0)}$	0.9426	0.9851	0.9166	0.9415	0.9132	0.9970
			T5			
$\hat{\pi}^{(2)}$	0.3473	0.3404	0.3480	0.3473	0.3576	0.3747
$\hat{\pi}^{(1)}$	0.3521	0.3383	0.3498	0.3526	0.3601	0.3620
$\hat{\pi}^{(0)}$	0.3462	0.3426	0.3473	0.3573	0.3622	0.3672
$\bar{\pi}^{(2)}$	0.8952	0.8512	0.8679	0.8400	0.8387	0.8497
$\bar{\pi}^{(1)}$	0.8704	0.8551	0.8690	0.8528	0.8337	0.8557
$\bar{\pi}^{(0)}$	0.9097	0.8591	0.8697	0.8583	0.8465	0.8518

In order to evaluate the computation efficiency, we report the running times for computing $\hat{\boldsymbol{\beta}}_r^*$ by using $\hat{\boldsymbol{\pi}}^{(2)}$,$\bar{\boldsymbol{\pi}}^{(2)}$ in Tables 4.9-4.10. The experiment is carried out using R programming language. Those values were computed on a desktop with Intel i5 processor and 8GB memory. We recorded the CPU times for 1000 repetitions, then average the time to make the comparison fair. We observe that the $\hat{\pi}^{(2)}$ require more time than $\bar{\pi}^{(2)}$ method. All the proposed methods have significant less computing times than the full data. In Table(4.11), we can see all the proposed methods have similar number of iterations, indicating smaller subsample sizes do not necessarily increase the iterations for Newton's method.

Table 4.9.
The CPU times in seconds for GA in Poisson regression using the A-optimal Scoring method with pre-subsample size $r_0 = 500$, $n = 50,000$, $p = 50$.

r	500	1000	1500	2000	2500	5000
r/n	1%	2%	3%	4%	5%	10%
$\hat{\pi}^{(2)}$	4.191	4.205	4.226	4.241	4.567	4.632
$\bar{\pi}^{(2)}$	2.313	2.334	2.356	2.395	3.025	3.564
Full data CPU seconds 5.872						

Table 4.10.
The CPU times in seconds using Newton's method of the different full sample sizes for GA in Poisson regression with $r_0 = 500$ and $r = 2000$.

r	10^4	10^5	10^6	0.5×10^7
$\hat{\pi}^{(2)}$	0.70	4.67	26.30	98.06
$\bar{\pi}^{(2)}$	0.64	3.50	15.22	49.22
Full	0.76	6.59	58.26	299.18

Table 4.11.
Averaged iterations using Newton's method for GA in Poisson regression with $r_0 = 500$ and various r. The iterations for full data set are 8.4.

	$\hat{\boldsymbol{\pi}}^{(2)}$		$\bar{\boldsymbol{\pi}}^{(2)}$		Uniform
r	Step1	Step2	Step1	Step2	
500	8.89	8.77	8.67	8.49	8.40
1000	8.75	8.56	8.56	8.23	8.80
1500	8.56	8.32	8.59	8.39	8.54
2000	8.55	8.01	8.58	8.53	8.34
2500	8.60	8.91	8.62	8.85	8.27

5. FULL SAMPLE REAL DATA ANALYSIS: BIKE SHARING DATA

5.1 Introduction of the Real Data

This data set is available from the UCI Machine Learning Repository website. Bike sharing systems can be considered new generation of old-fashioned bike rentals. The use of this system is not restricted to rentals and returns at the same docking station, bikes can be returned to any docking station after usage. Predicting the hourly bike request will help in planing, expanding and maintaining adequate number of bikes. In United States, the bike sharing system has been proved to be very successful in major cities, including Washington, DC, New York, Chicago, Los Angles, where bikes sharing has become a popular transportation option. Our goal in this example is to build statistical models to predict hourly request of bikes in Washington DC area. There are totally 17,389 observations in this data set, the response variables are the counts of casual rentals, registered rentals, and total rented bikes including both casual and registered. We split the data into two sets, using the 2011 year data to build the models and the 2012 data to calculate the prediction error. The predictor variables are season, workingday, daytime, weathersit, temp, hum and windspeed. The season variables include spring indicator, summer indicator and fall indicator, winter is the reference level. Variable workingday indicates whether a day is a working day, the reference level is weekend of holiday. Variable daytime indicates if the time is between 7am to 22pm, with the referencing time range from 0am to 5am representing the reference level night time. Weathersit varaible has 4 categories, the first category represents clear, few clouds, and partly cloudy; the second category represents mist plus cloudy, mist plus broken clouds, and mist; the third category represents light snow, light rain; and the fourth category represents heavy rain, and snow plus frog.

Since the fourth category only has three observations, we combine the third and fourth categories. We choose the first category as the reference. Temp is normalized temperature in Celsius. Hum is normalized humidity. Windspeed is normalized wind speed. There are 11 regression coefficients for the predictor variables including the intercept.

5.2 Explanatory Data Analysis

The response variables are count of hourly rental bikes that are non-negative integers, so we first use Poisson Regression to fit the model. We fit a Poisson Regression model for each of the 3 response variables: the casual rental, the registered rental, and the combined rental. Overdispersion test performs the following linear regressions for Poisson regression and Negative Binomial regression respectively. The test can be found in R package AER, see Zeileis and Kleiber (2008).

$$\frac{(y_i - \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})) - y_i}{\exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})} = a + \epsilon_i, \tag{5.2.1}$$

$$\frac{(y_i - \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})) - y_i}{\exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})} = a \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) + \epsilon_i, \tag{5.2.2}$$

where ϵ_i is the error term. When the estimate of a is close to 0, there is no overdispersion; a positive estimate of a in (5.2.1) indicates overdispersion with $V(y_i) = \phi\mu_i$ for the Quasipoisson regression. In (5.2.2), $V(y_i) = \mu_i + \alpha\mu_i^2$ represents Negative Binomial regression.

Table 5.1.
Dispersion tests for Poisson regression models with the casual bike rental, the registered bike rental, and the combined bike rental as the responses.

$H_0 : \phi = 0 \quad vs \quad H_1 : \phi > 0$			
Response	Test Statistics	P-value	Decision
Casual	34.582	< 0.0001	Reject
Register	50.561	< 0.0001	Reject
Combined	54.021	< 0.0001	Reject
$H_0 : \alpha = 0 \quad vs \quad H_1 : \alpha > 0$			
Response	Test Statistics	P-value	Decision
Casual	40.644	< 0.0001	Reject
Register	53.086	< 0.0001	Reject
Combined	53.300	< 0.0001	Reject

From the test results, we find overdispersion indeed exists, hence will apply Quasipoisson and Negative Binomial regression to the data set. We perform univariate analysis to see the effect of a single covariate on the response variables. We report the results in Tables 5.2-5.4 for Quasipoisson regression. Negative Binomial regression results are given in Tables 5.5-5.7. All the covariates are significant in Table 5.2, which means all can be viewed as good candidate variables to be included into the multiple regression models.

Table 5.2.
Univariate analysis in Quasipoisson regression model with the casual bike rental as the response variable using the full sample, $n = 8,645$.

	Estimate	SE	Z	P-value
Intercept	3.27112	0.01760	185.88680	< 0.0001
Summer	0.29016	0.03142	9.23530	< 0.0001
Intercept	3.16550	0.01838	172.24255	< 0.0001
Fall	0.58663	0.02958	19.83260	< 0.0001
Intercept	3.39663	0.01646	206.35876	< 0.0001
Winter	-0.18787	0.03561	-5.27502	< 0.0001
Intercept	3.85326	0.01718	224.23642	< 0.0001
Workingday	-0.85639	0.02484	-34.47975	< 0.0001
Intercept	1.68446	0.04779	35.24761	< 0.0001
Daytime	2.00169	0.04934	40.57321	< 0.0001
Intercept	3.40448	0.01642	207.34006	< 0.0001
W2_cloudy	-0.21564	0.03520	-6.12661	< 0.0001
Intercept	3.40818	0.01468	232.14225	< 0.0001
W3_rain	-0.88912	0.07408	-12.00141	< 0.0001
Intercept	1.37418	0.04378	31.39153	< 0.0001
Temp	0.68555	0.01312	52.25234	< 0.0001
Intercept	4.61196	0.04057	113.69199	< 0.0001
Hum	-0.40293	0.01314	-30.65711	< 0.0001
Intercept	3.21356	0.02757	116.54572	< 0.0001
Windspeed	0.08701	0.01415	6.14945	< 0.0001

Table 5.3.
Univariate analysis in Quasipoisson regression model with the registered bike rental as the response variable using the full sample, $n = 8,645$.

	Estimate	SE	Z	P-value
Intercept	4.72485	0.01196	394.98062	< 0.0001
Summer	0.08284	0.02298	3.60404	0.00032
Intercept	4.65266	0.01227	379.20036	< 0.0001
Fall	0.32222	0.02150	14.98425	< 0.0001
Intercept	4.70926	0.01198	393.07036	< 0.0001
Winter	0.14340	0.02287	6.27049	< 0.0001
Intercept	4.54697	0.01964	231.45786	< 0.0001
Workingday	0.28004	0.02282	12.27094	< 0.0001
Intercept	3.24840	0.02874	113.03048	< 0.0001
Daytime	1.81863	0.02985	60.92212	< 0.0001
Intercept	4.75842	0.01177	404.13371	< 0.0001
W2_cloudy	-0.04681	0.02366	-1.97885	0.04786
Intercept	4.78161	0.01046	457.21914	< 0.0001
W3_rain	-0.47836	0.04340	-11.02192	< 0.0001
Intercept	3.75916	0.02907	129.30906	< 0.0001
Temp	0.36190	0.00945	38.30035	< 0.0001
Intercept	5.46915	0.03182	171.85133	< 0.0001
Hum	-0.22455	0.00975	-23.04188	< 0.0001
Intercept	4.62696	0.01920	240.93188	< 0.0001
Windspeed	0.07474	0.00992	7.53327	< 0.0001

Table 5.4.
Univariate analysis in Quasipoisson regression model with the combined bike rental as the response variable using the full sample, $n = 8,645$.

	Estimate	SE	Z	P-value
Intercept	4.93486	0.01177	419.19987	< 0.0001
Summer	0.12555	0.02227	5.63736	< 0.0001
Intercept	4.85643	0.01207	402.29876	< 0.0001
Fall	0.37652	0.02078	18.12201	< 0.0001
Intercept	4.94758	0.01165	424.60922	< 0.0001
Winter	0.08174	0.02275	3.59326	0.00033
Intercept	4.95224	0.01794	276.07754	< 0.0001
Workingday	0.02352	0.02161	1.08826	0.27651
Intercept	3.43845	0.02818	122.00062	< 0.0001
Daytime	1.85281	0.02924	63.36785	< 0.0001
Intercept	4.98812	0.01147	434.85772	< 0.0001
W2_cloudy	-0.07921	0.02333	-3.39535	0.00069
Intercept	5.00734	0.01020	491.08878	< 0.0001
W3_rain	-0.54885	0.04375	-12.54628	< 0.0001
Intercept	3.80265	0.02810	135.30674	< 0.0001
Temp	0.42250	0.00898	47.03470	< 0.0001
Intercept	5.80008	0.03029	191.50467	< 0.0001
Hum	-0.25999	0.00938	-27.72792	< 0.0001
Intercept	4.84471	0.01881	257.52297	< 0.0001
Windspeed	0.07719	0.00971	7.95262	< 0.0001

Table 5.5.
Univariate analysis in Negative Binomial regression model with the casual bike rental as the response variable using the full sample, $n = 8,645$.

	Estimate	SE	Z	P-value
Intercept	3.27112	0.01708	191.56465	< 0.0001
Summer	0.29016	0.03376	8.59423	< 0.0001
Intercept	3.16550	0.01742	181.73144	< 0.0001
Fall	0.58663	0.03410	17.20469	< 0.0001
Intercept	3.39663	0.01688	201.19889	< 0.0001
Winter	-0.18787	0.03403	-5.52119	< 0.0001
Intercept	3.85326	0.02147	179.46812	< 0.0001
Workingday	-0.85639	0.02606	-32.86379	< 0.0001
Intercept	1.68446	0.02393	70.37895	< 0.0001
Daytime	2.00169	0.02871	69.71907	< 0.0001
Intercept	3.40448	0.01684	202.13835	< 0.0001
W2_cloudy	-0.21564	0.03331	-6.47470	< 0.0001
Intercept	3.40818	0.01537	221.69212	< 0.0001
W3_rain	-0.88912	0.05170	-17.19623	< 0.0001
Intercept	0.95319	0.03731	25.54995	< 0.0001
Temp	0.83235	0.01347	61.79263	< 0.0001
Intercept	4.92418	0.04851	101.50140	< 0.0001
Hum	-0.50304	0.01396	-36.02495	< 0.0001
Intercept	3.20047	0.02715	117.89963	< 0.0001
Windspeed	0.09515	0.01459	6.52211	< 0.0001

Table 5.6.
Univariate analysis in Negative Binomial regression model with the registered bike rental as the response variable using the full sample, $n = 8,645$.

	Estimate	SE	Z	P-value
Intercept	4.72485	0.01184	398.94777	< 0.0001
Summer	0.08284	0.02345	3.53180	0.00041
Intercept	4.65266	0.01181	393.82080	< 0.0001
Fall	0.32222	0.02318	13.89859	< 0.0001
Intercept	4.70926	0.01179	399.37898	< 0.0001
Winter	0.14340	0.02372	6.04516	< 0.0001
Intercept	4.54697	0.01767	257.26277	< 0.0001
Workingday	0.28004	0.02136	13.10794	< 0.0001
Intercept	3.24840	0.01628	199.53840	< 0.0001
Daytime	1.81863	0.01971	92.27911	< 0.0001
Intercept	4.75842	0.01184	401.93259	< 0.0001
W2_cloudy	-0.04681	0.02338	-2.00249	0.04526
Intercept	4.78161	0.01074	445.09665	< 0.0001
W3_rain	-0.47836	0.03581	-13.35798	< 0.0001
Intercept	3.69587	0.02642	139.87948	< 0.0001
Temp	0.38514	0.00963	40.00190	< 0.0001
Intercept	5.56386	0.03580	155.40347	< 0.0001
Hum	-0.25402	0.01028	-24.72116	< 0.0001
Intercept	4.61681	0.01892	243.98390	< 0.0001
Windspeed	0.08107	0.01018	7.96779	< 0.0001

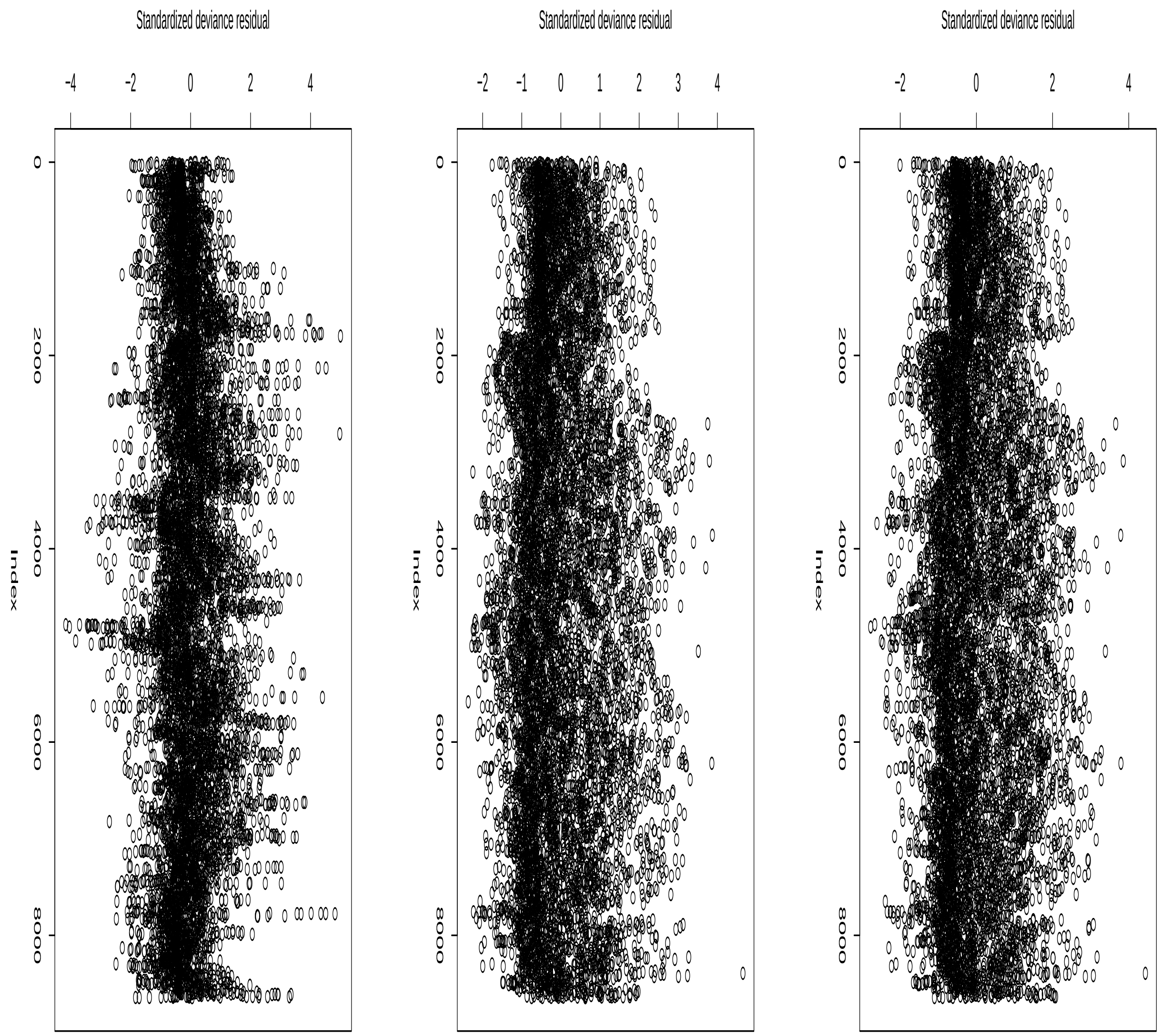
Table 5.7.
Univariate analysis in Quasipoisson regression model with the combined bike rental as the response variable using the full sample, $n = 8,645$.

	Estimate	SE	Z	P-value
Intercept	4.93486	0.01160	425.46044	< 0.0001
Summer	0.12555	0.02297	5.46624	< 0.0001
Intercept	4.85643	0.01158	419.45794	< 0.0001
Fall	0.37652	0.02272	16.57128	< 0.0001
Intercept	4.94758	0.01155	428.46837	< 0.0001
Winter	0.08174	0.02324	3.51796	0.00044
Intercept	4.95224	0.01779	278.30409	< 0.0001
Workingday	0.02352	0.02152	1.09303	0.27441
Intercept	3.43845	0.01589	216.43345	< 0.0001
Daytime	1.85281	0.01925	96.25135	< 0.0001
Intercept	4.98812	0.01157	430.94757	< 0.0001
W2_cloudy	-0.07921	0.02286	-3.46555	0.00053
Intercept	5.00734	0.01052	475.83106	< 0.0001
W3_rain	-0.54885	0.03508	-15.64651	< 0.0001
Intercept	3.71539	0.02519	147.47580	< 0.0001
Temp	0.45419	0.00918	49.48160	< 0.0001
Intercept	5.92241	0.03460	171.16480	< 0.0001
Hum	-0.29828	0.00993	-30.03632	< 0.0001
Intercept	4.83398	0.01853	260.84706	< 0.0001
Windspeed	0.08388	0.00997	8.41675	< 0.0001

Table 5.8.
Durbin-Watson test for autocorrelation with the casual bike rental, the registered bike rental, and the combined bike rental as response variable.

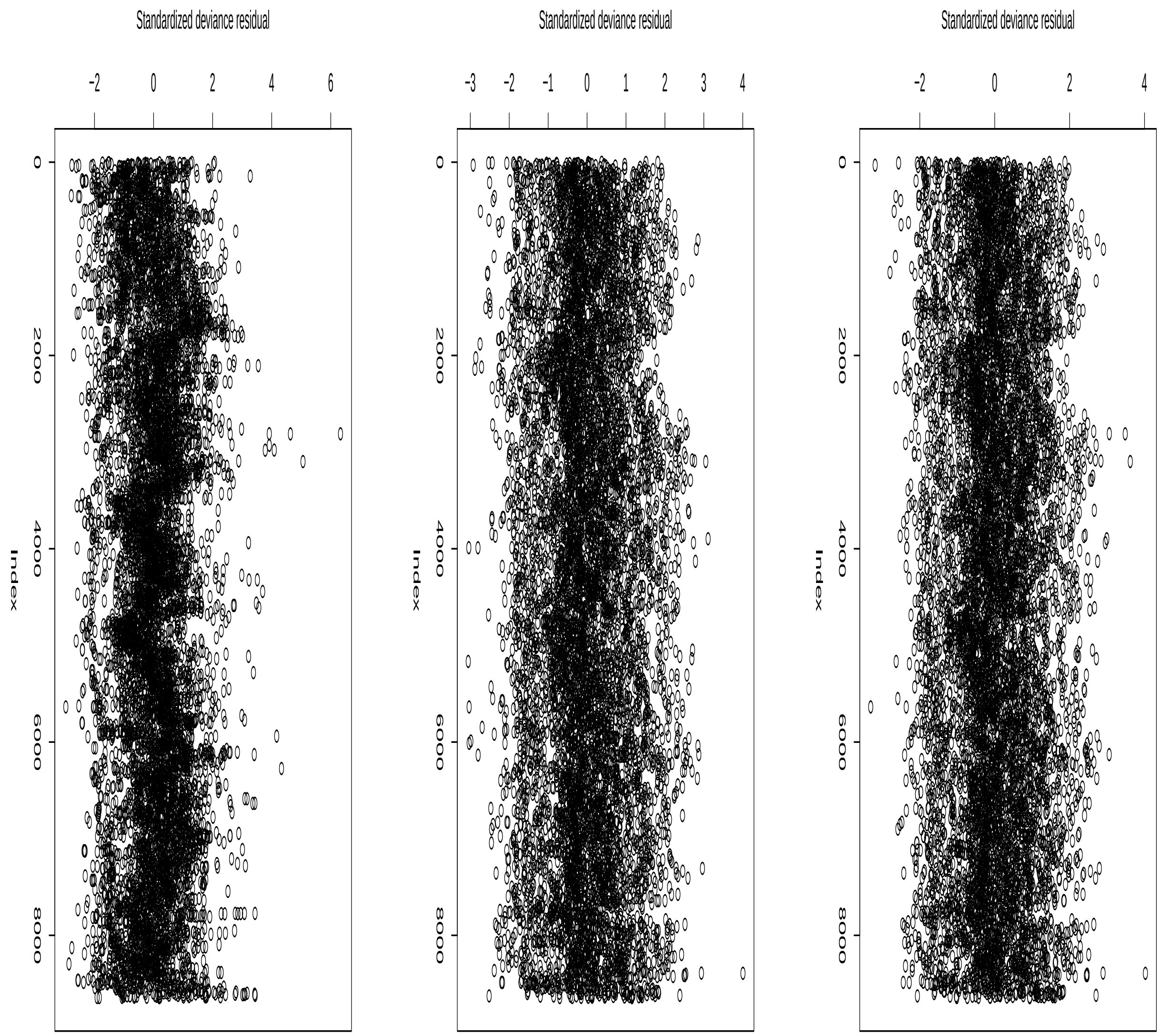
$H_0 : \rho = 0 \quad vs \quad H_1 : \rho \neq 0$			
Response	Test Statistics	P-value	Decision
Daytime indicator model			
Casual	0.3219	0.7603	Fail to reject
Register	0.7723	0.7521	Fail to reject
Combined	0.6952	0.7369	Fail to reject
24 hour indicator model			
Casual	0.4683	0.8432	Fail to reject
Register	0.8914	0.8320	Fail to reject
Combined	0.8571	0.8051	Fail to reject

To analyze this data set, researchers suggested that it is appropriate to assume independence among observations, see Fanaee-T, *et al.* (2013). We also conduct DurbinWatson test to see if there are auto correlations. The test results are reported in Table 5.8. The test suggests no auto-correlation for any of the 3 responses. We also visually examined the residual plots if there are some trends and auto correlations in Figures 5.1-5.2, suggesting no auto-correlation or obvious trends.



(a) Casual (b) Registered (c) Combined

Figure 5.1. Standarded deviance residuals in Quasipoisson regression model with the casual bike rental, the registered bike rental, and the combined bike rental as response variable.



(a) Casual (b) Registered (c) Combined

Figure 5.2. Standarded deviance residuals for Negative Binomial regression model with the casual bike rental, the registered bike rental, and the combined bike rental as response variable.

5.3 Model Fitting

We fit Poisson, Quasipoisson, and Negative Binomial regression models to each response. Results are given in Tables 5.9-5.11.

Table 5.9.
The estimates, standard errors, and P-values based on Poisson, Quasipoisson, and Negative Binomial regression. The response variable is the casual bike rental using the full sample, $n = 8,645$.

	Poisson	SE	P-value	Quasipoisson	SE	P-value	NB	SE	P-value
Intercept	1.22868	0.01532	< 0.0001	1.22868	0.05250	< 0.0001	0.85056	0.05519	< 0.0001
Summer	0.58595	0.00948	< 0.0001	0.58595	0.03248	< 0.0001	0.52711	0.03450	< 0.0001
Fall	0.34977	0.01080	< 0.0001	0.34977	0.03702	< 0.0001	0.16412	0.04313	0.00014
Winter	0.61975	0.00874	< 0.0001	0.61975	0.02994	< 0.0001	0.49915	0.03095	< 0.0001
Workingdays	-0.91764	0.00405	< 0.0001	-0.91764	0.01388	< 0.0001	-0.84928	0.01959	< 0.0001
Daytime	1.62495	0.00870	< 0.0001	1.62495	0.02981	< 0.0001	1.60662	0.02280	< 0.0001
W2_cloudy	0.01800	0.00520	0.00054	0.01800	0.01783	0.31261	0.01413	0.02262	0.53227
W3_rain	-0.38293	0.01108	< 0.0001	-0.38293	0.03796	< 0.0001	-0.45920	0.03789	< 0.0001
Temp	0.52674	0.00340	< 0.0001	0.52674	0.01164	< 0.0001	0.73047	0.01505	< 0.0001
Hum	-0.20214	0.00251	< 0.0001	-0.20214	0.00859	< 0.0001	-0.22215	0.01126	< 0.0001
Windspeed	0.00355	0.00210	0.09147	0.00355	0.00721	0.62237	-0.03622	0.00972	0.00019

Table 5.10.
The estimates, standard errors, and P-values based on Poisson, Quasipoisson, and Negative Binomial regression. The response variable is the registered bike rental using the full sample, $n = 8,645$.

	Poisson	SE	P-value	Quasipoisson	SE	P-value	NB	SE	P-value
Intercept	2.39368	0.00715	< 0.0001	2.39368	0.04649	< 0.0001	2.38957	0.04766	< 0.0001
Summer	0.43961	0.00422	< 0.0001	0.43961	0.02747	< 0.0001	0.38515	0.02982	< 0.0001
Fall	0.43632	0.00501	< 0.0001	0.43632	0.03260	< 0.0001	0.37417	0.03790	< 0.0001
Winter	0.62715	0.00376	< 0.0001	0.62715	0.02445	< 0.0001	0.60795	0.02655	< 0.0001
Workingdays	0.26597	0.00230	< 0.0001	0.26597	0.01496	< 0.0001	0.19477	0.01751	< 0.0001
Daytime	1.73391	0.00400	< 0.0001	1.73391	0.02605	< 0.0001	1.71301	0.01928	< 0.0001
W2_cloudy	-0.05769	0.00252	< 0.0001	-0.05769	0.01637	0.00043	-0.02610	0.01995	0.19062
W3_rain	-0.44990	0.00473	< 0.0001	-0.44990	0.03079	< 0.0001	-0.45104	0.03220	< 0.0001
Temp	0.18323	0.00164	< 0.0001	0.18323	0.01067	< 0.0001	0.24903	0.01325	< 0.0001
Hum	-0.03787	0.00124	< 0.0001	-0.03787	0.00804	< 0.0001	-0.05594	0.00997	< 0.0001
Windspeed	-0.00375	0.00105	0.00036	-0.00375	0.00684	0.58354	-0.01609	0.00859	0.06096

Table 5.11.
The estimates, standard errors, and P-values based on Poisson, Quasipoisson, and Negative Binomial regression. The response variable is the combined bike rental using the full sample, $n = 8,645$.

	Poisson	SE	P-value	Quasipoisson	SE	P-value	NB	SE	P-value
Intercept	2.72555	0.00644	< 0.0001	2.72555	0.04396	< 0.0001	2.62116	0.04524	< 0.0001
Summer	0.45425	0.00384	< 0.0001	0.45425	0.02624	< 0.0001	0.39737	0.02832	< 0.0001
Fall	0.40304	0.00452	< 0.0001	0.40304	0.03088	< 0.0001	0.34114	0.03598	< 0.0001
Winter	0.61406	0.00344	< 0.0001	0.61406	0.02351	< 0.0001	0.58316	0.02522	< 0.0001
Workingdays	0.00058	0.00195	0.76516	0.00058	0.01330	0.96510	-0.00506	0.01661	0.76068
Daytime	1.71376	0.00364	< 0.0001	1.71376	0.02482	< 0.0001	1.72794	0.01830	< 0.0001
W2_cloudy	-0.03973	0.00226	< 0.0001	-0.03973	0.01546	0.01018	-0.02622	0.01894	0.16629
W3_rain	-0.42498	0.00434	< 0.0001	-0.42498	0.02964	< 0.0001	-0.45953	0.03059	< 0.0001
Temp	0.25366	0.00147	< 0.0001	0.25366	0.01005	< 0.0001	0.32049	0.01258	< 0.0001
Hum	-0.07065	0.00111	< 0.0001	-0.07065	0.00755	< 0.0001	-0.07773	0.00947	< 0.0001
Windspeed	-0.00560	0.00094	< 0.0001	-0.00560	0.00642	0.38298	-0.02057	0.00815	0.01165

5.4 Conclusions

We find the full sample parameter estimates for Poisson and Quasipoisson regression models are the same, but the standard errors are different. Compared to Poisson and Quasipoisson, the Negative Binomial regression models have different parameter estimates and different standard errors. The parameter estimates of *workingday* variable for casual response is negative and for registered response is positive. This may be due to the reason that casual bike rentals are related to tourists on weekend and holiday, registered bike rentals are related to people using rental bikes as transportation tool to go to work on working days. The *workingday* variable is significant in casual and registered data model, but not significant in combined data model.

6. A-OPTIMAL SUBSAMPLING FOR REAL DATA ANALYSIS: BIKE SHARING DATA

We now apply our proposed subsampling methods to conduct Quasipoisson regression and Negative Binomial regression. The subsampling probabilities are caculated as follows:

$$\hat{\pi}_i^{(2)} = \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\||\hat{e}_i|},$$
$$\hat{\pi}_i^{(1)} = \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\||\hat{e}_i|},$$
$$\hat{\pi}_i^{(0)} = \frac{\|\mathbf{x}_i\||\hat{e}_i|}{\sum_{i=1}^n \|\mathbf{x}_i\||\hat{e}_i|},$$
$$\bar{\pi}_i^{(2)} = \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-1}\mathbf{x}_i\|\hat{g}_i},$$
$$\bar{\pi}_i^{(1)} = \frac{\|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|(\mathbf{X}^\top \mathbf{W}(\hat{\boldsymbol{\beta}})\mathbf{X})^{-\frac{1}{2}}\mathbf{x}_i\|\hat{g}_i},$$
$$\bar{\pi}_i^{(0)} = \frac{\|\mathbf{x}_i\|\hat{g}_i}{\sum_{i=1}^n \|\mathbf{x}_i\|\hat{g}_i}, \quad i = 1, \cdots, n.$$

For Quasioisson regression,

$$W(\hat{\boldsymbol{\beta}}) = \mathrm{Diag}(\frac{\hat{\mu}_i}{\hat{\phi}}), \quad \hat{\mu}_i = \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}),$$
$$\hat{e}_i = \frac{y_i - \hat{\mu}_i}{\hat{\phi}}, \quad \hat{g}_i = \sqrt{\frac{\hat{\mu}_i}{\hat{\phi}}} \quad i = 1, \cdots, n.$$

For Negative Binomial regression,

$$W(\hat{\boldsymbol{\beta}}) = \mathrm{Diag}(\frac{\hat{\mu}_i}{1+\hat{\alpha}\hat{\mu}_i}), \quad \hat{\mu}_i = \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}),$$
$$\hat{e}_i = \frac{y_i - \hat{\mu}_i}{1+\hat{\alpha}\hat{\mu}_i}, \quad \hat{g}_i = \sqrt{\frac{\hat{\mu}_i}{1+\hat{\alpha}\hat{\mu}_i}} \quad i = 1, \cdots, n,$$

where $\hat{\phi}$ and $\hat{\alpha}$ are the moment estimators based on full sample. We take a uniform subsample of size $r_0 = 200$ to get an initial estimates $\hat{\boldsymbol{\beta}}^*_{r_0}$; then plug in to the formulas to approximate the A-optimal subsampling distributions; take subsample of size r according to the approximate distributions to get $\hat{\boldsymbol{\beta}}_r^*$; We calculate the empirical mean square errors for different subsample sizes r for each of $B = 1000$ subsamples using the formula:

$$MSE = \frac{1}{B}\sum_{b=1}^{B} \|\hat{\boldsymbol{\beta}}_{r,b}^* - \hat{\boldsymbol{\beta}}\|^2,$$

where $\hat{\boldsymbol{\beta}}_{r,b}^*$ is the estimate from the b^{th} subsample with subsample size r.

6.1 Casual Bike Rentals

6.1.1 Quasipoisson Regression Model for Casual Data

Table 6.1.
Averaged estimates, theoretical standard errors(Tse), empirical standard errors(Ese), and P-values based on 1000 subsamples in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$				$\hat{\boldsymbol{\pi}}^{(1)}$				$\hat{\boldsymbol{\pi}}^{(0)}$			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	1.2297	0.1735	0.1748	< 0.001	1.2280	0.1859	0.1863	< 0.001	1.2294	0.2219	0.2217	< 0.001
Summer	0.5862	0.1091	0.1098	< 0.001	0.5850	0.1209	0.1195	< 0.001	0.5853	0.1417	0.1399	< 0.001
Fall	0.3495	0.1292	0.1286	0.0069	0.3503	0.1379	0.1385	0.0111	0.3502	0.1560	0.1551	0.0248
Winter	0.6195	0.1032	0.1012	< 0.001	0.6192	0.1117	0.1130	< 0.001	0.6197	0.1323	0.1323	< 0.001
Workingdays	-0.9169	0.0575	0.0592	< 0.001	-0.9185	0.0501	0.0486	< 0.001	-0.9166	0.0486	0.0481	< 0.001
Daytime	1.6246	0.1025	0.1034	< 0.001	1.6252	0.1107	0.1100	< 0.001	1.6245	0.1315	0.1329	< 0.001
W2_cloudy	0.0190	0.0672	0.0662	0.7777	0.0189	0.0608	0.0606	0.7554	0.0183	0.0617	0.0606	0.7672
W3_rain	-0.3821	0.1169	0.1169	0.0011	-0.3839	0.1244	0.1232	0.0020	-0.3837	0.1501	0.1516	0.0105
Temp	0.5273	0.0432	0.0413	< 0.001	0.5258	0.0417	0.0430	< 0.001	0.5262	0.0448	0.0450	< 0.001
Hum	-0.2028	0.0317	0.0302	< 0.001	-0.2014	0.0297	0.0310	< 0.001	-0.2011	0.0308	0.0311	< 0.001
Windspeed	0.0036	0.0273	0.0284	0.8946	0.0040	0.0243	0.0234	0.8690	0.0038	0.0251	0.0267	0.8795

Table 6.1, continued

	$\bar{\boldsymbol{\pi}}^{(2)}$				$\bar{\boldsymbol{\pi}}^{(1)}$				$\bar{\boldsymbol{\pi}}^{(0)}$				unif			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	1.2297	0.2310	0.2308	< 0.001	1.2278	0.2408	0.2421	< 0.001	1.2295	0.2728	0.2735	< 0.001	1.2288	0.2732	0.2722	< 0.001
Summer	0.5860	0.1575	0.1570	0.0002	0.5855	0.1737	0.1744	0.0008	0.5852	0.1975	0.1986	0.0030	0.5864	0.2054	0.2069	0.0043
Fall	0.3490	0.1875	0.1864	0.0627	0.3504	0.1994	0.1994	0.0788	0.3504	0.2196	0.2180	0.1106	0.3488	0.2378	0.2378	0.1424
Winter	0.6204	0.1458	0.1465	< 0.001	0.6196	0.1559	0.1571	0.0001	0.6203	0.1787	0.1778	0.0005	0.6195	0.1803	0.1785	0.0006
Workingdays	-0.9176	0.0807	0.0824	< 0.001	-0.9174	0.0688	0.0680	< 0.001	-0.9182	0.0647	0.0661	< 0.001	-0.9175	0.0820	0.0802	< 0.001
Daytime	1.6242	0.1294	0.1295	< 0.001	1.6243	0.1358	0.1343	< 0.001	1.6256	0.1535	0.1535	< 0.001	1.6244	0.1286	0.1304	< 0.001
W2_cloudy	0.0172	0.0901	0.0887	0.8484	0.0189	0.0798	0.0779	0.8128	0.0187	0.0781	0.0764	0.8108	0.0189	0.0950	0.0943	0.8423
W3_rain	-0.3828	0.1617	0.1624	0.0179	-0.3838	0.1687	0.1681	0.0229	-0.3831	0.1963	0.1960	0.0510	-0.3832	0.2245	0.2255	0.0878
Temp	0.5277	0.0616	0.0632	< 0.001	0.5277	0.0578	0.0593	< 0.001	0.5260	0.0592	0.0603	< 0.001	0.5268	0.0722	0.0723	< 0.001
Hum	-0.2020	0.0443	0.0440	< 0.001	-0.2020	0.0405	0.0410	< 0.001	-0.2029	0.0401	0.0385	< 0.001	-0.2025	0.0498	0.0479	< 0.001
Windspeed	0.0031	0.0377	0.0390	0.9335	0.0034	0.0332	0.0320	0.9176	0.0036	0.0336	0.0338	0.9154	0.0037	0.0425	0.0431	0.9306

All the P-values are based on theoretical standard errors in this thesis. In the above model, the Fall season's rental effect is detected by all the $\hat{\boldsymbol{\pi}}^{(k)}$ methods, but not detected by any of the $\bar{\boldsymbol{\pi}}^{(k)}$ methods or the uniform method at 0.05 level. Our common sense suggests that the casual rentals may decrease in rainy days. Such decreased rainy day effect is detected by all $\hat{\boldsymbol{\pi}}^{(k)}$, most of the $\bar{\boldsymbol{\pi}}^{(k)}$, but not by the uniform at 0.05 level. Proposed methods decreased standard errors, which in turn increased the power of tests to detect variables' effects.

Table 6.2.
The length ratios of the 95% confidence intervals of proposed subsampling to uniform subsampling in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$, subsample size $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.6349	0.6804	0.8121	0.8452	0.8814	0.9984
Summer	0.5312	0.5883	0.6898	0.7669	0.8457	0.9614
Fall	0.5435	0.5798	0.6562	0.7885	0.8384	0.9234
Winter	0.5724	0.6197	0.7339	0.8089	0.8648	0.9914
Workingdays	0.7005	0.6102	0.5929	0.9840	0.8387	0.7888
Daytime	0.7974	0.8609	1.0227	1.0068	1.0564	1.1937
W2_cloudy	0.7076	0.6400	0.6495	0.9489	0.8407	0.8219
W3_rain	0.5209	0.5543	0.6685	0.7203	0.7516	0.8746
Temp	0.5981	0.5775	0.6209	0.8541	0.8007	0.8203
Hum	0.6374	0.5972	0.6183	0.8907	0.8139	0.8066
Windspeed	0.6426	0.5712	0.5913	0.8863	0.7804	0.7898

Table 6.3.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	unif	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.9470	0.9436	0.9553	0.9484	0.9537	0.9517	0.9515
Summer	0.9440	0.9433	0.9583	0.9423	0.9466	0.9573	0.9411
Fall	0.9562	0.9433	0.9382	0.9491	0.9524	0.9387	0.9510
Winter	0.9421	0.9503	0.9353	0.9419	0.9392	0.9476	0.9409
Workingdays	0.9553	0.9335	0.9326	0.9373	0.9534	0.9510	0.9375
Daytime	0.9472	0.9473	0.9521	0.9505	0.9467	0.9522	0.9437
W2_cloudy	0.9522	0.9496	0.9565	0.9440	0.9398	0.9469	0.9387
W3_rain	0.9477	0.9555	0.9439	0.9357	0.9416	0.9531	0.9478
Temp	0.9398	0.9565	0.9448	0.9535	0.9550	0.9500	0.9465
Hum	0.9505	0.9434	0.9348	0.9525	0.9439	0.9390	0.9375
Windspeed	0.9511	0.9469	0.9485	0.9364	0.9424	0.9438	0.9380

Table 6.2 contains the confidence interval length ratios of the proposed methods versus uniform method. Table 6.3 reports the percentages that the confidence interval catches the corresponding full sample MLE. Table 6.4 reports the summer effect ($\hat{\beta}_2$) and examines the change in percentage that the confidence interval catches the full sample MLE $\hat{\beta}_2$ when subsample size r increases. Figure 6.1 is the plot of Table 6.4. First of all, most of the values in Table 6.2 is less than 1, indicating that the 95% confidence intervals constructed by our proposed methods have shorter length compared to uniform subsampling. Second, Table 6.4 shows that the coverage probabilities of proposed methods achieve the nominal 95% in both small and large subsample sizes. That is, our methods are more accurate while maintaining the nominal coverage probability.

Table 6.4.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$ in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
unif	0.9428	0.9440	0.9474	0.9391	0.9396	0.9423
$\hat{\boldsymbol{\pi}}^{(2)}$	0.9467	0.9433	0.9578	0.9553	0.9406	0.9478
$\hat{\boldsymbol{\pi}}^{(1)}$	0.9451	0.9583	0.9363	0.9465	0.9492	0.9530
$\hat{\boldsymbol{\pi}}^{(0)}$	0.9470	0.9423	0.9477	0.9520	0.9540	0.9484
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9434	0.9466	0.9332	0.9528	0.9424	0.9530
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9397	0.9573	0.9461	0.9373	0.9372	0.9382
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9472	0.9411	0.9364	0.9416	0.9361	0.9480

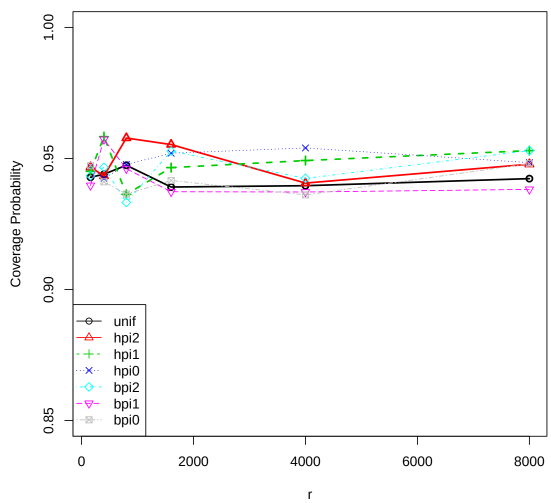


Figure 6.1. Simulated percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$ in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.5.
MSE ratios of the proposed subsampling to uniform subsampling in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
$\hat{\boldsymbol{\pi}}^{(2)}$	0.3556	0.3567	0.3578	0.3827	0.4173	0.3739
$\hat{\boldsymbol{\pi}}^{(1)}$	0.3961	0.4033	0.4020	0.4298	0.3955	0.4144
$\hat{\boldsymbol{\pi}}^{(0)}$	0.5393	0.5409	0.5550	0.5712	0.6105	0.6278
$\bar{\boldsymbol{\pi}}^{(2)}$	0.6717	0.6711	0.6771	0.6727	0.7308	0.7945
$\bar{\boldsymbol{\pi}}^{(1)}$	0.7252	0.7268	0.7333	0.7389	0.7330	0.8532
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9156	0.9160	0.9287	0.9416	0.9254	0.9062

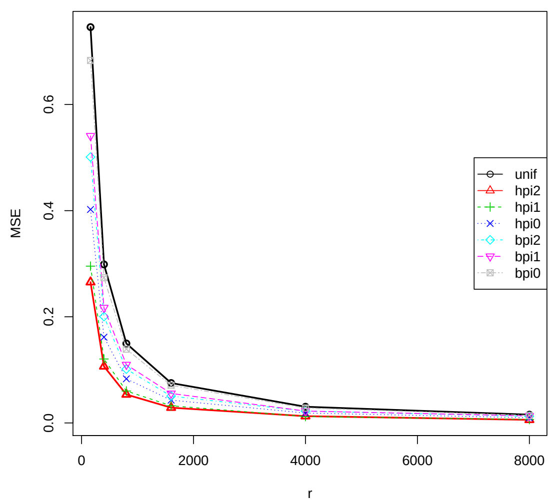


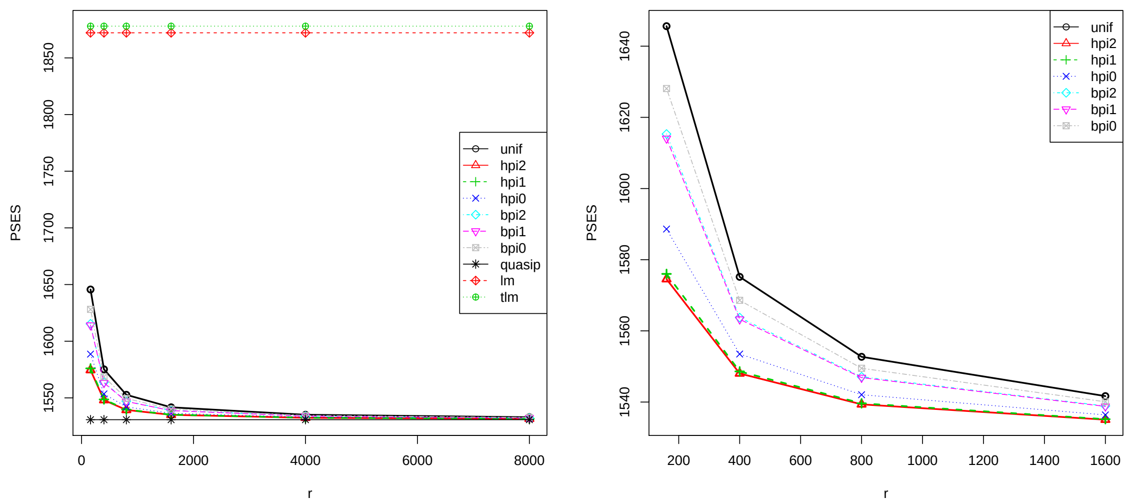
Figure 6.2. MSE plots in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.5 contains the MSE ratio of the proposed sampling methods to uniform sampling for different subsample sizes r. In Table 6.5, all the values are less than

1, indicating our proposed sampling distributions produce smaller MSE than the uniform sampling. The highest reduction is about 65%. $\hat{\boldsymbol{\pi}}^{(k)}$ produce smaller MSEs than $\bar{\boldsymbol{\pi}}^{(k)}$, $\hat{\boldsymbol{\pi}}^{(2)}$ is the best among $\hat{\boldsymbol{\pi}}^{(k)}$, and $\bar{\boldsymbol{\pi}}^{(2)}$ is the best among $\bar{\boldsymbol{\pi}}^{(k)}$, $k = 0, 1, 2$. Figure 6.2 is the MSE plots as subsample size r increases. In Figure 6.2, we found that the largest MSE's over different subsample sizes are given by the uniform methods, the smallest MSE's over different subsample sizes are given by the $\hat{\boldsymbol{\pi}}^{(2)}$ sampling.

Table 6.6.
Averages of the sum of squared predicted errors in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$. The sum of the squared prediction errors are 1,530.7560, 1,872.1331, and 1,877.9969 for the full sample Quasipoisson, linear regression and the log-transformed linear regression respectively.

r	160	400	800	1600	4000	8000
unif	1645.6281	1575.1841	1552.7189	1541.6749	1535.1085	1532.9298
$\hat{\boldsymbol{\pi}}^{(2)}$	1574.5322	1548.0304	1539.3540	1535.0452	1532.4693	1531.6122
$\hat{\boldsymbol{\pi}}^{(1)}$	1576.0130	1548.6081	1539.6404	1535.1878	1532.5262	1531.6407
$\hat{\boldsymbol{\pi}}^{(0)}$	1588.6148	1553.4949	1542.0583	1536.3904	1533.0057	1531.8802
$\bar{\boldsymbol{\pi}}^{(2)}$	1615.3228	1563.7336	1547.1043	1538.8951	1534.0032	1532.3782
$\bar{\boldsymbol{\pi}}^{(1)}$	1614.0602	1563.2572	1546.8707	1538.7795	1533.9573	1532.3553
$\bar{\boldsymbol{\pi}}^{(0)}$	1628.0999	1568.5875	1549.4888	1540.0768	1534.4734	1532.6129



(a) (b)

Figure 6.3. Averages of the sum of squared predicted errors and the first several averages of the sum of squared predicted errors plot in Quasipoisson regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

The formula to calculate the averages of the sum of squared predicted errors is as follows:

For full sample Quasipossion regression:

$$PSES = \frac{1}{n}\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i^\top \hat{\beta}_{qp}))^2.$$

For full sample Negative Binomial regression:

$$PSES = \frac{1}{n}\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i^\top \hat{\beta}_{nb}))^2.$$

For full sample linear regression model:

$$PSES = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \hat{\beta}_{ols})^2.$$

For full sample log-transformed linear regression model:

$$PSES = \frac{1}{n}\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i^\top \hat{\beta}_{tols}))^2.$$

For subsample Quasipossion and Negative Binomial regression:

$$PSES = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i^\top \hat{\beta}^*_{r,b}))^2.$$

Table 6.6 summarizes squared prediction errors using validation data. Here the transformed linear regression takes logarithm on the response variable, then use transformed response to fit the linear regression model. Table 6.6 shows that transformed linear regression model generates the largest full sample prediction error. The linear regression model is the second largest, indicating that using linear regression models for count data is not as good as Quasipoisson regression model in this example.

Figures 6.3 are based on Table 6.6. They show that when subsample size r increases, the averages of sum of squared prediction errors of different methods will converge to that of the full data Quasipoisson model. The uniform sampling generates the largest prediction errors for different r; the proposed $\hat{\boldsymbol{\pi}}^{(2)}$ sampling generates the smallest prediction errors.

6.1.2 Negative Binomial Regression Model for Casual Data

Table 6.7.
Averaged estimates, theoretical standard errors(Tse), empirical standard errors(Ese), and P-values based on 1000 subsamples in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$				$\hat{\boldsymbol{\pi}}^{(1)}$				$\hat{\boldsymbol{\pi}}^{(0)}$			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	0.8509	0.1654	0.1645	< 0.001	0.8506	0.1771	0.1768	< 0.001	0.8503	0.1999	0.2000	< 0.001
Summer	0.5279	0.1097	0.1092	< 0.001	0.5263	0.1168	0.1163	< 0.001	0.5270	0.1254	0.1269	< 0.001
Fall	0.1633	0.1381	0.1390	0.2372	0.1650	0.1451	0.1459	0.2556	0.1645	0.1515	0.1519	0.2776
Winter	0.4994	0.1024	0.1035	< 0.001	0.4991	0.1062	0.1058	< 0.001	0.4991	0.1133	0.1145	< 0.001
Workingdays	-0.8494	0.0690	0.0707	< 0.001	-0.8485	0.0656	0.0643	< 0.001	-0.8494	0.0668	0.0658	< 0.001
Daytime	1.6065	0.0849	0.0856	< 0.001	1.6068	0.0816	0.0831	< 0.001	1.6071	0.0837	0.0839	< 0.001
W2_cloudy	0.0144	0.0763	0.0743	0.8500	0.0138	0.0732	0.0748	0.8505	0.0147	0.0746	0.0759	0.8433
W3_rain	-0.4591	0.1188	0.1175	0.0001	-0.4590	0.1217	0.1211	0.0002	-0.4585	0.1330	0.1326	0.0006
Temp	0.7308	0.0475	0.0468	< 0.001	0.7301	0.0482	0.0498	< 0.001	0.7304	0.0510	0.0523	< 0.001
Hum	-0.2214	0.0355	0.0349	< 0.001	-0.2223	0.0353	0.0361	< 0.001	-0.2221	0.0380	0.0384	< 0.001
Windspeed	-0.0354	0.0315	0.0318	0.2619	-0.0365	0.0301	0.0309	0.2244	-0.0363	0.0315	0.0307	0.2493

	$\bar{\boldsymbol{\pi}}^{(2)}$				$\bar{\boldsymbol{\pi}}^{(1)}$				$\bar{\boldsymbol{\pi}}^{(0)}$				unif			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	0.8507	0.2155	0.2174	0.0001	0.8509	0.2321	0.2323	0.0002	0.8499	0.2633	0.2647	0.0012	0.8498	0.2360	0.2373	0.0003
Summer	0.5274	0.1595	0.1576	0.0009	0.5268	0.1702	0.1707	0.0020	0.5263	0.1836	0.1852	0.0041	0.5279	0.1736	0.1754	0.0024
Fall	0.1644	0.1914	0.1897	0.3904	0.1651	0.2048	0.2052	0.4203	0.1643	0.2172	0.2178	0.4494	0.1647	0.2101	0.2109	0.4332
Winter	0.4991	0.1383	0.1402	0.0003	0.4984	0.1456	0.1449	0.0006	0.4983	0.1571	0.1559	0.0015	0.4987	0.1452	0.1461	0.0006
Workingdays	-0.8498	0.0940	0.0958	< 0.001	-0.8493	0.0903	0.0915	< 0.001	-0.8485	0.0930	0.0940	< 0.001	-0.8486	0.0910	0.0903	< 0.001
Daytime	1.6073	0.1270	0.1266	< 0.001	1.6069	0.1218	0.1218	< 0.001	1.6061	0.1247	0.1260	< 0.001	1.6071	0.1198	0.1187	< 0.001
W2_cloudy	0.0149	0.1032	0.1049	0.8848	0.0133	0.0998	0.1015	0.8943	0.0135	0.1025	0.1029	0.8951	0.0149	0.0995	0.1001	0.8808
W3_rain	-0.4588	0.1741	0.1747	0.0084	-0.4587	0.1797	0.1786	0.0107	-0.4588	0.1979	0.1963	0.0204	-0.4594	0.2007	0.1997	0.0221
Temp	0.7306	0.0609	0.0600	< 0.001	0.7295	0.0631	0.0648	< 0.001	0.7311	0.0675	0.0672	< 0.001	0.7311	0.0657	0.0676	< 0.001
Hum	-0.2229	0.0463	0.0466	< 0.001	-0.2217	0.0463	0.0479	< 0.001	-0.2215	0.0499	0.0500	< 0.001	-0.2229	0.0478	0.0479	< 0.001
Windspeed	-0.0360	0.0400	0.0418	0.3685	-0.0372	0.0391	0.0400	0.3421	-0.0364	0.0418	0.0402	0.3837	-0.0367	0.0409	0.0391	0.3699

Table 6.7 shows that all the subsampling methods can not detect the Fall effect, however, our proposed $\hat{\boldsymbol{\pi}}^{(2)}$ method has the smallest P-values.

Table 6.8.
The length ratios of the 95% confidence intervals of proposed subsampling to uniform subsampling in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$, subsample size $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.7008	0.7504	0.8468	0.9130	0.9836	1.1155
Summer	0.6318	0.6732	0.7226	0.9190	0.9808	1.0576
Fall	0.6574	0.6906	0.7211	0.9109	0.9747	1.0336
Winter	0.7056	0.7314	0.7803	0.9527	1.0030	1.0819
Workingdays	0.7586	0.7208	0.7337	1.0326	0.9929	1.0217
Daytime	0.7091	0.6815	0.6985	1.0602	1.0170	1.0408
W2_cloudy	0.7671	0.7364	0.7498	1.0370	1.0031	1.0303
W3_rain	0.5917	0.6064	0.6629	0.8676	0.8953	0.9859
Temp	0.7235	0.7331	0.7760	0.9266	0.9598	1.0267
Hum	0.7413	0.7379	0.7937	0.9685	0.9679	1.0427
Windspeed	0.7712	0.7348	0.7703	0.9779	0.9567	1.0229

Table 6.9.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	unif	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.9485	0.9466	0.9388	0.9383	0.9474	0.9380	0.9337
Summer	0.9317	0.9487	0.9497	0.9481	0.9517	0.9456	0.9508
Fall	0.9483	0.9351	0.9567	0.9470	0.9367	0.9398	0.9557
Winter	0.9385	0.9444	0.9560	0.9515	0.9486	0.9437	0.9392
Workingdays	0.9463	0.9352	0.9381	0.9408	0.9339	0.9554	0.9376
Daytime	0.9331	0.9360	0.9382	0.9468	0.9489	0.9379	0.9527
W2_cloudy	0.9453	0.9564	0.9534	0.9442	0.9474	0.9372	0.9553
W3_rain	0.9416	0.9525	0.9499	0.9501	0.9483	0.9374	0.9333
Temp	0.9431	0.9442	0.9473	0.9458	0.9494	0.9513	0.9472
Hum	0.9400	0.9455	0.9490	0.9426	0.9438	0.9498	0.9464
Windspeed	0.9395	0.9375	0.9345	0.9456	0.9504	0.9504	0.9347

Table 6.10.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$ in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
unif	0.9472	0.9317	0.9454	0.9494	0.9543	0.9495
$\hat{\boldsymbol{\pi}}^{(2)}$	0.9372	0.9487	0.9499	0.9441	0.9422	0.9493
$\hat{\boldsymbol{\pi}}^{(1)}$	0.9409	0.9497	0.9365	0.9495	0.9494	0.9448
$\hat{\boldsymbol{\pi}}^{(0)}$	0.9373	0.9481	0.9462	0.9434	0.9488	0.9510
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9395	0.9517	0.9294	0.9360	0.9418	0.9462
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9508	0.9456	0.9337	0.9514	0.9378	0.9442
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9358	0.9508	0.9474	0.9459	0.9397	0.9531

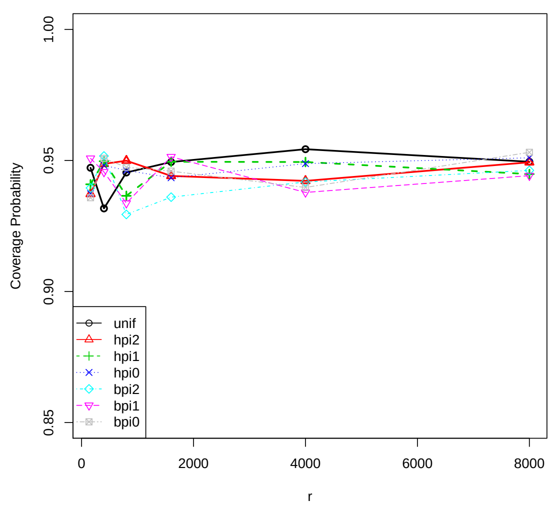


Figure 6.4. Simulated percentages of the 95% confidence intervals which caught the full sample MLE plot of $\hat{\beta}_2$ in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.11.
MSE ratios of the proposed subsampling to uniform subsampling in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
$\hat{\boldsymbol{\pi}}^{(2)}$	0.4568	0.4554	0.4576	0.4903	0.5632	0.5260
$\hat{\boldsymbol{\pi}}^{(1)}$	0.4901	0.4987	0.5014	0.4912	0.5848	0.5532
$\hat{\boldsymbol{\pi}}^{(0)}$	0.5669	0.5760	0.5702	0.5677	0.6101	0.5612
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8659	0.8720	0.8768	0.8848	0.9622	0.9322
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9476	0.9520	0.9567	0.9374	1.0483	1.1338
$\bar{\boldsymbol{\pi}}^{(0)}$	1.1107	1.1147	1.1243	1.1302	1.1674	1.2180

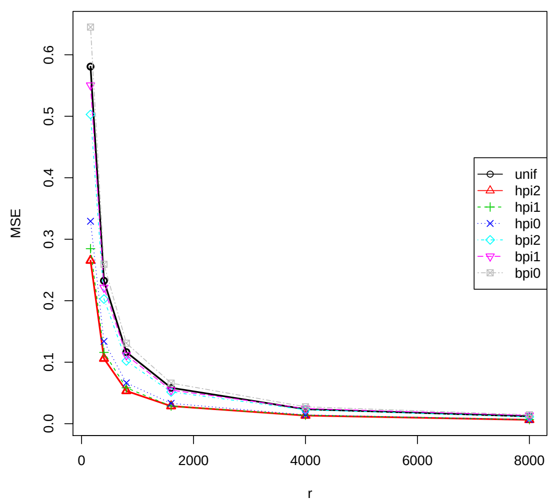
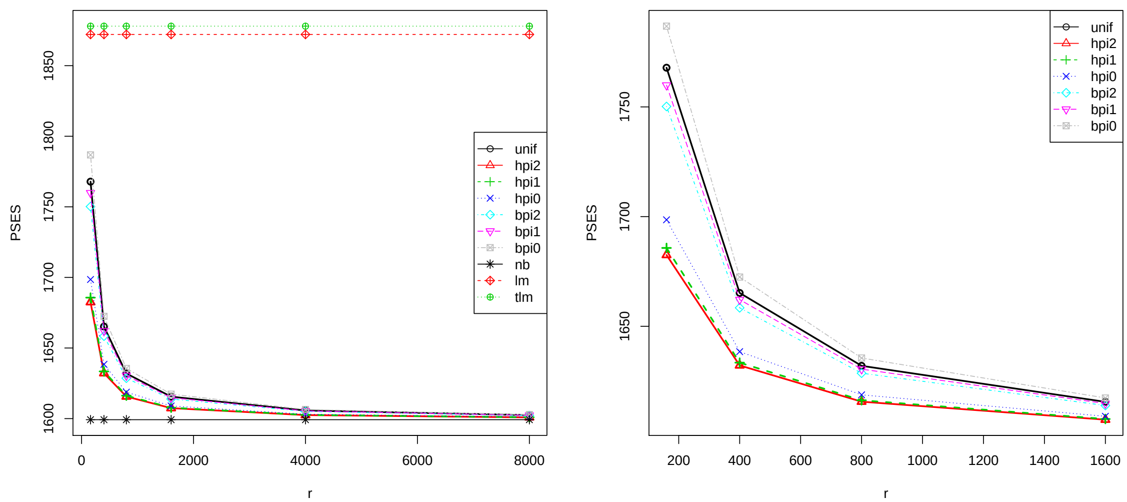


Figure 6.5. MSE plots in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.12.
Averages of the sum of squared predicted errors in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$. The sum of the squared prediction errors are 1,599.2348, 1,872.1331, and 1,877.9969 for the full sample Negative Binomial regression, linear regression and the log-transformed linear regression, respectively.

r	160	400	800	1600	4000	8000
unif	1767.8634	1665.2026	1631.9733	1615.5428	1605.7433	1602.4866
$\hat{\boldsymbol{\pi}}^{(2)}$	1682.4842	1632.1596	1615.6349	1607.4193	1602.5048	1600.8692
$\hat{\boldsymbol{\pi}}^{(1)}$	1685.7090	1633.4224	1616.2618	1607.7316	1602.6295	1600.9315
$\hat{\boldsymbol{\pi}}^{(0)}$	1698.5348	1638.4273	1618.7435	1608.9673	1603.1225	1601.1778
$\bar{\boldsymbol{\pi}}^{(2)}$	1750.1890	1658.4123	1628.6242	1613.8797	1605.0808	1602.1558
$\bar{\boldsymbol{\pi}}^{(1)}$	1759.8961	1662.1474	1630.4673	1614.7953	1605.4456	1602.3379
$\bar{\boldsymbol{\pi}}^{(0)}$	1786.8081	1672.4398	1635.5357	1617.3101	1606.4469	1602.8378



(a) (b)

Figure 6.6. Averages of the sum of squared predicted errors and the first several averages of the sum of squared predicted errors plot in Negative Binomial regression model with the casual bike rental as the response variable, the pre-subsample size $r_0 = 200$.

The behaviors of MSE ratios,the 95% confidence interval, coverage probabilities, prediction errors of our proposed sampling distributions in Negative Binomial regression model are similar to Quasipoisson regression model, but the $\bar{\boldsymbol{\pi}}^{(k)}$, $k = 0, 1, 2$, are not as good as in Quasipoisson regression model, $\hat{\boldsymbol{\pi}}^{(k)}$, $k = 0, 1, 2$, still work very well.

6.2 Registered Bike Rentals

In the below tables and figures, we report the results of our proposed method for the registered bike rental as response.

6.2.1 Quasipoisson Regression Model for Registered Data

Table 6.13.
Averaged estimates, theoretical standard errors(Tse), empirical standard errors(Ese), and P-values based on 1000 subsamples in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$				$\hat{\boldsymbol{\pi}}^{(1)}$				$\hat{\boldsymbol{\pi}}^{(0)}$			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	2.3944	0.1588	0.1606	< 0.001	2.3931	0.1708	0.1691	< 0.001	2.3946	0.1978	0.1979	< 0.001
Summer	0.4398	0.0906	0.0891	< 0.001	0.4392	0.0956	0.0962	< 0.001	0.4404	0.1068	0.1068	< 0.001
Fall	0.4368	0.1124	0.1112	0.0001	0.4358	0.1158	0.1141	0.0002	0.4370	0.1242	0.1239	0.0004
Winter	0.6263	0.0823	0.0818	< 0.001	0.6263	0.0858	0.0839	< 0.001	0.6280	0.0965	0.0948	< 0.001
Workingdays	0.2656	0.0557	0.0571	< 0.001	0.2654	0.0510	0.0526	< 0.001	0.2657	0.0526	0.0542	< 0.001
Daytime	1.7347	0.0982	0.0988	< 0.001	1.7342	0.1066	0.1074	< 0.001	1.7344	0.1250	0.1231	< 0.001
W2_cloudy	-0.0581	0.0642	0.0635	0.3655	-0.0577	0.0598	0.0584	0.3346	-0.0572	0.0602	0.0609	0.3420
W3_rain	-0.4508	0.1040	0.1046	< 0.001	-0.4495	0.1078	0.1093	< 0.001	-0.4496	0.1219	0.1213	0.0002
Temp	0.1830	0.0385	0.0404	< 0.001	0.1827	0.0378	0.0392	< 0.001	0.1839	0.0401	0.0405	< 0.001
Hum	-0.0374	0.0300	0.0280	0.2114	-0.0384	0.0288	0.0293	0.1815	-0.0378	0.0301	0.0293	0.2089
Windspeed	-0.0030	0.0265	0.0263	0.9111	-0.0045	0.0240	0.0242	0.8525	-0.0045	0.0247	0.0243	0.8546

	$\bar{\boldsymbol{\pi}}^{(2)}$				$\bar{\boldsymbol{\pi}}^{(1)}$				$\bar{\boldsymbol{\pi}}^{(0)}$				unif			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	2.3940	0.1969	0.1981	< 0.001	2.3934	0.2038	0.2022	< 0.001	2.3931	0.2201	0.2215	< 0.001	2.3935	0.2079	0.2093	< 0.001
Summer	0.4396	0.1140	0.1158	0.0001	0.4386	0.1165	0.1183	0.0002	0.4404	0.1241	0.1240	0.0004	0.4397	0.1235	0.1228	0.0004
Fall	0.4371	0.1467	0.1461	0.0029	0.4371	0.1465	0.1481	0.0028	0.4358	0.1496	0.1482	0.0036	0.4365	0.1576	0.1589	0.0056
Winter	0.6278	0.1032	0.1012	< 0.001	0.6265	0.1043	0.1053	< 0.001	0.6271	0.1121	0.1139	< 0.001	0.6281	0.1087	0.1095	< 0.001
Workingdays	0.2656	0.0669	0.0652	0.0001	0.2668	0.0600	0.0586	< 0.001	0.2665	0.0593	0.0607	< 0.001	0.2665	0.0626	0.0626	< 0.001
Daytime	1.7339	0.1169	0.1150	< 0.001	1.7346	0.1232	0.1229	< 0.001	1.7348	0.1383	0.1372	< 0.001	1.7330	0.1096	0.1106	< 0.001
W2_cloudy	-0.0574	0.0849	0.0859	0.4984	-0.0580	0.0774	0.0787	0.4535	-0.0568	0.0744	0.0725	0.4452	-0.0586	0.0829	0.0816	0.4796
W3_rain	-0.4490	0.1494	0.1489	0.0027	-0.4506	0.1529	0.1547	0.0032	-0.4505	0.1672	0.1659	0.0071	-0.4493	0.1826	0.1828	0.0139
Temp	0.1839	0.0501	0.0498	0.0002	0.1838	0.0480	0.0478	0.0001	0.1840	0.0486	0.0496	0.0002	0.1841	0.0533	0.0515	0.0006
Hum	-0.0377	0.0398	0.0405	0.3438	-0.0372	0.0372	0.0386	0.3177	-0.0373	0.0370	0.0382	0.3135	-0.0389	0.0409	0.0412	0.3427
Windspeed	-0.0047	0.0347	0.0358	0.8922	-0.0039	0.0306	0.0325	0.8976	-0.0035	0.0300	0.0308	0.9066	-0.0038	0.0340	0.0324	0.9105

Compared to full quasipossion regression model, all proposed method doesn't detect the effect of Hum, however, our proposed method have smaller P-value than uniform method.

Table 6.14.
The length ratios of the 95% confidence intervals of proposed subsampling to uniform subsampling in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$, subsample size $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.7640	0.8217	0.9515	0.9470	0.9802	1.0587
Summer	0.7337	0.7742	0.8644	0.9232	0.9428	1.0047
Fall	0.7133	0.7346	0.7879	0.9309	0.9297	0.9496
Winter	0.7567	0.7895	0.8870	0.9488	0.9596	1.0309
Workingdays	0.8901	0.8152	0.8404	1.0686	0.9590	0.9478
Daytime	0.8960	0.9726	1.1397	1.0657	1.1236	1.2609
W2_cloudy	0.7739	0.7215	0.7260	1.0237	0.9331	0.8972
W3_rain	0.5696	0.5906	0.6674	0.8183	0.8375	0.9158
Temp	0.7209	0.7091	0.7518	0.9395	0.9006	0.9109
Hum	0.7316	0.7024	0.7341	0.9721	0.9088	0.9034
Windspeed	0.7804	0.7058	0.7249	1.0192	0.8993	0.8828

Table 6.15.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	unif	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.9355	0.9485	0.9471	0.9321	0.9435	0.9556	0.9506
Summer	0.9411	0.9455	0.9335	0.9365	0.9433	0.9524	0.9360
Fall	0.9354	0.9375	0.9476	0.9383	0.9488	0.9439	0.9375
Winter	0.9566	0.9529	0.9569	0.9405	0.9500	0.9479	0.9376
Workingdays	0.9484	0.9545	0.9474	0.9413	0.9497	0.9460	0.9447
Daytime	0.9428	0.9504	0.9486	0.9358	0.9398	0.9299	0.9300
W2_cloudy	0.9477	0.9493	0.9422	0.9485	0.9491	0.9362	0.9413
W3_rain	0.9506	0.9489	0.9443	0.9461	0.9493	0.9398	0.9515
Temp	0.9539	0.9447	0.9498	0.9455	0.9535	0.9546	0.9447
Hum	0.9422	0.9552	0.9524	0.9362	0.9478	0.9402	0.9469
Windspeed	0.9529	0.9398	0.9540	0.9459	0.9500	0.9445	0.9435

Table 6.16.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$ in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
unif	0.9517	0.9411	0.9412	0.9393	0.9469	0.9435
$\hat{\boldsymbol{\pi}}^{(2)}$	0.9391	0.9455	0.9404	0.9545	0.9559	0.9337
$\hat{\boldsymbol{\pi}}^{(1)}$	0.9324	0.9335	0.9420	0.9525	0.9375	0.9477
$\hat{\boldsymbol{\pi}}^{(0)}$	0.9406	0.9365	0.9523	0.9445	0.9365	0.9561
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9375	0.9433	0.9342	0.9379	0.9487	0.9417
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9448	0.9524	0.9386	0.9457	0.9519	0.9332
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9370	0.9360	0.9474	0.9406	0.9451	0.9539

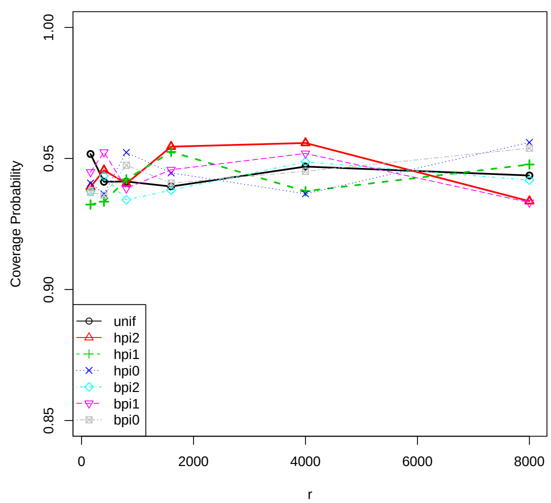


Figure 6.7. Simulated percentages of the 95% confidence intervals which caught the full sample MLE plot of $\hat{\beta}_2$ in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.17.
MSE ratios of the proposed subsampling to uniform subsampling in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
$\hat{\boldsymbol{\pi}}^{(2)}$	0.5347	0.5449	0.5434	0.5971	0.6652	0.7931
$\hat{\boldsymbol{\pi}}^{(1)}$	0.5875	0.5808	0.6066	0.6047	0.5741	0.8126
$\hat{\boldsymbol{\pi}}^{(0)}$	0.7374	0.7411	0.7565	0.7490	0.7524	0.9830
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8742	0.8873	0.9041	0.9411	0.9828	1.1221
$\bar{\boldsymbol{\pi}}^{(1)}$	0.8933	0.9053	0.8914	0.9397	0.8861	0.8972
$\bar{\boldsymbol{\pi}}^{(0)}$	1.0174	1.0206	1.0207	1.0220	1.0617	1.1821

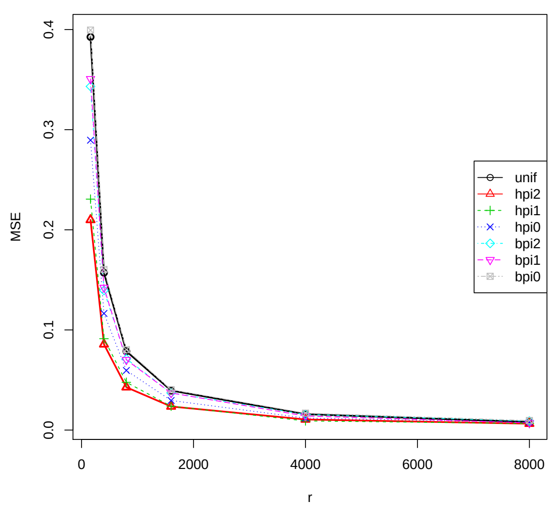
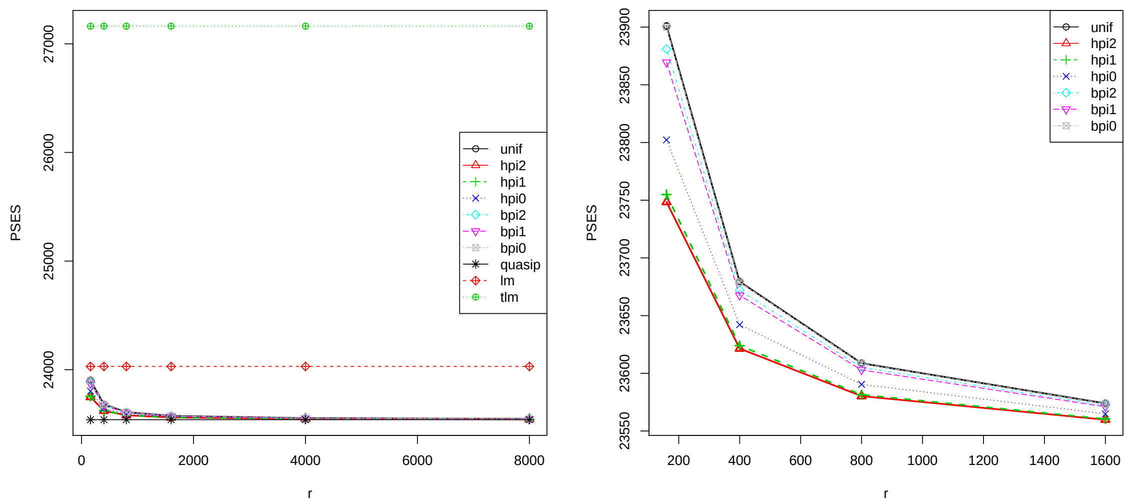


Figure 6.8. MSE plots in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.18.
Averages of the sum of squared predicted errors in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$. The sum of the squared prediction errors are 23,539.5308, 24,029.5526, and 27,162.6674 for the full sample Quasipoisson, linear regression and the log-transformed linear regression respectively.

r	160	400	800	1600	4000	8000
unif	23900.8137	23679.3619	23608.6705	23573.9071	23553.2349	23546.3751
$\hat{\boldsymbol{\pi}}^{(2)}$	23748.4020	23621.4508	23580.2203	23559.8080	23547.6255	23543.5754
$\hat{\boldsymbol{\pi}}^{(1)}$	23754.8661	23623.9449	23581.4521	23560.4201	23547.8694	23543.6972
$\hat{\boldsymbol{\pi}}^{(0)}$	23802.2565	23642.0976	23590.3953	23564.8585	23549.6368	23544.5796
$\bar{\boldsymbol{\pi}}^{(2)}$	23881.0383	23671.9234	23605.0292	23572.1059	23552.5191	23546.0180
$\bar{\boldsymbol{\pi}}^{(1)}$	23869.4407	23667.5691	23602.8991	23571.0525	23552.1005	23545.8092
$\bar{\boldsymbol{\pi}}^{(0)}$	23899.9932	23679.0730	23608.5324	23573.8397	23553.2083	23546.3619



(a) (b)

Figure 6.9. Averages of the sum of squared predicted errors and the first several averages of the sum of squared predicted errors plot in Quasipoisson regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

6.2.2 Negative Binomial Regression for Registered Data

Table 6.19.
Averaged estimates, theoretical standard errors(Tse), empirical standard errors(Ese), and P-values based on 1000 subsamples in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$				$\hat{\boldsymbol{\pi}}^{(1)}$				$\hat{\boldsymbol{\pi}}^{(0)}$			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	2.3900	0.1526	0.1518	< 0.001	2.3893	0.1637	0.1621	< 0.001	2.3894	0.1834	0.1821	< 0.001
Summer	0.3850	0.0962	0.0972	0.0001	0.3849	0.1018	0.1008	0.0002	0.3849	0.1089	0.1104	0.0004
Fall	0.3742	0.1270	0.1267	0.0032	0.3745	0.1320	0.1338	0.0045	0.3749	0.1367	0.1374	0.0061
Winter	0.6087	0.0904	0.0892	< 0.001	0.6074	0.0932	0.0915	< 0.001	0.6074	0.0991	0.0972	< 0.001
Workingdays	0.1944	0.0628	0.0611	0.0020	0.1945	0.0602	0.0588	0.0012	0.1938	0.0618	0.0627	0.0017
Daytime	1.7123	0.0773	0.0775	< 0.001	1.7133	0.0752	0.0754	< 0.001	1.7121	0.0774	0.0782	< 0.001
W2_cloudy	-0.0262	0.0732	0.0743	0.7202	-0.0270	0.0704	0.0695	0.7012	-0.0251	0.0715	0.0702	0.7255
W3_rain	-0.4510	0.1080	0.1067	< 0.001	-0.4502	0.1105	0.1120	< 0.001	-0.4512	0.1197	0.1209	0.0002
Temp	0.2500	0.0441	0.0450	< 0.001	0.2485	0.0450	0.0468	< 0.001	0.2485	0.0476	0.0465	< 0.001
Hum	-0.0565	0.0336	0.0330	0.0925	-0.0562	0.0338	0.0324	0.0960	-0.0565	0.0364	0.0375	0.1212
Windspeed	-0.0164	0.0299	0.0314	0.5826	-0.0152	0.0285	0.0282	0.5936	-0.0159	0.0298	0.0306	0.5946

	$\hat{\boldsymbol{\pi}}^{(2)}$				$\hat{\boldsymbol{\pi}}^{(1)}$				$\hat{\boldsymbol{\pi}}^{(0)}$				unif			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	2.3886	0.1915	0.1911	< 0.001	2.3893	0.2051	0.2037	< 0.001	2.3902	0.2285	0.2298	< 0.001	2.3900	0.2102	0.2106	< 0.001
Summer	0.3861	0.1227	0.1211	0.0016	0.3857	0.1293	0.1276	0.0029	0.3848	0.1374	0.1389	0.0051	0.3850	0.1330	0.1328	0.0038
Fall	0.3749	0.1663	0.1668	0.0241	0.3743	0.1724	0.1728	0.0300	0.3741	0.1778	0.1768	0.0353	0.3745	0.1767	0.1760	0.0341
Winter	0.6086	0.1165	0.1158	< 0.001	0.6077	0.1203	0.1220	< 0.001	0.6079	0.1277	0.1273	< 0.001	0.6082	0.1218	0.1212	< 0.001
Workingdays	0.1948	0.0802	0.0800	0.0152	0.1948	0.0771	0.0775	0.0115	0.1951	0.0790	0.0790	0.0135	0.1950	0.0775	0.0759	0.0119
Daytime	1.7120	0.1067	0.1082	< 0.001	1.7138	0.1036	0.1048	< 0.001	1.7133	0.1066	0.1049	< 0.001	1.7132	0.1048	0.1055	< 0.001
W2_cloudy	-0.0270	0.0946	0.0961	0.7752	-0.0260	0.0907	0.0902	0.7740	-0.0257	0.0918	0.0917	0.7795	-0.0252	0.0912	0.0895	0.7825
W3_rain	-0.4520	0.1526	0.1516	0.0031	-0.4502	0.1564	0.1547	0.0040	-0.4503	0.1688	0.1686	0.0076	-0.4513	0.1746	0.1764	0.0098
Temp	0.2496	0.0557	0.0559	< 0.001	0.2484	0.0569	0.0562	< 0.001	0.2491	0.0600	0.0585	< 0.001	0.2482	0.0589	0.0593	< 0.001
Hum	-0.0564	0.0425	0.0423	0.1843	-0.0566	0.0427	0.0410	0.1853	-0.0550	0.0458	0.0447	0.2295	-0.0554	0.0444	0.0441	0.2123
Windspeed	-0.0156	0.0372	0.0369	0.6757	-0.0166	0.0357	0.0357	0.6411	-0.0158	0.0373	0.0384	0.6712	-0.0151	0.0373	0.0369	0.6854

Our proposed $\hat{\boldsymbol{\pi}}^{(2)}$ and $\hat{\boldsymbol{\pi}}^{(1)}$ sampling can detect the hum effect at significance level 10% level, while the uniform method can not.

Table 6.20.
The length ratios of the 95% confidence intervals of proposed subsampling to uniform subsampling in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$, subsample size $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.7259	0.7786	0.8721	0.9109	0.9754	1.0867
Summer	0.7237	0.7657	0.8190	0.9228	0.9727	1.0331
Fall	0.7185	0.7468	0.7732	0.9407	0.9756	1.0057
Winter	0.7424	0.7652	0.8141	0.9563	0.9875	1.0488
Workingdays	0.8101	0.7771	0.7971	1.0352	0.9949	1.0190
Daytime	0.7378	0.7173	0.7384	1.0188	0.9890	1.0172
W2_cloudy	0.8021	0.7710	0.7837	1.0369	0.9935	1.0064
W3_rain	0.6185	0.6327	0.6855	0.8740	0.8956	0.9667
Temp	0.7492	0.7629	0.8078	0.9458	0.9657	1.0185
Hum	0.7562	0.7597	0.8202	0.9568	0.9620	1.0304
Windspeed	0.8024	0.7651	0.7993	0.9974	0.9579	0.9995

Table 6.21.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	unif	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.9533	0.9462	0.9451	0.9490	0.9480	0.9448	0.9484
Summer	0.9370	0.9464	0.9389	0.9463	0.9403	0.9487	0.9349
Fall	0.9395	0.9422	0.9300	0.9514	0.9354	0.9360	0.9378
Winter	0.9466	0.9458	0.9392	0.9407	0.9518	0.9484	0.9556
Workingdays	0.9347	0.9531	0.9487	0.9334	0.9424	0.9456	0.9341
Daytime	0.9464	0.9536	0.9439	0.9441	0.9491	0.9389	0.9479
W2_cloudy	0.9531	0.9422	0.9506	0.9463	0.9404	0.9477	0.9394
W3_rain	0.9385	0.9358	0.9430	0.9506	0.9385	0.9363	0.9529
Temp	0.9552	0.9450	0.9462	0.9531	0.9429	0.9519	0.9384
Hum	0.9456	0.9343	0.9526	0.9540	0.9442	0.9533	0.9387
Windspeed	0.9430	0.9376	0.9361	0.9493	0.9486	0.9498	0.9396

Table 6.22.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$ in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
unif	0.9493	0.9370	0.9444	0.9460	0.9523	0.9570
$\hat{\boldsymbol{\pi}}^{(2)}$	0.9410	0.9464	0.9470	0.9464	0.9484	0.9388
$\hat{\boldsymbol{\pi}}^{(1)}$	0.9516	0.9389	0.9329	0.9373	0.9520	0.9399
$\hat{\boldsymbol{\pi}}^{(0)}$	0.9515	0.9463	0.9378	0.9422	0.9431	0.9524
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9425	0.9403	0.9487	0.9406	0.9411	0.9481
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9358	0.9487	0.9483	0.9498	0.9551	0.9426
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9510	0.9349	0.9381	0.9431	0.9418	0.9524

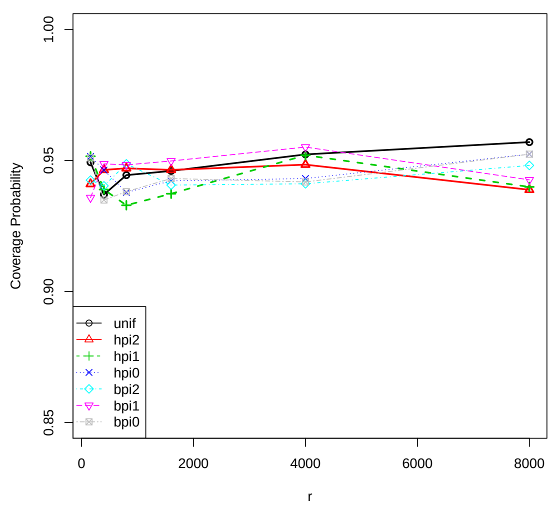


Figure 6.10. Simulated percentages of the 95% confidence intervals which caught the full sample MLE plot of $\hat{\beta}_2$ in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.23.
MSE ratios of the proposed subsampling to uniform subsampling in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
$\hat{\boldsymbol{\pi}}^{(2)}$	0.5193	0.5252	0.5227	0.5192	0.6590	0.7754
$\hat{\boldsymbol{\pi}}^{(1)}$	0.5498	0.5487	0.5683	0.6053	0.6262	0.7408
$\hat{\boldsymbol{\pi}}^{(0)}$	0.6324	0.6299	0.6318	0.6894	0.6642	0.7615
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8804	0.8806	0.9026	0.8948	0.9574	1.1030
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9313	0.9328	0.9365	0.9496	1.0160	0.8911
$\bar{\boldsymbol{\pi}}^{(0)}$	1.0597	1.0573	1.0608	1.0942	1.0647	1.1387

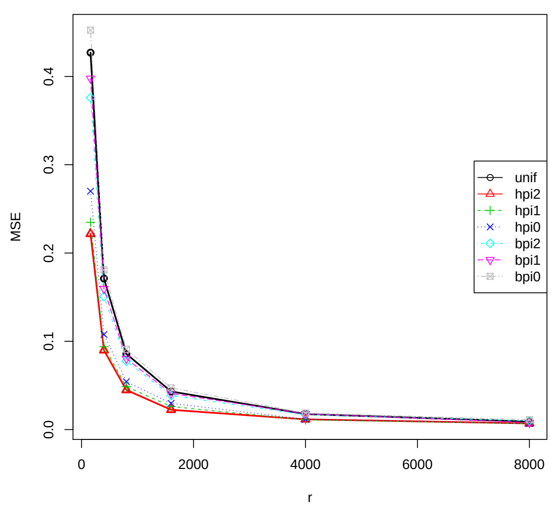


Figure 6.11. MSE plots in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.24.
Averages of the sum of squared predicted errors in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$. The sum of the squared prediction errors are 23,310.4025, 24,029.5526, and 27,162.6674 for the full sample Negative Binomial regression, linear regression and the log-transformed linear regression, respectively.

r	160	400	800	1600	4000	8000
unif	23802.6693	23500.7719	23404.5048	23357.1837	23329.0502	23319.7156
$\hat{\boldsymbol{\pi}}^{(2)}$	23578.6567	23415.6688	23362.6977	23336.4657	23320.8075	23315.6016
$\hat{\boldsymbol{\pi}}^{(1)}$	23588.4010	23419.4302	23364.5557	23337.3891	23321.1755	23315.7854
$\hat{\boldsymbol{\pi}}^{(0)}$	23628.0393	23434.6437	23372.0562	23341.1128	23322.6586	23316.5259
$\bar{\boldsymbol{\pi}}^{(2)}$	23759.2923	23484.4594	23396.5200	23353.2340	23327.4806	23318.9324
$\bar{\boldsymbol{\pi}}^{(1)}$	23775.7709	23490.6835	23399.5713	23354.7445	23328.0812	23319.2321
$\bar{\boldsymbol{\pi}}^{(0)}$	23838.2567	23514.0837	23411.0083	23360.3975	23330.3267	23320.3523

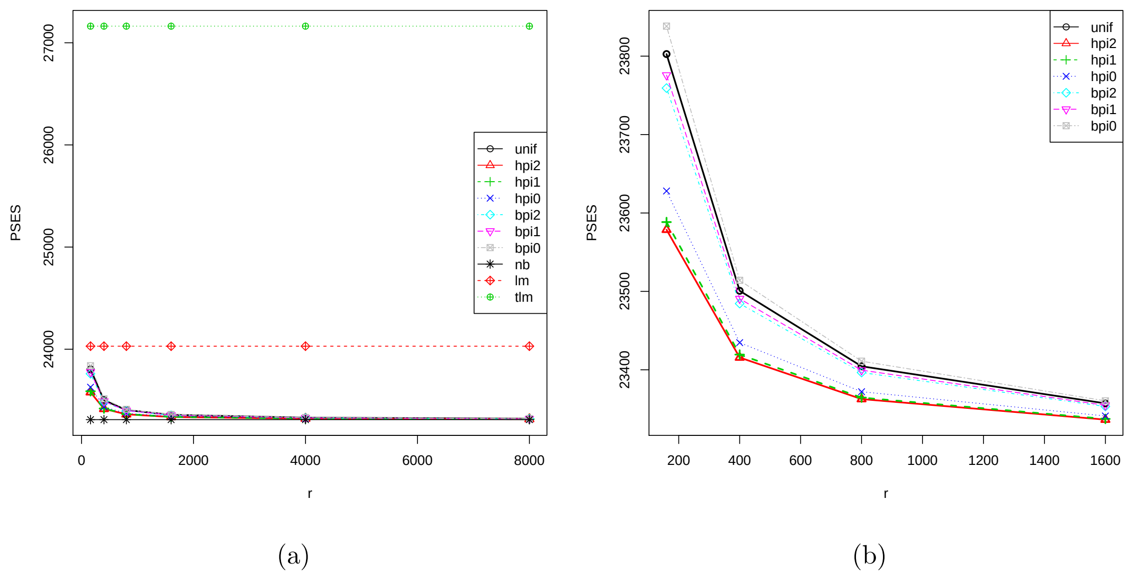


Figure 6.12. Averages of the sum of squared predicted errors and the first several averages of the sum of squared predicted errors plot in Negative Binomial regression model with the registered bike rental as the response variable, the pre-subsample size $r_0 = 200$.

6.3 Combined Bike Rentals

In the below tables and figures, we show the result of our proposed method for the combined bike rental as response.

6.3.1 Quasipoisson Regression Model for Combined Data

Table 6.25.
Averaged estimates, theoretical standard errors(Tse), empirical standard errors(Ese), and P-values based on 1000 subsamples in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$				$\hat{\boldsymbol{\pi}}^{(1)}$				$\hat{\boldsymbol{\pi}}^{(0)}$			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	2.7255	0.1513	0.1528	< 0.001	2.7246	0.1627	0.1621	< 0.001	2.7265	0.1890	0.1871	< 0.001
Summer	0.4551	0.0876	0.0891	< 0.001	0.4548	0.0932	0.0917	< 0.001	0.4536	0.1048	0.1030	< 0.001
Fall	0.4038	0.1076	0.1075	0.0002	0.4024	0.1113	0.1117	0.0003	0.4034	0.1204	0.1201	0.0008
Winter	0.6144	0.0799	0.0797	< 0.001	0.6144	0.0839	0.0841	< 0.001	0.6148	0.0951	0.0950	< 0.001
Workingdays	-0.0001	0.0535	0.0538	0.9991	0.0011	0.0483	0.0470	0.9821	0.0008	0.0487	0.0490	0.9877
Daytime	1.7144	0.0939	0.0947	< 0.001	1.7145	0.1021	0.1016	< 0.001	1.7139	0.1197	0.1200	< 0.001
W2_cloudy	-0.0399	0.0607	0.0588	0.5108	-0.0404	0.0564	0.0545	0.4735	-0.0399	0.0569	0.0585	0.4827
W3_rain	-0.4247	0.1005	0.1003	< 0.001	-0.4248	0.1048	0.1034	0.0001	-0.4245	0.1197	0.1203	0.0004
Temp	0.2534	0.0368	0.0383	< 0.001	0.2543	0.0362	0.0382	< 0.001	0.2527	0.0384	0.0378	< 0.001
Hum	-0.0717	0.0285	0.0271	0.0118	-0.0698	0.0272	0.0275	0.0104	-0.0703	0.0284	0.0269	0.0133
Windspeed	-0.0057	0.0252	0.0265	0.8216	-0.0049	0.0227	0.0228	0.8298	-0.0053	0.0233	0.0215	0.8206

	$\bar{\boldsymbol{\pi}}^{(2)}$				$\bar{\boldsymbol{\pi}}^{(1)}$				$\bar{\boldsymbol{\pi}}^{(0)}$				unif			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	2.7257	0.1848	0.1854	< 0.001	2.7249	0.1912	0.1919	< 0.001	2.7263	0.2088	0.2106	< 0.001	2.7253	0.1949	0.1949	< 0.001
Summer	0.4541	0.1094	0.1087	< 0.001	0.4552	0.1129	0.1147	0.0001	0.4550	0.1222	0.1235	0.0002	0.4538	0.1192	0.1198	0.0001
Fall	0.4032	0.1390	0.1370	0.0037	0.4032	0.1396	0.1414	0.0039	0.4035	0.1450	0.1462	0.0054	0.4039	0.1499	0.1512	0.0071
Winter	0.6136	0.0992	0.0988	< 0.001	0.6147	0.1012	0.1024	< 0.001	0.6142	0.1103	0.1097	< 0.001	0.6142	0.1045	0.1048	< 0.001
Workingdays	0.0003	0.0668	0.0669	0.9958	0.0004	0.0590	0.0596	0.9952	0.0002	0.0574	0.0574	0.9967	0.0007	0.0629	0.0636	0.9905
Daytime	1.7146	0.1109	0.1122	< 0.001	1.7133	0.1170	0.1157	< 0.001	1.7132	0.1317	0.1312	< 0.001	1.7136	0.1042	0.1025	< 0.001
W2_cloudy	-0.0389	0.0783	0.0798	0.6190	-0.0389	0.0710	0.0728	0.5845	-0.0405	0.0689	0.0673	0.5564	-0.0403	0.0760	0.0757	0.5958
W3_rain	-0.4256	0.1424	0.1420	0.0028	-0.4255	0.1463	0.1444	0.0036	-0.4245	0.1624	0.1634	0.0090	-0.4243	0.1754	0.1762	0.0156
Temp	0.2541	0.0475	0.0471	< 0.001	0.2527	0.0453	0.0441	< 0.001	0.2529	0.0462	0.0469	< 0.001	0.2546	0.0504	0.0490	< 0.001
Hum	-0.0710	0.0370	0.0383	0.0553	-0.0700	0.0345	0.0331	0.0425	-0.0699	0.0344	0.0362	0.0424	-0.0714	0.0379	0.0368	0.0597
Windspeed	-0.0055	0.0323	0.0331	0.8650	-0.0057	0.0284	0.0280	0.8411	-0.0054	0.0281	0.0300	0.8468	-0.0057	0.0318	0.0309	0.8582

Table 6.26.
The length ratios of the 95% confidence intervals of proposed subsampling to uniform subsampling in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$, subsample size $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.7763	0.8348	0.9699	0.9483	0.9812	1.0714
Summer	0.7353	0.7819	0.8795	0.9179	0.9475	1.0251
Fall	0.7177	0.7425	0.8029	0.9271	0.9314	0.9669
Winter	0.7648	0.8030	0.9097	0.9495	0.9678	1.0553
Workingdays	0.8513	0.7677	0.7741	1.0623	0.9374	0.9122
Daytime	0.9011	0.9799	1.1488	1.0651	1.1233	1.2641
W2_cloudy	0.7996	0.7425	0.7486	1.0305	0.9352	0.9063
W3_rain	0.5731	0.5975	0.6828	0.8121	0.8340	0.9261
Temp	0.7316	0.7178	0.7630	0.9426	0.9000	0.9175
Hum	0.7513	0.7186	0.7497	0.9772	0.9103	0.9092
Windspeed	0.7920	0.7139	0.7333	1.0159	0.8942	0.8848

Table 6.27.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	unif	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.9435	0.9364	0.9393	0.9458	0.9385	0.9326	0.9447
Summer	0.9487	0.9460	0.9545	0.9343	0.9451	0.9419	0.9495
Fall	0.9459	0.9357	0.9329	0.9385	0.9442	0.9495	0.9381
Winter	0.9429	0.9415	0.9390	0.9568	0.9421	0.9332	0.9443
Workingdays	0.9519	0.9316	0.9394	0.9384	0.9385	0.9445	0.9348
Daytime	0.9513	0.9451	0.9517	0.9426	0.9365	0.9503	0.9322
W2_cloudy	0.9454	0.9372	0.9426	0.9403	0.9362	0.9350	0.9411
W3_rain	0.9352	0.9402	0.9456	0.9422	0.9410	0.9377	0.9396
Temp	0.9469	0.9487	0.9515	0.9529	0.9389	0.9324	0.9385
Hum	0.9448	0.9336	0.9468	0.9425	0.9392	0.9440	0.9518
Windspeed	0.9552	0.9324	0.9435	0.9485	0.9365	0.9443	0.9349

Table 6.28.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$ in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
unif	0.9529	0.9487	0.9435	0.9453	0.9372	0.9500
$\hat{\boldsymbol{\pi}}^{(2)}$	0.9316	0.9460	0.9391	0.9530	0.9422	0.9362
$\hat{\boldsymbol{\pi}}^{(1)}$	0.9341	0.9545	0.9572	0.9356	0.9492	0.9508
$\hat{\boldsymbol{\pi}}^{(0)}$	0.9419	0.9343	0.9362	0.9358	0.9413	0.9381
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9337	0.9451	0.9391	0.9455	0.9476	0.9519
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9439	0.9419	0.9507	0.9396	0.9497	0.9453
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9391	0.9495	0.9461	0.9376	0.9333	0.9481

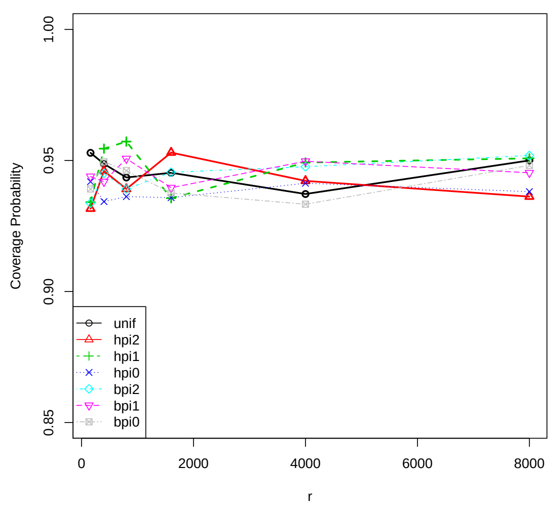


Figure 6.13. Simulated percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$ plot in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.29.
MSE ratios of the proposed subsampling to uniform subsampling in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
$\hat{\boldsymbol{\pi}}^{(2)}$	0.5454	0.5446	0.5578	0.6156	0.7129	0.7618
$\hat{\boldsymbol{\pi}}^{(1)}$	0.5941	0.5967	0.6197	0.6572	0.5966	0.9150
$\hat{\boldsymbol{\pi}}^{(0)}$	0.7577	0.7706	0.7582	0.8077	0.9320	0.8597
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8676	0.8731	0.8939	0.8732	0.9098	1.0556
$\bar{\boldsymbol{\pi}}^{(1)}$	0.8971	0.9113	0.9216	0.9662	0.9773	1.1354
$\bar{\boldsymbol{\pi}}^{(0)}$	1.0391	1.0563	1.0347	1.0967	1.0786	1.0132

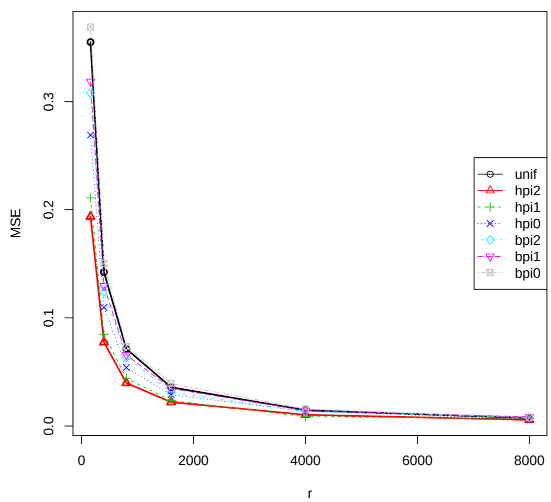
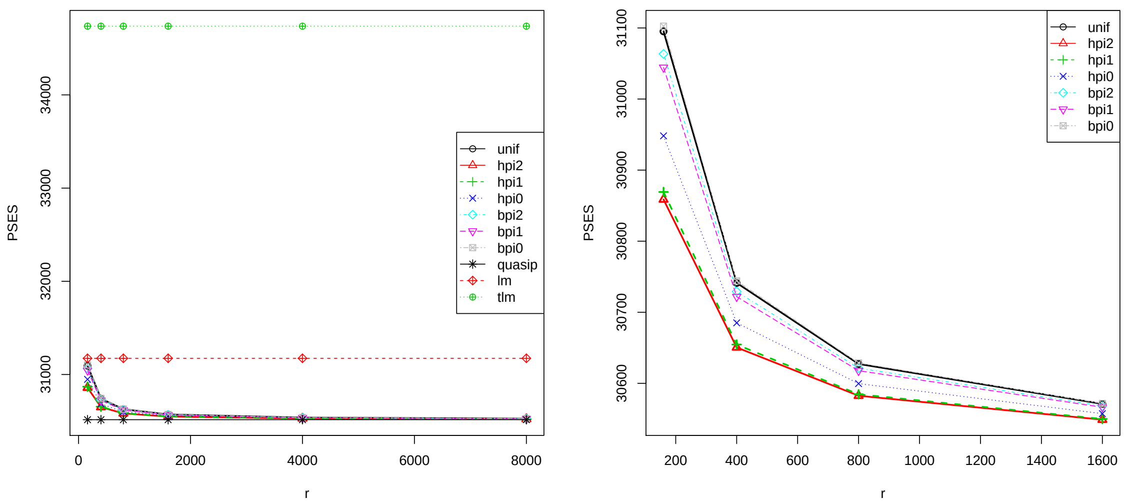


Figure 6.14. MSE plots in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.30.
Averages of the sum of squared predicted errors in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$. The sum of the squared prediction errors are 30,515.2950, 31,173.9273, and 34,737.2255 for the full sample Quasipoisson, linear regression and the log-transformed linear regression, respectively.

r	160	400	800	1600	4000	8000
unif	31095.2082	30741.2343	30627.2654	30571.0308	30537.5295	30526.4023
$\hat{\boldsymbol{\pi}}^{(2)}$	30858.6727	30650.4646	30582.5173	30548.8156	30528.6815	30521.9846
$\hat{\boldsymbol{\pi}}^{(1)}$	30869.1906	30654.5505	30584.5401	30549.8220	30529.0829	30522.1851
$\hat{\boldsymbol{\pi}}^{(0)}$	30948.1753	30685.0437	30599.6042	30557.3085	30532.0666	30523.6751
$\bar{\boldsymbol{\pi}}^{(2)}$	31063.4838	30729.1532	30621.3255	30568.0860	30536.3576	30525.8173
$\bar{\boldsymbol{\pi}}^{(1)}$	31044.4196	30721.9070	30617.7652	30566.3215	30535.6556	30525.4669
$\bar{\boldsymbol{\pi}}^{(0)}$	31102.5002	30744.0272	30628.6413	30571.7137	30537.8014	30526.5380



(a) (b)

Figure 6.15. Averages of the sum of squared predicted errors and the first several averages of the sum of squared predicted errors plot in Quasipoisson regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

6.3.2 Negative Binomial Regression for Combined Data

Table 6.31.
Averaged estimates, theoretical standard errors(Tse), empirical standard errors(Ese), and P-values based on 1000 subsamples in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$				$\hat{\boldsymbol{\pi}}^{(1)}$				$\hat{\boldsymbol{\pi}}^{(0)}$			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	2.6209	0.1470	0.1450	< 0.001	2.6219	0.1577	0.1567	< 0.001	2.6212	0.1767	0.1782	< 0.001
Summer	0.3981	0.0937	0.0941	< 0.001	0.3982	0.0992	0.1010	0.0001	0.3974	0.1060	0.1067	0.0002
Fall	0.3421	0.1230	0.1218	0.0054	0.3418	0.1280	0.1266	0.0076	0.3415	0.1326	0.1337	0.0100
Winter	0.5826	0.0879	0.0897	< 0.001	0.5838	0.0906	0.0895	< 0.001	0.5838	0.0963	0.0980	< 0.001
Workingdays	-0.0055	0.0617	0.0602	0.9294	-0.0044	0.0591	0.0575	0.9412	-0.0055	0.0605	0.0588	0.9270
Daytime	1.7289	0.0745	0.0760	< 0.001	1.7271	0.0724	0.0722	< 0.001	1.7284	0.0744	0.0744	< 0.001
W2_cloudy	-0.0262	0.0705	0.0717	0.7104	-0.0262	0.0677	0.0687	0.6986	-0.0253	0.0688	0.0705	0.7131
W3_rain	-0.4602	0.1047	0.1043	< 0.001	-0.4591	0.1071	0.1053	< 0.001	-0.4599	0.1161	0.1171	0.0001
Temp	0.3198	0.0429	0.0439	< 0.001	0.3198	0.0437	0.0425	< 0.001	0.3208	0.0463	0.0462	< 0.001
Hum	-0.0779	0.0324	0.0326	0.0163	-0.0782	0.0326	0.0326	0.0163	-0.0785	0.0352	0.0346	0.0256
Windspeed	-0.0207	0.0289	0.0278	0.4747	-0.0215	0.0276	0.0289	0.4352	-0.0212	0.0288	0.0274	0.4615

	$\bar{\boldsymbol{\pi}}^{(2)}$				$\bar{\boldsymbol{\pi}}^{(1)}$				$\bar{\boldsymbol{\pi}}^{(0)}$				unif			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	2.6216	0.1826	0.1840	< 0.001	2.6219	0.1955	0.1969	< 0.001	2.6210	0.2187	0.2184	< 0.001	2.6219	0.2008	0.2006	< 0.001
Summer	0.3980	0.1199	0.1213	0.0009	0.3976	0.1265	0.1266	0.0017	0.3964	0.1346	0.1366	0.0032	0.3977	0.1302	0.1307	0.0023
Fall	0.3416	0.1603	0.1619	0.0331	0.3402	0.1668	0.1657	0.0414	0.3418	0.1727	0.1714	0.0478	0.3404	0.1714	0.1702	0.0470
Winter	0.5837	0.1135	0.1141	< 0.001	0.5832	0.1171	0.1180	< 0.001	0.5829	0.1243	0.1241	< 0.001	0.5824	0.1186	0.1197	< 0.001
Workingdays	-0.0057	0.0783	0.0794	0.9423	-0.0042	0.0751	0.0744	0.9558	-0.0044	0.0769	0.0768	0.9549	-0.0050	0.0756	0.0749	0.9474
Daytime	1.7284	0.1029	0.1017	< 0.001	1.7272	0.0997	0.0997	< 0.001	1.7272	0.1025	0.1018	< 0.001	1.7278	0.1008	0.1004	< 0.001
W2_cloudy	-0.0256	0.0906	0.0892	0.7772	-0.0258	0.0868	0.0868	0.7664	-0.0256	0.0880	0.0867	0.7716	-0.0269	0.0873	0.0881	0.7577
W3_rain	-0.4594	0.1481	0.1472	0.0019	-0.4600	0.1517	0.1525	0.0024	-0.4598	0.1640	0.1659	0.0051	-0.4592	0.1695	0.1703	0.0067
Temp	0.3207	0.0539	0.0550	< 0.001	0.3201	0.0552	0.0562	< 0.001	0.3196	0.0584	0.0576	< 0.001	0.3211	0.0572	0.0570	< 0.001
Hum	-0.0770	0.0408	0.0401	0.0589	-0.0784	0.0410	0.0404	0.0556	-0.0781	0.0440	0.0446	0.0756	-0.0778	0.0426	0.0420	0.0683
Windspeed	-0.0200	0.0356	0.0361	0.5749	-0.0216	0.0343	0.0341	0.5294	-0.0208	0.0359	0.0344	0.5620	-0.0199	0.0359	0.0371	0.5790

Table 6.32.
The length ratios of the 95% confidence intervals of proposed subsampling to uniform subsampling in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$, subsample size $r = 400$.

	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.7321	0.7852	0.8801	0.9097	0.9740	1.0891
Summer	0.7195	0.7617	0.8144	0.9212	0.9717	1.0342
Fall	0.7178	0.7469	0.7738	0.9355	0.9735	1.0079
Winter	0.7417	0.7642	0.8118	0.9568	0.9872	1.0486
Workingdays	0.8161	0.7815	0.8003	1.0364	0.9937	1.0172
Daytime	0.7394	0.7184	0.7384	1.0207	0.9892	1.0170
W2_cloudy	0.8067	0.7754	0.7878	1.0370	0.9935	1.0081
W3_rain	0.6181	0.6322	0.6851	0.8741	0.8949	0.9680
Temp	0.7502	0.7640	0.8088	0.9423	0.9644	1.0205
Hum	0.7604	0.7640	0.8245	0.9562	0.9610	1.0310
Windspeed	0.8061	0.7689	0.8033	0.9922	0.9549	1.0004

Table 6.33.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$, $r = 400$.

	unif	$\hat{\boldsymbol{\pi}}^{(2)}$	$\hat{\boldsymbol{\pi}}^{(1)}$	$\hat{\boldsymbol{\pi}}^{(0)}$	$\bar{\boldsymbol{\pi}}^{(2)}$	$\bar{\boldsymbol{\pi}}^{(1)}$	$\bar{\boldsymbol{\pi}}^{(0)}$
Intercept	0.9537	0.9442	0.9481	0.9383	0.9502	0.9335	0.9313
Summer	0.9408	0.9448	0.9425	0.9534	0.9513	0.9512	0.9410
Fall	0.9543	0.9494	0.9481	0.9497	0.9394	0.9538	0.9496
Winter	0.9568	0.9555	0.9458	0.9486	0.9493	0.9366	0.9535
Workingdays	0.9453	0.9322	0.9473	0.9435	0.9392	0.9416	0.9327
Daytime	0.9345	0.9376	0.9443	0.9491	0.9474	0.9469	0.9501
W2_cloudy	0.9477	0.9575	0.9563	0.9392	0.9430	0.9400	0.9423
W3_rain	0.9374	0.9412	0.9485	0.9461	0.9471	0.9513	0.9554
Temp	0.9483	0.9361	0.9426	0.9391	0.9430	0.9400	0.9511
Hum	0.9524	0.9391	0.9445	0.9480	0.9400	0.9354	0.9503
Windspeed	0.9536	0.9336	0.9469	0.9467	0.9493	0.9394	0.9449

Table 6.34.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
unif	0.9341	0.9408	0.9529	0.9354	0.9489	0.9502
$\hat{\boldsymbol{\pi}}^{(2)}$	0.9571	0.9448	0.9499	0.9465	0.9401	0.9469
$\hat{\boldsymbol{\pi}}^{(1)}$	0.9521	0.9425	0.9515	0.9402	0.9405	0.9416
$\hat{\boldsymbol{\pi}}^{(0)}$	0.9421	0.9534	0.9535	0.9389	0.9563	0.9407
$\bar{\boldsymbol{\pi}}^{(2)}$	0.9407	0.9513	0.9462	0.9369	0.9484	0.9563
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9505	0.9512	0.9384	0.9371	0.9416	0.9500
$\bar{\boldsymbol{\pi}}^{(0)}$	0.9401	0.9410	0.9362	0.9359	0.9504	0.9479

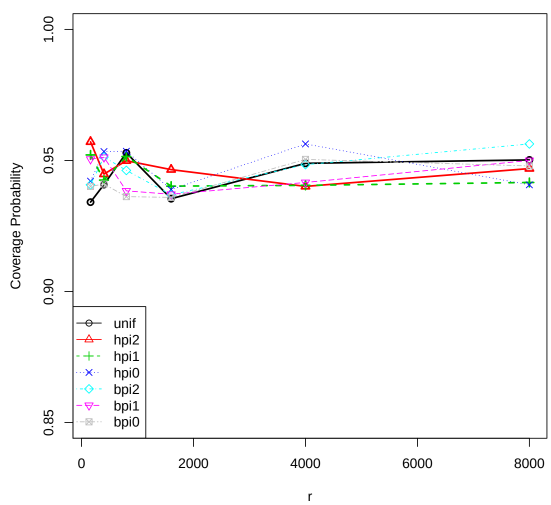


Figure 6.16. Simulated percentages of the 95% confidence intervals which caught the full sample MLE $\hat{\beta}_2$ plot in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

Table 6.35.
MSE ratios of the proposed subsampling to uniform subsampling in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

r	160	400	800	1600	4000	8000
$\hat{\boldsymbol{\pi}}^{(2)}$	0.5167	0.5174	0.5197	0.5243	0.6330	0.6594
$\hat{\boldsymbol{\pi}}^{(1)}$	0.5536	0.5620	0.5592	0.6185	0.6032	0.8555
$\hat{\boldsymbol{\pi}}^{(0)}$	0.6395	0.6382	0.6548	0.6920	0.7972	0.6552
$\bar{\boldsymbol{\pi}}^{(2)}$	0.8779	0.8878	0.8931	0.8908	0.8728	0.9790
$\bar{\boldsymbol{\pi}}^{(1)}$	0.9285	0.9364	0.9437	0.9790	0.9666	1.0724
$\bar{\boldsymbol{\pi}}^{(0)}$	1.0620	1.0655	1.0604	1.0995	1.1721	1.0456

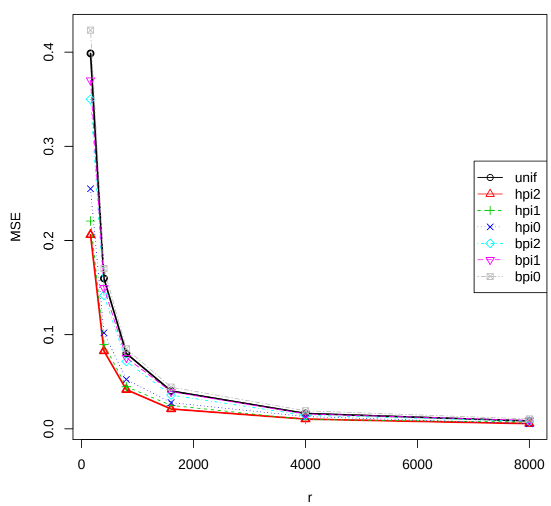
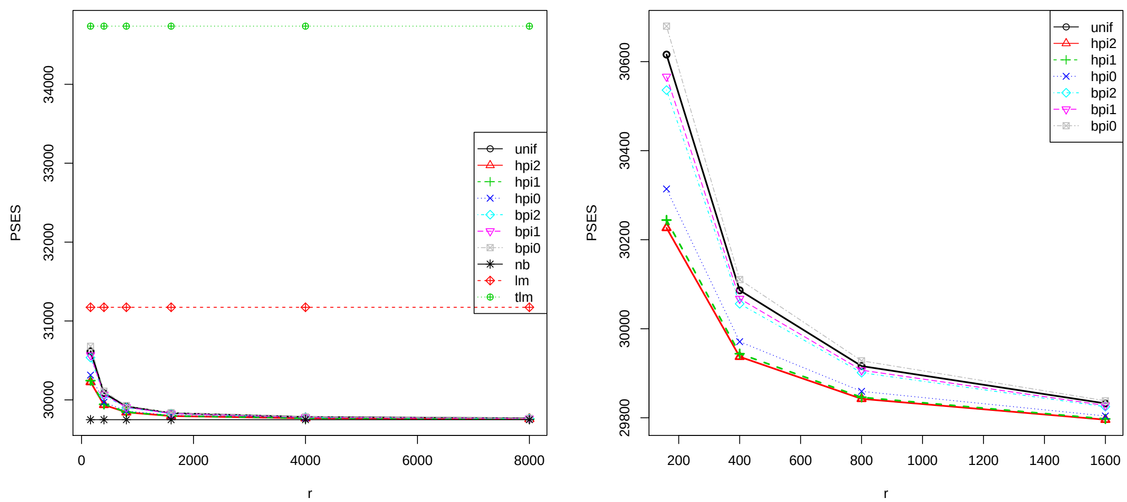


Figure 6.17. MSE plots in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 =$ 200.

Table 6.36.
Averages of the sum of squared predicted errors in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$. The sum of the squared prediction errors are 29,749.0368, 31,173.9273 and 34,737.2255 for the full sample Negative Binomial regression, linear regression, and the log-transformed linear regression, respectively.

r	160	400	800	1600	4000	8000
unif	30615.9223	30086.3590	29916.1353	29832.1963	29782.2072	29765.6065
$\hat{\boldsymbol{\pi}}^{(2)}$	30226.5013	29937.0480	29842.5484	29795.6692	29767.6602	29758.3436
$\hat{\boldsymbol{\pi}}^{(1)}$	30244.1503	29943.9024	29845.9415	29797.3573	29768.3334	29758.6799
$\hat{\boldsymbol{\pi}}^{(0)}$	30314.0683	29970.9335	29859.3019	29803.9988	29770.9807	29760.0020
$\bar{\boldsymbol{\pi}}^{(2)}$	30536.1013	30056.0057	29901.2189	29824.8030	29779.2654	29764.1381
$\bar{\boldsymbol{\pi}}^{(1)}$	30566.4065	30067.5692	29906.9084	29827.6247	29780.3886	29764.6988
$\bar{\boldsymbol{\pi}}^{(0)}$	30679.7756	30110.5270	29927.9925	29838.0684	29784.5425	29766.7718



(a) (b)

Figure 6.18. Averages of the sum of squared predicted errors and the first several averages of the sum of squared predicted errors plot in Negative Binomial regression model with the combined bike rental as the response variable, the pre-subsample size $r_0 = 200$.

6.4 Conclusions

First, by looking at the *workingday* variable, we infer that casual renters contain many tourists, and registered renters contain many people using bike rental as transportation tool, so the casual and registered bike renters are different. Second, the full sample prediction errors of combined data are larger than the prediction errors of casual data plus registered data, which indicates modeling the combined data leads to larger prediction errors than modeling the casual and registered data respectively. Casual and registered data should not be combined. Third, using prediction errors as criterion, we recommend Quasipoisson regression model for casual data and Negative Binomial regression model for registered data. Given weather and other covariates information, we can apply our models to the new data to make prediction of hourly bike rentals, this answers our research question. Fourth, our proposed subsampling methods give superior results to the uniform subsampling and $\hat{\boldsymbol{\pi}}^{(2)}$ method is the best.

7. A-OPTIMAL SUBSAMPLING FOR REAL DATA ANALYSIS: BLOG FEEDBACK DATA

In this chapter, we apply the proposed subsampling method to analyze the Blog Feedback data set, available from the UCI machine learning repository. This data was collected and processed from raw html of the blog posts. The goal is to predict the number of comments in the upcoming 24 hours relative to the base time. The base time was chosen from the past, and the blog posts selected were published within 72 hours before base time. The features were recorded at the base time based on the selected blog posts.

There are 52,397 observations in the training data set, and 7,624 observations in the test data set. We use the training data set to build the model, and the test data set to calculate the prediction errors. The 23 features are total number of comments before base time (Tc), number of comments in the 24 hours right before the base time (Cl24), number of comments in the time period between T1 and T2 (Ct1t2), where T1 denotes the date time 48 hours before base time, T2 denotes the date time 24 hours before base time, number of comments in 24 hours immediately after publication of the post but before base time (Cf24), total number of trackbacks before base time (Tt), number of trackbacks in the last 24 hours before the base time (Tl24), number of trackbacks between T1 and T2 (Tt1t2), where T1 is the time point 48 hours before basetime and T2 the time point 24 hours before basetime, number of trackbacks within 24 hours immediately after publication of the post but before basetime (Tf24), the length of time between the publication of the blog post and base time (Ltime), the length of the blog post (Lbp), indicators (0 or 1) for whether Monday to whether Saturday of the base time (Mbt, Tbt, Wbt, THbt, Fbt, Sbt), indicators (0 or 1) for whether Monday to whether Saturday of the blog publication date (Mpb, Tpb, Wpb, THpb, Fpb, Spb), number of parent pages(Ppage).

Poisson regression model is not appropriate for this data set because of observed overdispersion in data and inflated number of zeros. Quasipoisson regression model has the same parameter estimates as the Poisson regression model and does not accommodate zero-inflation, so it is not a good choice either. Zero-inflated Poisson regression model allows inflated zeros hence is an appropriate choice.

As 64.05% of the values in the response variable are 0, we shall consider fitting the zero-inflated Poisson regression model for this data set. The estimating equation of zero-inflated Poisson regression contains the parameter $0 \leq \rho \leq 1$, which accounts for the amount of positive structural zeros beyond the sampling zeros explained by the Poisson distribution f_{poi}. In the literature, ρ can be modeled as a function of the predictor variables, for example, via the logistic link. Here for simplifying the estimating process, we shall estimate ρ first. As Y follows the zero inflated model (2.2.2), we have

$$P(Y=0) = \rho + (1-\rho)\exp(-\mu).$$

On the other hand, $E(Y) = (1-\rho)\mu$. Thus $\mu = E(Y)/(1-\rho)$ and we get

$$P(Y=0) = \rho + (1-\rho)\exp(-E(Y)/(1-\rho)).$$

The moment equation of this is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[Y_i = 0] = \rho + (1-\rho)\exp(-\bar{Y}/(1-\rho)).$$

Since $\bar{y} = 6.765$, $\hat{p} = 0.6405$, we solve the equtions to get $\hat{\rho} \approx \hat{p} = 0.6405$.

To compare Poisson, Quasipoisson, with zero-inflated Poisson regression models, we report the full sample estimates, standard errors, P-values for these three models in Table (7.1). Many parameters in the Quasipoisson model are not significant, while these parameters in zero-inflated Poisson model are significant. Also, we perform proposed subsampling method in the zero-inflated Poisson regression model.

Table 7.1.
The estimates, standard errors, and P-values based on Poisson, Quasipoisson, and zero-inflated Poisson regression using the full sample, $n = 52,397$.

	Poisson	SE	P-value	Quasipoisson	SE	P-value	ZIPoisson	SE	P-value
Intercept	2.70536	0.01058	< 0.0001	2.70536	0.08167	< 0.0001	3.42978	0.01085	< 0.0001
Tc	0.00371	0.00004	< 0.0001	0.00371	0.00030	< 0.0001	0.00312	0.00004	< 0.0001
Cl24	0.00282	0.00004	< 0.0001	0.00282	0.00030	< 0.0001	0.00276	0.00004	< 0.0001
Ct1t2	0.00013	0.00005	0.00373	0.00013	0.00036	0.70717	0.00025	0.00005	< 0.0001
Cf24	-0.00236	0.00002	< 0.0001	-0.00236	0.00019	< 0.0001	-0.00254	0.00003	< 0.0001
Tt	0.18007	0.00482	< 0.0001	0.18007	0.03719	< 0.0001	0.15279	0.00471	< 0.0001
Tl24	-0.09276	0.00280	< 0.0001	-0.09276	0.02165	0.00002	-0.09377	0.00267	< 0.0001
Tt1t2	-0.03809	0.00313	< 0.0001	-0.03809	0.02412	0.11438	-0.04378	0.00298	< 0.0001
Tf24	-0.06000	0.00456	< 0.0001	-0.06000	0.03520	0.08830	-0.03660	0.00445	< 0.0001
Ltime	-0.06277	0.00014	< 0.0001	-0.06277	0.00107	< 0.0001	-0.05235	0.00015	< 0.0001
Lbp	0.00005	< 0.0001	< 0.0001	0.00005	0.00001	< 0.0001	0.00004	0.00001	< 0.0001
Mbt	0.19249	0.00912	< 0.0001	0.19249	0.07040	0.00626	0.09339	0.00933	< 0.0001
Tbt	0.07939	0.01072	< 0.0001	0.07939	0.08276	0.33744	-0.06151	0.01122	< 0.0001
Wbt	0.02238	0.01104	0.04267	0.02238	0.08523	0.79289	-0.13030	0.01155	< 0.0001
THbt	0.05547	0.01067	< 0.0001	0.05547	0.08238	0.50077	-0.09195	0.01108	< 0.0001
Fbt	-0.24868	0.00977	< 0.0001	-0.24868	0.07542	0.00098	-0.31279	0.01002	< 0.0001
Sbt	-0.23916	0.00794	< 0.0001	-0.23916	0.06128	0.00010	-0.22643	0.00805	< 0.0001
Mpb	0.18675	0.00992	< 0.0001	0.18675	0.07658	0.01474	0.15946	0.01051	< 0.0001
Tpb	0.23210	0.01107	< 0.0001	0.23210	0.08547	0.00662	0.22193	0.01169	< 0.0001
Wpb	0.05575	0.01158	< 0.0001	0.05575	0.08935	0.53271	0.08395	0.01204	< 0.0001
THpb	0.36164	0.01134	< 0.0001	0.36164	0.08755	0.00004	0.29686	0.01174	< 0.0001
Fpb	0.47488	0.01037	< 0.0001	0.47488	0.08004	< 0.0001	0.33577	0.01060	< 0.0001
Spb	0.19624	0.00984	< 0.0001	0.19624	0.07599	0.00982	0.09328	0.01011	< 0.0001
Ppage	-0.17265	0.00389	< 0.0001	-0.17265	0.03005	< 0.0001	-0.11498	0.00363	< 0.0001

Table 7.2.
Averaged estimates, theoretical standard errors(Tse), empirical standard errors(Ese), and P-values based on 1000 subsamples in the zero-inflated Poisson regression model, $r_0 = 2500$, $r = 5000$.

	unif				$\hat{\boldsymbol{\pi}}^{(2)}$			
	Estimate	Tse	Ese	P-value	Estimate	Tse	Ese	P-value
Intercept	3.31604	0.60943	0.56943	< 0.0001	3.39144	0.08391	0.07782	< 0.0001
Tc	0.00499	0.00463	0.00271	0.28070	0.00318	0.00036	0.00029	< 0.0001
Cl24	0.00298	0.00246	0.00172	0.22524	0.00270	0.00031	0.00029	< 0.0001
Ct1t2	-0.00006	0.00275	0.00212	0.98193	0.00024	0.00037	0.00034	0.51017
Cf24	-0.00407	0.00332	0.00266	0.22057	-0.00252	0.00026	0.00019	< 0.0001
Tt	0.13274	0.60876	0.32564	0.82739	0.15511	0.04384	0.03117	0.00040
Tl24	-0.08212	0.12111	0.12792	0.49776	-0.09500	0.01955	0.02092	< 0.0001
Tt1t2	-0.04429	0.13443	0.14660	0.74182	-0.04496	0.02124	0.02190	0.03431
Tf24	-0.02871	0.62134	0.32695	0.96314	-0.03759	0.04335	0.02892	0.38585
Ltime	-0.05948	0.00787	0.00744	< 0.0001	-0.05443	0.00168	0.00164	< 0.0001
Lbp	0.00003	0.00001	0.00001	0.02618	0.00004	0.00001	0.00001	< 0.0001
Mbt	0.15348	0.49478	0.47033	0.75641	0.13700	0.06758	0.06715	0.04264
Tbt	0.01812	1.01089	0.59248	0.98570	-0.07086	0.09689	0.10023	0.46461
Wbt	-0.15888	0.94287	0.61652	0.86618	-0.15383	0.10749	0.10257	0.15239
THbt	-0.11398	0.85801	0.59002	0.89431	-0.08562	0.10579	0.10892	0.41830
Fbt	-0.25691	0.72860	0.53337	0.72438	-0.25842	0.09592	0.11123	0.00706
Sbt	-0.25243	0.62211	0.43792	0.68491	-0.23805	0.08079	0.08066	0.00321
Mpb	0.18671	0.73179	0.49846	0.79861	0.22473	0.07984	0.10999	0.00488
Tpb	0.36845	0.76743	0.57054	0.63115	0.30397	0.10215	0.11611	0.00292
Wpb	0.23372	0.71303	0.59162	0.74307	0.12763	0.10816	0.11870	0.23797
THpb	0.33222	0.64228	0.61171	0.60499	0.29644	0.10121	0.12326	0.00340
Fpb	0.49398	0.60383	0.56036	0.41331	0.40595	0.08629	0.09056	< 0.0001
Spb	0.24244	0.57235	0.50639	0.67186	0.14735	0.07293	0.07362	0.04333
Ppage	-0.13385	0.10897	0.12224	0.21934	-0.11631	0.03527	0.03816	0.00097

In Table (7.2), we find that the standard errors of uniform method are bigger than those of $\hat{\boldsymbol{\pi}}^{(2)}$ method, the averaged parameter estimates of $\hat{\boldsymbol{\pi}}^{(2)}$ are closer to the full sample estimates than uniform method. For the uniform subsampling method, the comparisons between theoretical and empirical standard errors show large differences. This means the empirical performance of the uniform subsampling menthod does not reach the theoretical results in presence of inflated zeros, when $r = 5000$. Theoretical and the empirical standard errors of the $\hat{\boldsymbol{\pi}}^{(2)}$ method show that the empirical performance is consistent with theoretical result in presence of infalated zeros.

The comparison between uniform and $\hat{\boldsymbol{\pi}}^{(2)}$ subsampling methods suggests that for many variables the P-value of $\hat{\boldsymbol{\pi}}^{(2)}$ method is significant while that of uniform method are not. For example, effects of Tc, Cl24, Cf24, Tt, Tl24, Tt1t2, Mbt, Fbt, Sbt, Mpb, Tpb, THpb, Fpb, Spb, and Ppage are detected by $\hat{\boldsymbol{\pi}}^{(2)}$ method but are not detected by the uniform method. This means our proposed method reduces standard error hence increases power of test for regression coefficients.

Table 7.3.
The length ratios of the 95% confidence intervals of proposed $\hat{\boldsymbol{\pi}}^{(2)}$ subsampling methods to uniform subsampling method in zero-inflated Poisson regression model, pre-subsample size $r_0 = 2500$.

r	1000	2500	5000	10000	25000	50000
Intercept	0.1950	0.1368	0.1441	0.1499	0.1263	0.1315
Tc	0.0901	0.0976	0.1069	0.0953	0.0930	0.1012
Cl24	0.1452	0.1532	0.1792	0.1583	0.1571	0.1568
Ct1t2	0.1440	0.1481	0.1676	0.1448	0.1491	0.1469
Cf24	0.0599	0.0633	0.0720	0.0767	0.0736	0.0837
Tt	0.0848	0.0761	0.0915	0.0863	0.0792	0.0810
Tl24	0.1048	0.1152	0.1579	0.1522	0.1641	0.1751
Tt1t2	0.1055	0.1129	0.1522	0.1498	0.1555	0.1645
Tf24	0.0932	0.0776	0.0890	0.0797	0.0734	0.0786
Ltime	0.2348	0.2432	0.2191	0.2124	0.2308	0.2093
Lbp	0.3005	0.2605	0.2890	0.3476	0.2928	0.2687
Mbt	0.2015	0.1284	0.1478	0.1625	0.1280	0.1324
Tbt	0.1950	0.1338	0.1526	0.1611	0.1092	0.1218
Wbt	0.1829	0.1543	0.1669	0.1694	0.1312	0.1289
THbt	0.2104	0.1747	0.1678	0.1635	0.1480	0.1399
Fbt	0.2024	0.1512	0.1632	0.1571	0.1486	0.1556
Sbt	0.2066	0.1591	0.1737	0.1632	0.1438	0.1602
Mpb	0.1602	0.1746	0.1666	0.1742	0.1548	0.1285
Tpb	0.1756	0.1622	0.1721	0.1910	0.1795	0.1407
Wpb	0.1782	0.1923	0.1725	0.1979	0.1810	0.1415
THpb	0.1922	0.1623	0.1647	0.1770	0.1549	0.1568
Fpb	0.1597	0.1447	0.1662	0.1782	0.1426	0.1423
Spb	0.1664	0.1336	0.1333	0.1430	0.1242	0.1210
Ppage	0.2388	0.2534	0.3325	0.2951	0.2876	0.3433

Table 7.4.
Simulated percentages of the 95% confidence intervals which caught the full sample MLE in zero-inflated Poisson regression model, pre-subsample size $r_0 = 2500$.

r	1000	2500	5000	10000	25000	50000
Intercept	0.9989	0.9989	0.9919	0.9955	0.9924	0.9917
Tc	0.9905	0.9956	0.9902	0.9979	0.9915	0.9999
Cl24	0.9998	0.9916	0.9947	0.9941	0.9995	0.9938
Ct1t2	0.9965	0.9986	0.9989	0.9940	0.9923	0.9928
Cf24	0.9981	0.9973	0.9977	0.9977	0.9971	0.9958
Tt	0.9959	0.9942	0.9933	0.9991	0.9941	0.9922
Tl24	0.9901	0.9936	0.9951	0.9947	0.9998	0.9979
Tt1t2	0.9916	0.9998	0.9950	0.9916	0.9928	0.9961
Tf24	0.9994	0.9976	0.9949	0.9986	0.9920	0.9918
Ltime	0.9907	0.9944	0.9922	0.9944	0.9917	0.9984
Lbp	0.9903	0.9998	0.9997	0.9936	0.9934	0.9948
Mbt	0.9940	0.9903	0.9971	0.9932	0.9908	0.9948
Tbt	0.9992	0.9998	0.9972	0.9922	0.9989	0.9970
Wbt	0.9952	0.9916	0.9938	0.9927	0.9926	0.9979
THbt	0.9931	0.9918	0.9905	0.9914	0.9947	0.9930
Fbt	0.9983	0.9987	0.9949	0.9962	0.9934	0.9955
Sbt	0.9990	0.9978	0.9932	0.9949	0.9914	0.9995
Mpb	0.9978	0.9986	0.9936	1.0000	0.9999	0.9911
Tpb	0.9918	0.9961	0.9944	0.9987	0.9906	0.9990
Wpb	0.9950	0.9900	0.9919	0.9922	0.9974	0.9951
THpb	0.9917	0.9958	0.9945	0.9945	0.9963	0.9984
Fpb	0.9979	0.9986	0.9932	0.9966	0.9957	0.9998
Spb	0.9988	0.9929	0.9925	0.9994	0.9996	0.9905
Ppage	0.9953	0.9959	0.9917	0.9907	0.9980	0.9958

Table (7.3) and Table (7.4) are confidence interval length ratios and coverage probabilities. In table (7.3), all the values are smaller than 1, indicating that the lengths of 95% confidence intervals created by $\hat{\boldsymbol{\pi}}^{(2)}$ method are smaller than those of uniform method.

Table 7.5.
MSE ratios of the $\hat{\boldsymbol{\pi}}^{(2)}$ subsampling to the uniform subsampling method in zero-inflated Poisson regression model, pre-subsample size $r_0 = 2500$.

r	1000	2500	5000	10000	25000	50000
$\hat{\boldsymbol{\pi}}^{(2)}$	0.0287	0.0360	0.0373	0.0396	0.05796	0.0823

Table (7.5) shows the MSE ratios of $\hat{\boldsymbol{\pi}}^{(2)}$ method to uniform subsampling method. The values are smaller than 0.1, which means the MSE of our proposed method is less than 10% percent that of uniform subsampling.

Table 7.6.
Averages of the sum of squared predicted errors in zero-inflated Poisson regression model, pre-subsample size $r_0 = 2500$, the sum of the squared prediction error is 1,407.4712 for the full sample zero-inflated Poisson regression.

r	1000	2500	5000	10000	25000	50000
uniform	5215.3313	2876.1691	2653.2441	2323.7320	1811.6740	1598.1760
$\hat{\boldsymbol{\pi}}^{(2)}$	1599.7506	1525.9297	1524.9536	1509.2128	1500.5681	1428.4280

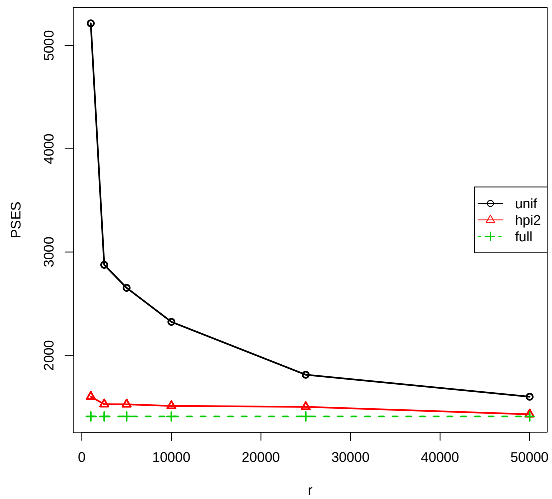


Figure 7.1. Averaged predicted sum of error squares plot in zero-inflated Poisson regression model, $r_0 = 2500$.

Table (7.6) reports averages of the sum of squared predicted errors, and Figure (7.1) is based on Table(7.6), we can find that when the subsample size r is small, the uniform method produces very large prediction error. The prediction errors of $\hat{\boldsymbol{\pi}}^{(2)}$ method are smaller than those of the uniform method.

REFERENCES

REFERENCES

[1] Avron, H., Maymounkov, P. and Toledo, S. (2010). Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, **32**: 1217–1236.

[2] Baxter, J., Jones, R., Lin, M. and Olsen, J. (2004). SLLN for Weighted Independent Identically Distributed Random Variables. *J. Theoret. Probab.*, **17**: 165–181. doi:10.1023/B:JOTP.0000020480.84425.8d.

[3] Barbe, P. and Bertail, P. (1995). *Weighted bootstrap*. Lecture Notes in Statist. Vol. 98, Springer, New York.

[4] Bhlmann, P., Kane M., Drineas P., Mark van der Laan (Eidotrs) (2016). *Handbook of Big Data*. Chapman and Hall/CRC.

[5] Chatterjee, S. and Bose, A. (2002). Dimension asymptotics for generalized bootstrap in linear regression. *Ann. Inst. Statist. Math.* **54** (2): 367–381.

[6] Cameron, C. and Trivedi, P. (1998). *Regression analysis of count data*. Cambridge University Press, Cambridge, United Kingdom.

[7] Chung, T., Peng, H. and Tan, F. (2018). A-optimal Subsampling For Big Data General Estimating Equations. *Manuscript*. Available at `https://www.math.iupui.edu/~hpeng/preprints_hp.html`.

[8] Chung, K.L. (2001). *A Course in Probability Theory*. Academic Press, San Diego, CA.

[9] Candés, E.J. and Tao, T. (2009). Exact Matrix Completion via Convex Optimization. *Found Comput Math* **9**: 717. doi:10.1007/s10208-009-9045-5.

[10] Dobson, A. and Barnett, A. (2002). *An Introduction to Generalized Linear Models*. CRC Press, Boca Raton, FL.

[11] Drineas P., Magdon-Ismail, M., Mahoney M.W. and Woodruff, D.P. (2012d). Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, **13**: 3475–3506.

[12] Drineas P., Kannan R. and Mahoney M.W. (2006a). Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, **36**: 132–157.

[13] Drineas, P., Mahoney,M.W., Muthukrishnan, S. and Sarlós, T. (2010). Faster least squares approximation. *Numerische Mathematik*, **117**(2): 219–249.

[14] Drineas P., Mahoney M.W. and Muthukrishnan S. (2006b). Sampling algorithms for ℓ_2 regression and applications. *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136.

[15] Fan, J., Han, F. and Liu, H. (2013). Challenges of big data analysis. *arXiv:1308.1479.*

[16] Freedman, D.A. (1981). Bootstrapping regression models. *Ann. Statist.* **9**(6): 1218–1228.

[17] Govindaraju, V., Raghavan, V.V. and Rao, C.R. (Editors) (2015). *Big Data Analytics: Handbook of statistics*. Volume 33. Publisher: Elsevier

[18] Lai, T. L. and C. Z. Wei (1982). A Law of the Iterated Logarithm for Double Arrays of Independent Random Variables with Applications to Regression and Time Series Models. *Ann. Probab.* **10**(2): 320–335.

[19] Ma, P. and Sun, X. (2014). Leveraging for big data regression. *Computational Statistics*. textbf7 (1): 70-76.

[20] Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *arXiv:1104.5557v3* [cs.DS]

[21] Mccullagh, P. and Nelder, J. (1984). *Generalized Linear Models.* Springer-Science+Business Media, New York, NY.

[22] Ma, P. , Mahoney, M.W, and Yu, B. (2015). A statistical perspective on algorithmic leveraging *Journal of Machine Learning Research.* **16** (April): 861–911.

[23] Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21** (4): 2053–2086.

[24] Peng, H. and Tan, F. (2018a). A Fast Algorithm For Computing The A-optimal Sampling Distributions In Big Data Linear Regression. *Preprint.* Available at `https://www.math.iupui.edu/~hpeng/preprints_hp.html`.

[25] Peng, H. and Tan, F. (2018b). Big Data Linear Regression Via A-optimal Subsampling. Submitted to *Ann. Statist.* Available at `https://www.math.iupui.edu/~hpeng/preprints_hp.html`.

[26] Sarlós, T. (2006). Improved approximation algorithms for large matrices via random projections. *In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152.

[27] Teicher, H.(1974). On the law of the iterated logarithm. *Ann. Probability* **2**: 714–728.

[28] Wang, C., Chen, M.-H., Schifano, E., Wu, J. and Yan, J. (2015). A Survey of Statistical Methods and Computing for Big Data. arXiv:1502.07989

[29] Wang, H., Zhu, R., and Ma, P. (2015). Optimal subsampling for large sample logistic regression. *JASA accepted.*

[30] Zhu, R., Ma, P., Mahoney, M.W. and Yu, B. (2015). Optimal subsampling Approaches for Large Sample Linear Regression. *arXiv:1509.0511.v1* [stat.ME].

VITA

VITA

My name is Xiaofeng Zhao. I received my Bachelor's degree from Donghua University in 2009. In 2012, I obtained my Master degree in Mathematics with concentration in Applied Statistics from the Department of Mathematical Sciences, IUPUI. Since then, I have continued my Ph.D. study in this Department.