Journal of Natural Sciences Research
ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online)
Vol.4, No.3, 2014


# Maximum Likelihood Normal Regression with Censored and Uncensored Data

Adel .A. Haddaw
Al Isira University – Amman- Jordan

**ABSTRACT**

A normal regression situation is considered in which we have data for n+m individuals. The values denote by $y_{n+1}$ , $y_{n+2,\dots,}y_{n+m}$ represent right- censored observations. Maximum likelihood estimation of the regression coefficients and residual variance for the normal case with censored and uncensored data is derived and assessed through simulation studies.

## 1- Introduction

Consider a regression situation in which we have data for n+m individuals. For the first n individuals the values of the response variable, say $y_1, y_2, \dots y_n$ represent uncensored observations while for the remaining m individual, the values denote by $y_{n+1}, y_{n+2}, \dots, y_{n+m}$ represent right- censored observations. Thus if $y_i$ is a random variable representing the response observation for the ith individuals, we have that

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

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

We shall suppose that the ith individual . So we have values $x_{i1}, x_{i2} \dots, x_{ik}$ on k explanatory variables. If we write

$$Y_i = \mu i + \varepsilon i \quad , i = 1 , \dots, n + m \quad .....(3)$$

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

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

Where xio = 1 for i = 1 , …, n+m . Then the usual assumptions that the true residuals have constant variance and are uncorrelated will also be made, that is ,

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

A number of authors (Draper and Smith,(1981) ; Ogah et al,(2011) 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 was considered in which the response variable has a type one extreme value distribution for smallest values. 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) ; Wei et al, (1990) ; Jin et al, (2005) ; Jin et al,(2006) , were considered least- squares regression with censored data.

The purpose of this paper was to derive maximum likelihood estimation of the regression coefficients and residual variance for the normal case with censored and uncensored data and its applications.

## 2- Theoretical framework (Maximum Likelihood Estimation of the Regression Coefficints and Residual Variance for the Normal Case)

Assuming that the **(**Ei) are IN( 0 , $\sigma^2$) random variables , the P.d . f of $Y_i$ is

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

Since

$$P( Y_i > y_i ) = 1/ \sigma\sqrt{2\pi} \int^{\infty}_{y_i} e^{-1/2 (y - \mu i/ \sigma)2} dy = 1- \Phi(y_i - \mu i/ \sigma) \dots(7)$$

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

The likelihood function is

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

We have

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

Thus

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

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

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

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 (10) in the form

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

Where

$$y_i^* = \begin{cases} y_i, & i = 1, 2, \ldots, n \\ \mu i + \sigma\, h(z_i), & i = n+1, \ldots n+m \end{cases} \qquad \ldots.(12)$$

We also have

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

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

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

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

and

$$\sum_{i=1}^{n} z_i^{\wedge 2} + \sum_{i=n+1}^{n+m} z_i^{\wedge}\, h(z_i^{\wedge}) = n \qquad \ldots\ldots(15)$$

Where

$$\mu^{\wedge}i = \sum_{j=0}^{n} \beta_j^{\wedge}\, x_{ij}, \; i=1, \ldots, n+m \quad \ldots \quad (16)$$

$$z_i^{\wedge} = (y_i - \mu^{\wedge}i)/\sigma^{\wedge}, \; i=1, \ldots, n+m \quad \ldots. \qquad (17)$$

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

Journal of Natural Sciences Research
ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online)
Vol.4, No.3, 2014 

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

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

Substituting

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

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

Where

$\underline{x}$ is a matrix of x,s

and $x_{i0} = 1$ for i= 1, …, n . From (20) we have the well- known result

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

Also from (15) , we have

$$\sum_{i=0}^{n} (y_i - \hat{\mu}_i / \hat{\sigma})^2 = n \quad \ldots \qquad (21)$$

leading to the estimator

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

$$= \sum_{i=1}^{n} \Big(y_i - \sum_{j=0}^{k} \hat{\beta}_j\, x_{ij}\Big)^2 / n \qquad \ldots\ldots \qquad (22)$$

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

$$\hat{\sigma}_0^2 = \sum_{i=1}^{n} \Big(y_i - \sum_{j=0}^{k} \hat{\beta}_j\, x_{ij}\Big)^2 / n-k-1 \qquad \ldots\ldots(23)$$

**3- Applied Side (Results and Discussion)**

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

As we mentioned that in introduction, a common application for the normal regression model occurs in life-testing when the response variable represents the time to failure. Right censoring of the observations is common in such cases because of the need for early termination of the investigation. Several forms of censoring are possible. Here we shall consider 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 $R = \sum r$ denote 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

$$Y_i = \beta o + \beta 1 xi + \varepsilon i \quad , i = 1 , \ldots, n \qquad \ldots \qquad (24)$$

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

$$f(y_i) = 1/\sigma\sqrt{2\pi}\, \exp[\, -1/2(y_i - \beta o + \beta 1 xi / \sigma)^2\, ] , \quad -\infty < y_i < \infty, \qquad (25)$$

Equally spaced values of(x) were used with xi = i – ½(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 β0= β1=0 in the regression model.

The ML estimates were obtained using a Minitab program. A run-size of 4000 was used in each case.

Values of the biases, variances of the ML estimators are shown in tables 1,2 for β0, β1 respectively.

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

| | P | Bias | Variance |
|---|---|---|---|
| β0 | 0.00<br>0.25<br>0.50 | 0.002<br>-0.003<br>-0.004 | 0.203<br>0.234<br>0.245 |
| β1 | 0.00<br>0.25<br>0.50 | 0.003<br>0.004<br>0.006 | 0.204<br>0.226<br>0.249 |

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

| | P | Bias | Variance |
|---|---|---|---|
| β0 | 0.00<br>0.25<br>0.50 | 0.003<br>- 0.004<br>- 0.005 | 0.201<br>0.224<br>0.252 |
| β1 | 0.00<br>0.25<br>0.50 | 0.002<br>0.003<br>0.004 | 0.203<br>0.227<br>0.258 |

From tables 1, 2 the main findings are as follows.

1- For estimation of β0 for n=5, 10, the bias of the ML estimator was negligible and a positive when no censoring was present. But with censoring there was a negative bias which became more pronounced as the degree of censoring increased. The variance of the ML estimator had large values when there was a heavy degree of censoring.

2- For estimation of β1 for n=5, 10, the biases of the ML estimators were a positive bias and negligible in all cases. The variance of the ML estimator had small value when there was no censoring.

**4-Conclusion**

From literature review, there are a numbers of authors considered the least square estimator and its applications with uncensored and censored data. In the paper, the ML estimator of the regression coefficients and residual variance for the normal case with censored and uncensored data was derived. For estimation of β0 and β1 for n=5, 10, the biases of the ML estimator were negligible, a negative and a positive in all cases respectively. The variance of the ML estimator of β0 and β1 for n=5, 10, had large values when there was a heavy degree of censoring.

**ACKNOWLEDGEMENT**

The author thanks the reviewers foe their helpful comments.

**REFERENCE**

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

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

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

Jin, Z.L et al, (2006).On Least Squares Regression With Censored Data.Biomerika,93,1.

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

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.

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