Subsampling Method for Big Data in Poisson Regression

A N I Pradana and A Sofro*

Mathematics Department, Universitas Negeri Surabaya, Ketintang, Surabaya 60231, East Java, Indonesia

* Email: ayuninsofro@unesa.ac.id

Abstract. Big Data is very large, too fast and complex that cannot be managed with traditional tools; and must use new methods and tools to process it. Subsampling method is one method that can be used to solve big data problems. This paper focuses on estimation in Poisson Regression using Maximum Likelihood Estimation (MLE). In this study, it will be discussed about the implementation of the Subsampling method of The Demand for Medical Care data. To understand the performance of subsampling method, the paper provided a scenario with different sample sizes. The results of this study obtained parameter estimates and Poisson regression models on each subsample. Based on the Akaike Information Criterion, the best model is providing the smallest value. From the scenario, the sample size of the subsampling which is close to sampling tend to have smaller value of AIC.

1. Introduction

Big data is a collection of data that is too large so that there is a lot of information that is almost unlimited in number and requires large computational costs in analyzing it. The simplest way to reduce the computational cost of a procedure is to subsample data before doing anything else. With the development of science and technology, the statistical strategy that can solve the big data problem is using the subsampling method which provides a more effective. The approach is taking subsamples from the original data set and using the subsamples to estimate models, predictions and statistical inferences. There are several variants of the subsampling algorithm, for example to solve Ordinary Least Square (OLS) in linear regression for big data [1] and Leverage techniques are used for more informative samples from complete data subsets for linear regression [2].

The subsampling technique can be applied in most of the statistical analysis approach. If we want to predict or find out the relationship between variables, then regression analysis can be used which has two types of variables, namely the response ($Y$) variable and predictor variable ($X$). In this paper, we will focus on Poisson regression as one method to describe the relationship between several independent variables and the dependent variable in the form of discrete data [3].

One of the count data is about the demand for medical data which is concerned with the number of requests medical care for individuals aged 66 years. The total observation is 4406 covered by Medicare public insurance obtained from the United States National Medical Expenditure Survey (NMES) [4]. If all data is used to model, it is very inefficient and takes a long time. How to analyze the data efficiently is becoming increasingly important. Therefore, we need a more powerful approach, one of them is using the subsampling method which can make it easier to analyze. It also can overcome computational problems because of large data [5].
2. Method

2.1 Big Data
Big data is data that exceeds the capacity process of a conventional database system. Data is too large, moving too fast or not following the existing database architecture structure. Big data is voluminous and complex data sets; and therefore, it is difficult to be analyzed [6]. To get the value of the data, we must choose an alternative way to process it [7]. There are three main dimensions in big data, namely 3V: Volume showing the amount of data, Velocity shows the speed at which data is generated and analyzed, Variety shows on many forms of data taken, both structured and unstructured [8].

2.2 Subsampling Method
The first Subsampling method was introduced by Carlstein (1986) as a tool for estimating the parameters of statistical sampling distribution calculated from the sample [9]. The subsampling method is a method for dealing with big data by taking a subsample from the original data set concerning the probability distribution and using this sample as a substitute of the original data to estimate the model, predictions and statistical inference. In subsample retrieval for parameter estimation can use simple random sampling [5].

2.3 Poisson Regression
Poisson regression is one of the nonlinear regression methods. Poisson regression analysis is one of the statistical analytical tools. Poisson regression aims to describe the relationship between response / dependent variables in the form of discrete data and several predictor / independent variables [3]. In Poisson must meet the assumption of equidispersion, which means that the variance value of the response variable Y given by $\sum_i x_i = x$ must be equal to the mean value ie $\text{Var}(x) = E(x) = \mu$. Suppose there is a response variable that $y_i$ is the number of occurrences of an event and is a measure of observation, then the probability function $y_i$ is:

$$f(y_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, y_i = 1,2,3,...n,$$

(1)

or can be written $y_i \sim \text{Poi} (\mu_i), i = 1,2,...,n$.

where are the parameters $\mu_i > 0$. If $Y$ Poisson distribution has the same mean and variance, it will be shown as follows[10]:

$$E(y) = \mu \text{ and } \text{var}(y) = \mu.$$

Furthermore, the Poisson regression model can be written as

$$y_i = \exp \exp (x_i^T \beta).$$

(2)

2.4 Maximum Likelihood Estimation (MLE)
The Maximum Likelihood Estimation (MLE) method is one method for estimating parameters. This method was first introduced by RA Fisher in 1912. This estimation method can be applied to most problems and has a strong intuitive appeal, and often produces a good predictor of parameters $\beta$.

Let $y_1, y_2, ..., y_n$ be a continuous random variable of size n with a probability function $f(y_1, \beta)$ and $\beta$ is an unknown parameter.

In Poisson regression, the likelihood function is as follows:

$$L(y, \beta) = \prod_{i=1}^{n} f(y_i, \beta)$$

$$= \prod_{i=1}^{n} \left[ \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \right] = \frac{\prod_{i=1}^{n} \mu_i^{y_i} \exp \left(- \sum_{i=1}^{n} \mu_i \right)}{\prod_{i=1}^{n} y_i!},$$

(3)

And the log-likelihood function can be written as follows:
\[ l(\beta) = \log \log L(y; \beta) = \log \log \prod_{i=1}^{n} \left[ \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \right] \\
= \sum_{i=1}^{n} y_i (x_i^T \beta) - \sum_{i=1}^{n} \exp(x_i^T \beta) - \log \sum_{i=1}^{n} y_i! , \quad (4) \]

Parameter values \( \beta \) get by maximizing the opportunity function. This is done by looking for the first partial derivative of the likelihood function for each parameter and then zero. So, MLE \( \hat{\beta} \) is the solution to the following equation:

\[ \frac{\partial \log \log L(y; \beta)}{\partial \beta} = 0. \]

2.5 Subsampling Methods in Poisson Regression

In the first sampling approach, which is choosing a small subset of complete data called the subsampling step, then using the sample to estimate the model parameters.

General Subsampling algorithm:

- **Subsampling**
  Taking a random subsample with a size \( r > p \) where \( p \) is the number of predictor variables of all data using sampling probability distribution \( \{\pi_i\}_{i=1}^{n} \). Record the data selected \( \{x_i^*, y_i^*\}_{i=1}^{r} \) as predictor variables and response variables, as well as \( \pi_i^* \), to \( i = 1, 2, ..., r \).

- **Estimation**
  Maximize the log-likelihood function to get estimates \( \beta \) based on the subsample.

Based on the log-likelihood function of Poisson regression in equation (4), the log-likelihood function in the subsample can be written as follows:

\[ l(\beta) = \frac{1}{r} \sum_{i=1}^{r} \left[ y_i^* (x_i^*^T \beta) - \exp(x_i^*^T \beta) \right] . \quad (5) \]

Maximization can be applied with the Newton Rapshon method, with some iterations to converge.

In the subsampling method, the subsample number is getting closer to the sample, the smaller the MSE value means that the formed model is better if the subsample number is close to the number of samples [11].

2.6 The Demand for Medical Care

Demand for Medical Care is a health care request for the elderly population covered by Medicare, a public insurance program that offers substantial protection against health care costs. The request for treatment includes the number of visits to doctors / non-doctors in the office and the number of visits to doctors / non-doctors in outpatient hospitals. To model the demand for medical care, the number of doctor visits as a response variable was used and there were some factors including the number of hospitalizations at the hospital, self-perceived health status, number of patients in chronic conditions, gender, number of years of education, private insurance indicators.

Data were obtained from the United States National Medical Expenditure Survey (NMES) conducted in 1987 and 1988 to provide an overview of how Americans use and pay for health services. NMES is based on representation, national probability samples of civilians, non-institutionalized populations, and individuals who received long-term care facilities during 1987. Under the NMES survey, more than 38,000 people in 15,000 households throughout the United States were interviewed every three months about insurance coverage health, services used, and the costs and sources of payment.

2.7 Akaike’s Information Criterion (AIC)

Akaike’s Information Criterion (AIC) is the best selection criterion among some models. The best model is a model that has a minimum AIC value among all other models.

Akaike’s Information Criterion (AIC) is defined as follows [12]:

\[ AIC = -2 \ln \ln L(\beta) + 2 k . \quad (6) \]

where \( L(\beta) \) is the likelihood value and \( k \) is the number of parameters.
3. Results and Discussion

3.1 Subsampling Method for Poisson Regression

3.1.1 Taking Subsample

The first step is taking subsample at random or random sampling without returning, each member has the same opportunity to be chosen as a subsample member. The subsample is chosen randomly without regard to levels using R software. Sampling is used because the sample size is too large. Subsample taken based on the subsampling algorithm is as much as $r > p$ where $p$ is the number of predictor variables and $r$ is the number of subsamples taken. In sample data, $y$ as a response variable and $x$ as a predictor variable while in subsample data, the response variable is expressed as $y^*$ and the predictor variable is expressed as $x^*$. In this study used 4406 observations and 7 variables and selected subsample numbers $r = 400$ and $r > p = 6$.

3.1.2 Parameter Estimation of the Poisson Regression Model

In the next step after taking subsample, parameter estimation is done by using Maximum Likelihood Estimation (MLE) on $r$ subsample data as the original data by maximizing equation (7) as follows:

$$l(\beta) = \frac{1}{r} \sum_{i=1}^{r} \left[ y_i^* (x_i^*)^T \beta - \exp(x_i^T \beta) \right].$$  \hspace{1cm} (7)

The results of the parameter estimation can be seen in Table 1.

**Table 1. Estimation of parameters of the Poisson Regression Model on Subsample data with $r = 400$**

| Parameter | Independent Variables | Estimation |
|-----------|-----------------------|------------|
| 0         | (intercept)           | 1.22569    |
| 1         | Number of hospital stay ($x_1$) | 0.31707 |
| 2         | Health status ($x_2$)  | -0.42079   |
| 3         | Number of chronic diseases ($x_3$) | 0.16563 |
| 4         | Gender ($x_4$)        | -0.16948   |
| 5         | Number of years of education ($x_5$) | 0.02108 |
| 6         | Insurance indicator ($x_6$) | -0.05957 |

From the results of the parameter estimation above, the Poisson regression model formed is:

$$y_i = e^{(1.22569 + 0.31707 x_1 - 0.42079 x_2 + 0.16563 x_3 - 0.16948 x_4 + 0.02108 x_5 - 0.05957 x_6)}$$

Based on the Poisson regression model above, each increase in the number of hospitalizations increases the number of physician visits by $\exp(0.31707)$ if the other variables are not included in the model. Each increase in the number of chronic illnesses suffered by a person increases the number of physician visits by $\exp(0.16563)$ if other variables are not included in the model. Each increase in one year of education increases the number of doctor visits by as much as $\exp(0.02108)$ if the other variables are not included in the model.

3.2 Scenario

In this section, we will discuss some subsample scenarios to investigate approach performance. The scenario is taking several subsampling with different sample sizes, to determine differences in the estimated results of subsample data parameters. It is also to find the best model among the several subsamples. The amount of $r$ taken based on the algorithm is $r > p$. Table 2 is about the results of parameter estimation for several subsamples.
Table 2. Parameter estimation of Poisson Regression model on subsample data $r = 600, \ r = 800, r = 1000$

| Parameter   | Independent Variables | Estimation |
|-------------|-----------------------|------------|
|             | 600                   | 800        | 1000       |
| 0           | (intercept)           | 1.17868    | 1.15744    | 1.15574    |
| 1           | Number of hospital stay ($x_1$) | 0.21098 | 0.20994 | 0.18079 |
| 2           | Health status ($x_2$) | -0.32223 | -0.26647 | -0.46835 |
| 3           | Number of chronic diseases($x_3$) | 0.20214 | 0.23144 | 0.16359 |
| 4           | Gender ($x_4$)        | -0.16222 | -0.07386 | -0.13272 |
| 5           | Number of years of education($x_5$) | 0.00701 | 0.00980 | 0.02679 |
| 6           | Insurance indicator ($x_6$) | 0.24328 | 0.09469 | 0.07794 |

Based on the table, the Poisson regression model formed on each subsample is as follows:

For $r = 600$ : $y_i = e^{(1.17868 + 0.21098x_1 - 0.32223x_2 + 0.20214x_3 - 0.16222x_4 + 0.00701x_5 + 0.24328x_6)}$

For $r = 800$ : $y_i = e^{(1.15744 + 0.20994x_1 - 0.26647x_2 + 0.23144x_3 - 0.07386x_4 + 0.00980x_5 + 0.09469x_6)}$

For $r = 1000$ : $y_i = e^{(1.15574 + 0.18079x_1 - 0.46835x_2 + 0.16359x_3 - 0.13272x_4 + 0.02679x_5 + 0.07794x_6)}$

3.3 Selection of the Best Model

Based on the results of processing data from some subsamples used in this research, the AIC value will be obtained to find out the best model. The AIC value is obtained based on the log-likelihood value. The value of the log-likelihood for each subsample ($r$) are as follows, for $r = 400$ is $-4.730267$, $r = 600$ is $8.002661$, $r = 800$ is $5.192549$ and $r = 1000$ is $4.988598$. The AIC calculation results can be seen in Table 3 below:

Table 3. AIC results on each subsample model

| Subsample ($r$) | 24.67021 | 24.3851 | 23.9772 |

It can be seen from Table 3 that at $r$ (the number of subsamples) 1000 has the smallest AIC values. It means that at 1000 number of subsample, the model provides the best among all of the others.

4. Conclusion

It can be concluded that the subsampling method is one approach that can deal with big data by taking a subsample from completed data. Meanwhile, when the number of subsamples tends to the number of samples, thus the value of AIC is getting smaller.

References

[1] Drineas P, Mahoney M W, and Sarlós S M T 2011 Numer. Math. 117 219
[2] Ma P and Sun X 2014 Wire’s Comp. Stat. 7 70
[3] Cameron A C and Trivedi P K 1998 Regression Analysis of Cound Data (Cambridge: Cambridge University Press)
[4] Deb P and Trivedi P K 1998 J. Appl. Econom. 12 313
[5] Meng C, Wang Y, Zhang X , Mandal A , and Ma P 2017 Effective Statistical Methods for Big Data Analytics (USA: IGI Global)
[6] Shofiyah F and Sofro A 2018 J. Phy IOP Conf. Ser. 1108 012107
[7] Dumbill E 2012 Planning for Big Data (New York: O'Reilly)
[8] Phillips-Wren G and Hoskisson A 2014 Frontiers in Artificial Intelligence and Applications (IOS Press Ebooks)
[9] Fukuchi J I 1999 Biom. 86 591
[10] Myers R H, Montgomery D C, Vining G G, and Robinson T J 2010 Generalized Linear Models with Applications in Engineering and the Sciences (New Jersey: John Wiley)
[11] Wang H, Zhu R, and Ma P 2017 J. of the Am. Stat. Assoc. 113 829
[12] Akaike H 1974 IEEE 19 716