Study on Response Loss under Linear Regression Model

Limin Song
School of Mathematics and Statistics, Shandong University of Technology, Zibo, Shandong, China

Abstract: This paper mainly studies the problem of missing response under the linear regression model, systematically summarizes the development situation, and introduces the research background and significance of missing response variables. The following focus: the first step is to build a missing model of linear regression response.In the second step, the two aspects of mean interpolation and observation data estimation are studied in detail respectively, and the specific mathematical formulas are given. In the third step, the estimate of regression coefficient, mean, distribution function, quantile under "complete sample" based on missing value and based on part number.

Keywords: *Linear Regression, Missing Response Variable, Mean Interpolation, Estimation Of Observation Data*

I. INTRODUCTION

In the process of daily research and learning, the absence of data is indispensable and inevitable. People usually think of using the traditional statistical analysis methods to compensate for the missing data, but due to the lack of data, these statistical methods cannot be completed smoothly. Now the question is transformed into how to deal with these missing data, so that we can make the statistical method continue, which has also become a hot topic, and has aroused the attention and research of many experts and scholars. Through this paper, we can accurately find the response missing model, and quickly solve the missing problem with mean interpolation and observation data estimation, and give the estimation of regression coefficient, mean, distribution function and quantile in three cases, so as to achieve the expected research purpose.

II. CONSTRUCTION OF THE RESPONSE MISSING MODEL

Consider the linear model as follows

$$Y = X\beta + \varepsilon \in R^{n\times 1}$$

among , $X=(X_1,X_2,...,X_p)\in R^{n\times p}$, $\beta=(\beta_1,\beta_2,...,\beta_p)^T\in R^{p\times 1}$ is an unknown parameter vector, Y is response variable, $\varepsilon=(\varepsilon_1,\varepsilon_2,...,\varepsilon_n)^T\in R^{n\times 1}$,error sequence $\{\varepsilon\}$ meet the independent and identical distribution,and $E(\varepsilon\,|\,X)=0, 0<\sigma^2=Var\varepsilon<\infty, Var(\varepsilon\,|\,X)=\delta^2$.

Incomplete samples with an independent and identical distribution are assumed $\{(X_i,Y_i,\delta_i),1\le i\le n\}$ the corresponding whole population, among $\{X_i,1\le i\le n\}$ can be all observed, $\{Y_i,1\le i\le n\}$ is missing, δ_i is indicating the Y_i missing variables,is a $1\times n$ vector, i. e

$$\delta_i=\begin{cases}0, if & Y_i \quad is \quad mis\sin g\\ 1, if & Y_i \quad is \quad not \quad mis\sin g\end{cases}.$$

Now use (X,Y,δ) instead $\{(X_i,Y_i,\delta_i),1\le i\le n\}$ the corresponding whole population. Therefore, it is assumed that $\{Y_i\}$ meet the MAR deletion mechanism, even $P(\delta=1\,|\,X,Y)=P(\delta=1\,|\,X)=P(X)$,that is, under the given X ,Y and δ independent from the conditions.

For convenience, we now introduce a notation:

$r=\sum_{i=1}^{n}\delta_i$,used to represent the number of cells without missing data;

$m=n-r$,used to represent the number of cells presence of missing data;

$s_r=\{i:\delta_i=1, i=1,\dots\dots,n\}$,used to represent the cell set without missing data;

$s_m=\{i:\delta_i=0, i=1,\dots\dots,n\}$,used to represent the set of the cells with missing data.

III. MEAN INTERPOLATION

Mean interpolation method is to replace the unobserved missing data with the mean of the observed data, which is more applicable when the observed data variables obey or approximate to the normal distribution. $Y=\{Y_{obs},Y_{mis}\}$, Y_{obs} is the observed data in the data matrix Y, Y_{mis} is the missing part in the Y, $n=n^{OB}+n^{NA}$,the single mean interpolation method is to average all the observed data and interpolate the missing data to obtain the unique mean as the interpolation value, namely $\overline{y_1}$,

$$\overline{y_1} = \frac{\sum_{i=1}^{n} \delta_i y_i}{n^{OB}}$$

The obtained population mean is estimated as:

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i y_i + (1-\delta_i)\bar{y}_1\right] = \frac{n^{OB}}{n}\bar{y}_1 + \frac{n-n^{OB}}{n}\bar{y}_1 = \bar{y}_1$$

It can be seen that the mean of the observed data is also an interpolated population mean estimate. s^2 is the variance of the overall data, s_1^2 is the variance of the observed data, $\bar{y}$ is the mean of the data, $\bar{y}_1$ is the mean value of the observed data, y_j is the value of the jth observation data, y_j' is the jth missing value, therefore

$$s^2 = \sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n-1} = \frac{\left[\sum_{i=1}^{n}\left(y_j-\bar{y}_1\right)^2 + \sum_{i=1}^{n}\left(y_j'-\bar{y}_1\right)^2\right]}{n-1}.$$

Using the mean value of the observed data with $\bar{y}_1$,instead of the unobserved y_j',therefore

$$s^2 = \sum_{i=1}^{n}\frac{(y_i-\bar{y}_1)^2}{n-1} = \frac{s_1^2\left(n^{OB}-1\right)}{n-1}.$$

IV. OBSERVE THE DATA FOR THE ESTIMATION (OE)

First, it can be obtained by the projection $\{Y_{obs}, Y_{mis}\}$,according to X_{obs}obtain $\hat{\beta}$,i.e. $\hat{\beta} = \left(X^T_{\text{obs}}X_{\text{obs}}\right)^{-1}X^T_{\text{obs}}Y_{\text{obs}}$.Then, based on complete observations $(X_i, Y_i)$for β weighted least squares (WLS) estimation, $i \in s_r$.

$$\hat{\beta}_r = \left(\sum_{i=1}^{n} \frac{\delta_i \tilde{X}_i' \tilde{X}_i}{v_0^2(X_i)}\right)^{-1} \sum_{i=1}^{n} \frac{\delta_i \tilde{X}_i Y_i}{v_0^2(X_i)},$$

Among $v_0(X)$ is a strictly positive, known function.

And then, by calculation $\hat{Y}_{mis}$,i.e. $\hat{Y}_{mis} = X_{mis}\hat{\beta}$.Using $\hat{Y}_{mis}$ to fill Y_{mis},then get $\{Y_{obs}, \hat{Y}_{mis}\}$. $\hat{Y}$ can be obtained after the expansion.

Making estimates under a "complete sample" with missing values, using the observed sample data for the pairs of {(X$_i$, Y$_i$), i ∈ s$_r$},given the estimates of the regression coefficients, mean, distribution function, quantile, i.e.,

$$\hat{\beta}_{n1} = \left(\sum_{i=1}^{n} \delta_i X_i' X_i\right)^{-1}\left(\sum_{i=1}^{n} \delta_i X_i' Y_i\right),$$

$$\bar{Y}_{n1} = \frac{1}{r}\sum_{i=1}^{n} \delta_i Y_i ,$$

$$\hat{F}_{n1}(Y) = \frac{1}{r}\sum_{i=1}^{n}\left\{\delta_i I(Y_i \le Y)\right\},$$

$$\hat{\theta}_{q1} = \inf\left\{\hat{F}_{n1}(u) \ge q\right\} = \hat{F}_{n1}^{-1}(q).$$

V. ESTIMATES BASED ON THE FIXED COMPLEMENT UNDER THE MISSING RESPONSE

Using fixed complement methods to fill in the missing response variables, whenY$_i$ is missing, it is complemented with its predicted value,that is $Y_{i1}^* = X_i\hat{\beta}_{n1}, i \in s_m$,is a $p \times n$ matrix ,as $Y_i, i \in s_m$ fill the value. After making up, $\tilde{Y}_{i1} = \delta_i Y_i + (1-\delta_i)Y_{i1}^*, i = 1,\ldots,n$, is a numerical value. Therefore, based on the estimation of the regression coefficient, mean, distribution function and quantile after fixed complement, that is,

$$\hat{\beta}_{n2} = \left(\sum_{i=1}^{n} X_i' X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i' \tilde{Y}_{i1}\right),$$

$$\bar{Y}_{n2} = \frac{1}{n}\sum_{i=1}^{n} \tilde{Y}_{i1} ,$$

$$\hat{F}_{n2}(Y) = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i I(Y_i \le Y) + (1-\delta_i) I\left(Y_{i1}^* \le Y\right)\right\},$$

$$\hat{\theta}_{q2} = \inf\left\{\hat{F}_{n2}(u) \ge q\right\} = \hat{F}_{n2}^{-1}(q).$$

VI. ESTIMATES FILLED BASED ON LINEAR REGRESSION FILLING WITH MISSING RESPONSE

Using the stochastic method, and learn from the part number linear regression filling method, that is, to use $Y_{i2}^{*} = X_i\hat{\beta}_{n1} + \varepsilon_i^{*}, i \in s_m$, is a $p \times n$ matrix, as $Y_i, i \in s_m$ fill the value, among $\varepsilon_i^{*} = J^{-1}\sum_{i=1}^{J}\varepsilon_{il}^{*}, J \geq 1, \{\varepsilon_{il}^{*}, l = 1,\ldots,J\}$ is from $\{Y_j - X_j^{'}\hat{\beta}_{n1}, j \in s_r\}$selected independently and repeatedly J sample. After making up, $\tilde{Y}_{i2} = \delta_i Y_i + (1-\delta_i)Y_{i2}^{*}, i = 1,\ldots,n$,is a numerical value.

Therefore, the regression coefficient, mean, distribution function and quantile based on random complement, that is,

$$\hat{\beta}_{n3} = \left(\sum_{i=1}^{n} X_i^{'} X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i^{'}\tilde{Y}_{i2}\right),$$

$$\bar{Y}_{n3} = \frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_{i2},$$

$$\hat{F}_{n3}(Y) = \frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{l}\delta_i I(Y_i \leq Y) + \\ (1-\delta_i)J^{-1}\sum_{l=1}^{J} I\left(X_i\hat{\beta}_{n1} + \varepsilon_{il}^{*} \leq Y\right)\end{array}\right\},$$

$$\hat{\theta}_{q3} = \inf\left\{\hat{F}_{n3}(u) \geq q\right\} = \hat{F}_{n3}^{-1}(q).$$

References

[1] Sanvesh Srivastava, Glen DePalma,Chuanhai Liu.An Asynchronous Distributed Expectation Maximization Algorithm For Massive Data:The DEM Algorithm [J].Journal of Computational & Graphical Statistics, 2018,1:34.

[2] Wang Q，Rao J N K. Empirical likelihood for linear regression models under imputation for missing responses[J].Canad J Statist，2001，29：597—608.

[3] Brick J M，Kalton G. Handling missing data in survey research[J]. Statist MethoDs Med Res，1996，5：215—238.