Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.6, No.12, 2016

The Comparison between the Simulation Variance for Censored and Uncensored Data for Maximum likelihood Normal Regression Model

Adel .A. Haddaw
Al-Mustansiriyah University –Iraq–Baghdad

ABSTRACT

A normal linear regression model is considered in which we have data for n+m individuals. For the first (n) individuals the values of the response variable, say $y_1, y_2, \ldots y_n$ represent uncensored observations while for the remaining (m) individuals, the values denote by y_{n+1}, $y_{n+2} \ldots y_{n+m}$ represent right-censored observations. Maximum likelihood estimation of the linear regression coefficients and residual variance for the normal case with censored and uncensored data is derived and assessed through simulation studies. The main findings result of the comparison between the simulation variance for censored and uncensored data is that for estimation of β_0 and β_1 for n=5, 10, the variance of the ml estimator had larger values than for estimation of β_0 and β_1 for n=5, 10 when there was censoring.

1- Introduction

In this section, it is important from literature review to present some books and papers related of this paper. A number of authors such as Wei, L.J. et al (1990) were considered linear regression analysis of censored survival data based on rank tests. (Haddaw, 1989) was derived and published a paper (Regression of Censored Data of Maximum Likelihood Regression Normal Case with Ungrouped Data). But it was only the derivation without any application by simulation or actual data. (Draper and Smith, 1981) ; (Ogah et al , 2011) were considered the least square estimator and its applications without censored data. Also a number of authors such as (Haddaw and Young, 1986), a regression model were considered in which the response variable has a type one extreme value distribution for smallest values and small sample moment properties of estimators of the regression coefficients and scale parameter, based on maximum likelihood estimation, ordinary least square and best linear unbiased estimation with censored and uncensored data ; (Kalbfleisch and Prentice, 2002) were considered the statistical analysis of failure time data ; (Jin et al, 2005) were considered rank regression analysis of multivariate failure time data based on marginal liner models.

There are two types of censoring, the first is that right-censored observations, while the second one is that left-censored observations. In this paper we were considered right-censored observations. There is important something to notice that censoring data is different from truncated data.

The purpose of this paper is to present maximum likelihood estimation of the regression coefficients and residual variance for the normal case with censored and uncensored data and its applications by simulation. The main purpose is to compare between the simulation variance for censored and uncensored data.

2- Theoretical Framework

Consider a regression model in which we have data for n+m individuals. For the first (n) individuals the values of the response variable, say $y_1, y_2, \ldots y_n$ represent uncensored

observations while for the remaining (m) individuals, the values denote by y_{n+1} , y_{n+2} , …, y_{n+m} represent right- censored observations. Thus if y_i is a random variable representing the response observation for the ith individuals, we have that

$$Y_i = y_i \quad , \qquad i= 1\dots n \quad .... \qquad (1)$$

$$Y_i > y_i \quad , \quad i = n+1\dots n+m \quad \dots \qquad (2)$$

We shall suppose that the ith individuals. So we have values x_{i1}, x_{i2} …, x_{ik} on k explanatory variables.

If we write

$$Y_i = \mu_i + \varepsilon_i \quad , i = 1\dots n + m \quad \dots \qquad (3)$$

Where Exp (ε_i) = 0, we shall assume that the usual multiple linear regression model with

$$\mu i = \sum_{j=0}^{K} \beta i \, xij \quad , i= 1 , \dots , n+m \quad \dots \qquad (4)$$

Where $x_{i0} = 1$ for i = 1… n+m . Then the usual assumptions that the true residuals have a constant variance and are uncorrelated, that is,

$$V(\varepsilon_i) = \sigma^2 \quad , Cov(\varepsilon_i, \varepsilon_i *) = 0 , i\neq i^*= 1 , \dots , n + m \dots. \qquad (5)$$

Assuming that the **(**ε_i) are IN (0, σ^2) random variables, the P.d. f of Yi is

$$f (y_i) = 1/ \sigma\sqrt{2\pi} \exp[-1/2(y_i - \mu i/ \sigma)^2] , \qquad - \infty < y < \infty \quad \dots. \qquad (6)$$

Since

$$P(Y_i > y_i) = 1/ \sigma\sqrt{2\pi} \int_{y_i}^{\infty} e^{-\frac{1}{2}(yi - \mu i / \sigma)2} \, dy = 1- \Phi(y_i - \mu i/ \sigma) \dots \qquad (7)$$

Where Φ(.) denote the c.d.f of the N(0, 1) distribution.

The likelihood function is

$$L = \{ \prod_{i = n+1}^{n} 1/\sigma\sqrt{2\pi} \exp[-1/2(y_i - \mu i/ \sigma)^2 \} \{ \prod_{i = 1}^{n+m} \{1-\Phi (y_i - \mu i/ \sigma) \quad \dots \qquad (8)$$

We have

$$\text{Log } L=- n/2 \log (2 \pi) – n \log \sigma – 1/2\sigma^2 \sum_{i=1}^{n} (y_i - \mu i)^2 + \sum_{i= n+1}^{n+m} \log(1- \Phi(y_i - \mu i/ \sigma) \qquad (9)$$

$$\partial \log L/ \partial \beta_j =1/ \sigma^2 \sum_{i=n+1}^{n} (y_i - \mu i) \partial \mu i/ \partial \beta_j + 1/ \sigma \sum_{i=1}^{n+m} \Phi(y_i - \mu i/ \sigma)/1- \Phi(y_i - \mu i/ \sigma) (\partial \mu i / \partial \beta_j)$$

$$= 1/\sigma^2 \{\sum_{i=1}^{n} (y_i - \mu i)\, x_{ij} + \sum_{i=n+1}^{n+m} \sigma\, \Phi(y_i - \mu i/\sigma)\, x_{ij} /1- \Phi(y_i - \mu i/\sigma) \}$$

$$= 1/\sigma^2 \{\sum_{i=1}^{n} (y_i - \mu i)\, x_{ij} + \sum_{i=n+1}^{n+m} \sigma\, x_{ij}\, h(y_i - \mu i/\sigma) \} \text{ , for } j = 0,1 , \ldots,k \quad \ldots (10)$$

Where

$$h(t) = \Phi(t)/\{ 1- \Phi(t)\} ,$$

is the hazard rate function for the N(0,1) .

Putting $z_i = (y_i - \mu i)/\sigma$, we may write formula (10) in the form

$$\partial \log L/ \partial \beta_j = 1/\sigma^2 \sum_{i=1}^{n+m} (y_i{}^* - \mu i)\, x_{ij} \text{ , } j = 0 , 1 , \ldots, k \quad \ldots \quad (11)$$

Where

$$y_i \quad , \quad i = 1 , 2 , \ldots, n$$

$$y_i{}^* = \begin{cases} \\ \mu i + \sigma\, h(z_i) \text{ , } i = n+1, \ldots n+m \end{cases} \quad \ldots. \quad (12)$$

We also have

$$\partial \log L/ \partial \sigma = -n/\sigma + \sum_{i=1}^{n} (y_i - \mu i)^2/\sigma^3 + 1/\sigma^2 \sum_{i=n+1}^{n+m} \Phi(y_i - \mu i/\sigma)/1- \Phi(y_i - \mu i/\sigma)$$

$$= 1/\sigma \{\sum_{i=1}^{n} z_i^{\,2} - n + \sum_{i=n+1}^{n+m} z_i\, h(z_i) \} \quad \ldots. \quad (13)$$

Equating $\partial \log L / \partial \beta_j$ and $\partial \log L/ \partial \sigma$ to zero, we see that the maximum likelihood estimates of the (β_j) and σ^2 satisfy the equations

$$\sum_{i=n+1}^{n+m} (y_i \wedge{}^* - \mu \wedge i)\, x_{ij} = 0 , \; j = 0 , 1 , \ldots, k \quad \ldots. \quad (14)$$

And

$$\sum_{i=1}^{n} \hat{z}_i^{\,2} + \sum_{i=n+1}^{n+m} \hat{z}_i\, h(\hat{z}_i) = n \quad \ldots \quad (15)$$

Where

$$\hat{\mu}_i = \sum_{j=0}^{n} \hat{\beta}_j x_{ij} \quad , i=1, \ldots, n+m \qquad \ldots \qquad (16)$$

$$\hat{z}_i = (y_i - \hat{\mu}_i) / \hat{\sigma} \quad , i=1, \ldots, n+m \quad \ldots. \qquad (17)$$

$$\hat{y}_i^* = \begin{cases} y_i & , \ i = 1, 2, \ldots, n \\ \hat{\mu}_i + \hat{\sigma}\, h(\hat{z}_i) & , \ i = n+1, \ldots, n+m \end{cases} \qquad \ldots \qquad (18)$$

In the case when there is no censoring(when m= 0) , we have $\hat{y}_i^* = y_i$, i= 0 ,1, .. k the set of equation(14) becomes

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i)\, x_{ij} = 0 \quad , j = 0, 1 \ldots k \qquad \ldots \qquad (19$$

Substituting

$\hat{\mu}_i = \sum_{j=0}^{k} \hat{\beta}_j x_{ij}$ and putting $\underline{\hat{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$, formula(19) in matrix form is

$$\underline{x}' \underline{x}\, \underline{\hat{\beta}} = \underline{x}' \underline{y} \qquad \ldots. \qquad (20)$$

Where

$\underline{x}$ is a matrix contains (n) rows and (n) columns of x's , and $x_{i0} = 1$ for i= 1… n. From formula (20) we have the well- known result:

$$\underline{\hat{\beta}} = (\underline{x}' \underline{x})^{-1} \underline{x}' \underline{y}$$

Also from formula (15), we have

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i / \hat{\sigma})^2 = n \qquad (21)$$

Leading to the estimator

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 / n$$

$$= \sum_{i=1}^{n} (y_i - \sum_{j=o}^{k} \hat{\beta}_j x_{ij})^2 / n \qquad (22)$$

The maximum likelihood (ML) estimator $\hat{\sigma}^2$ for the uncensored case is biased, an unbiased estimator being

$$\sigma^{\wedge 2} = \sum_{i=1}^{n} (y_i - \sum_{j=o}^{k} \beta_j \wedge x_{ij})^2 / n-k-1 \quad \ldots. \quad (23)$$

3- Applied Side Using Simulation

In this section we conducted simulation to assess the performance of maximum likelihood estimation of the regression coefficients and residual variance for the normal case.

Right censoring of the observations is common in many cases. Several forms of censoring are possible. Here we considered type 2 censoring. We suppose that the r smallest observations denote by $y_{(1)} < y_{(2)} < \ldots < y_{(r)}$ are observed, the remaining n- r observations being censored at the value $y_{(r)}$. The(r) is fixed integer satisfying $1 \leq r \leq n$. We let $R=\sum r$ denotes the total number of uncensored observations.

In order to examine the Ml estimators, a Monte Carlo simulation study was made for the case of a single explanatory variable, the model without censoring being

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , i = 1 , \ldots, n \quad \ldots \quad (24)$$

While with censoring being

$$Y_i = y_i \quad , \quad i = n+1 , \ldots , n+m$$

$E(\varepsilon_i) = 0$, $V \varepsilon_i) = \sigma^2$ and the Y_i are independently distributed with p.d.f for Y_i is given by

$$f(y_i) = 1/ \sigma\sqrt{2\pi} \exp[-1/2(y_i - \beta_0 + \beta_1 x_i / \sigma)^2] , \quad - \infty < y_i < \infty, \quad \ldots \quad (25)$$

Equally spaced values of(x) were used with $x_i = i - \frac{1}{2}(n+1)$, i=1, …n . Equal sample sizes n=5, 10 were used and equal censoring proportion p= 0.0, 0.25, 0.50 were applied. Without loss of generality, the y-observations were generated putting $\beta_0 = \beta_1 = 0$ in the regression model.

The ML estimates were obtained using a Minitab program. A run-size of 4000 was used in each case. Because of the big results (4000) values of the estimation for β_0 and β_1 so that Table 1 and table 2 does not contain the values of β_0 and β1 respectively.

Values of the variances of the ML estimators are shown in Table 1and Table 2 for β_0, β_1 respectively.

Table 1 Summary statistics for the simulation studies (n=5)

	P	Variance
β_0	0.00	0.203
	0.25	0.234
	0.50	0.245
β_1	0.00	0.205
	0.25	0.226
	0.50	0.249

Table 2 Summary statistics for the simulation studies (n=10)

	P	Variance
β_0	0.00	0.201
	0.25	0.224
	0.50	0.232
β_1	0.00	0.202
	0.25	0.224
	0.50	0.247

From Tables 1and 2, the main findings are as follows:

1- For estimation of β_0 for n=5, 10, the variance of the ML estimator had larger values than for estimation of β_0 for n=5, 10 when there was a heavy degree of censoring.

2- For estimation of β_1 for n=5, 10, the variance of the ML estimator had larger values than for estimation of β_1 for n=5, 10 when there was censoring.

3- For estimation of β_0 for n= 10, the variance of the ML estimator had smaller values than for estimation of β_0 for n=5 when there was no censoring and censoring.

4- For estimation of β_1 for n= 10, the variance of the ML estimator had smaller values than for estimation of β_1 for n=5 when there was no censoring and censoring.

4-Conclusion

In this paper, the ML estimator of the linear regression coefficients and residual variance for the normal case with censored and uncensored data was presented and simulated. For estimation of β_0 and β_1 for n=5, 10, the variance of the ML estimator had larger values than for estimation of β_0 and β_1 for n=5, 10 when there was censoring. For estimation of β_0 and β_1 for n= 10, the variance of the ML estimator had smaller values than for estimation of β_0 and β_1 for n=5 when there was no censoring and censoring.

ACKNOWLEDGMENT

The author thanks the reviewers for their helpful comments.

REFERENCES

1- Draper, N.R and Smith, H.(1982).Applied Regression Analysis 2nd , New York .John Wily and sons ,inc.

2- Haddaw, A. Ahmed.(1989). Regression of Censored Data of Maximum Likelihood Regression Normal Case with Ungrouped Data ..Journal of Management and Economics. Al-Mustansiriyah University.–Iraq–Baghdad. Faculty of Management and Economics, No.11

3- Haddow ,A .A and Young, D.H.(1986). Moment Properties of Estimators for A Type 1 Extreme -Value Regression Model. Communication.Statist.-Theory and Methods,15 (8).

4- Jin, Z.L and Ying, Z .(2005).Rank Regression analysis of Multivariate Failure Time Data Based On Marginal Linear Models. Scand. J. Statist.

5- Kalbfleisch, J.D. and Prentice, R.l.(2002). The Statistical Analysis of Failure Time Data, 2nd ed. Hoboken : Wiley.

6- Ogah ,D.M et al,.(2011).Relationship Between Body Measurements and Live Weight in Adult Muscovy Ducks Using Path Analysis. Trakia Journal of Sciences, Vol.9, No.1.

7- Wei, L.J. et al. (1990). Linear Regression analysis of Censored Survival Data Based on Rank Tests. Biometrika 77.