Influential Observation Detection in Bayesian Regression Using Conjugate Prior Distribution: An Application on Pavement Data

Semra Turkan[1], Gamze Ozel[2]

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

[1] Hacettepe University and Department of Statistics, Ankara, Turkey
sturkan@hacettepe.edu.tr

[2] Hacettepe University and Department of Statistics, Ankara, Turkey
gamzeozl@ hacettepe.edu.tr

Abstract

In this study, influential observation diagnostics based on case deletion approach with the ith case deleted are examined for bayesian regression. For this purpose, the formulas of Cooks distance, Welsch-Kuh distance and the Hadi measure which are major case deletion diagnostics in linear regression are derived for bayesian regression using conjugate prior distribution. To show the performance of proposed diagnostics on detection influential observations in bayesian regression using conjugate prior distribution, the pavement data is used.

Keywords: Bayesian Regression; Diagnostics; Influential Observations; Prior Information.

1 Introduction

Let $\mathbf{y}$ denote the vector containing data and let $\boldsymbol{\theta}$ denote the vector of unknown parameters of model. The distribution of y for a fixed value of $\boldsymbol{\theta}$ is called conditional distribution of y given $\boldsymbol{\theta}$. Before examining the data, what it is known about likely values for parameters is considered and this knowledge is translated into form of a probability distribution for $\boldsymbol{\theta}$. This is called the prior distribution of $\boldsymbol{\theta}$. Combining prior distribution of $\boldsymbol{\theta}$ and the conditional distribution of y given $\boldsymbol{\theta}$ to obtain the conditional distribution of θ given y is called posterior distribution of $\boldsymbol{\theta}$. Bayes' formula provides a means of combining the distributions of $\boldsymbol{\theta}$ and of y given $\boldsymbol{\theta}$ to obtain the distribution of $\boldsymbol{\theta}$ given y. It is expressed as

$$f(\boldsymbol{\theta}|y) = Cf(\boldsymbol{\theta})f(y|\boldsymbol{\theta}) \tag{1}$$

where C is a quantity that does not involve $\boldsymbol{\theta}$ [1, 2]. The Bayes estimates of regression coefficients are taken be the expectations of coefficients under posterior distribution. These turns out to be exactly the same as the least squares estimates.

The linear regression model is model for the conditional distribution of y given a vector of y given a vector of independent variables in x

$$y_i = x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad \text{or} \quad y \sim N(\boldsymbol{X}\beta, \ \sigma^2 I_n)$$

[1]. Bayes formula in regression usually is used to obtain the posterior distribution of all the parameters, β and σ^2. But in this study, we are interested in β not in σ^2. The Bayes formula with β in place of $\boldsymbol{\theta}$ and with all distributions conditional on σ^2 is expressed as

$$f(\boldsymbol{\beta}|y, \sigma^2) = Cf(\beta|\sigma^2)f(y|\beta, \sigma^2) \tag{2}$$

where C is a quantity that does not involve β. Since $f(\beta, \sigma^2)$ does not involve β, neither does $f(\boldsymbol{\beta}|\sigma^2)$, $f(\boldsymbol{\beta}|y, \sigma^2) = Cf(y|\beta, \sigma^2)$[2]. It is assuming that $f(y|\beta, \sigma^2)$ is the p.d.f. of a multivariate normal distribution with mean vector $X\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2 I$:

$$f(y|X, \beta, \sigma^2) = C \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right) \tag{3}$$

In least square analysis, the residual vector $y - X\hat{\beta}$ is perpendicular to all the columns of the regression matrix X, which implies that

$$(y - X\beta)'(y - X\beta) = (y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})$$

Therefore,

$$f(\beta|y, X, \sigma^2) = C \exp\left\{-\frac{1}{2\sigma^2}(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})\right\} \tag{4}$$

Let take the form of the prior distribution of β conditional on σ^2 to be multivariate normal, that is,

$$f(\beta|\sigma^2) = C \exp\left\{-\frac{1}{2\sigma^2}(\beta - b)'(V)^{-1}(\beta - b)\right\} \tag{5}$$

with some mean vector b and a variance-covariance matrix $\sigma^2 V$ proportional to σ^2. b and V are chosen to reflect prior knowledge. For this purpose, the information provided by previous study, gives some idea of what might be expected from the current study, could be used.

The Bayes estimate of parameter vector β is the expectation of β under the posterior distribution. This expectation is expressed as follow:

$$\begin{aligned} \hat{\beta}_{bayes} &= (V^{-1} + X'X)^{-1}V^{-1}b + (V^{-1} + X'X)^{-1}X'X\hat{\beta} \\ &= (V^{-1} + X'X)^{-1}V^{-1}b + (V^{-1} + X'X)^{-1}X'y \end{aligned} \tag{6}$$

[2].

2 Proposed influential diagnostics for the Bayesian Regression Analysis

In this section, we derived Cook's distance, Welsch-Kuh distance and Hadi measure for bayesian regression based on $\hat{\beta}_{bayes}$ in (6).

Cook's distance for Bayesian regression (D_i^{Bayes}): Cook's distance [3] measures distance between estimates of regression coefficients with ith observation $\hat{\beta}$ and without ith observation $\hat{\beta}_{-i}$.D_i^{Bayes} is proposed using case deletion approach and Sherman-Morrison –Woodbury Theorem such as

$$\begin{aligned} D_i^{Bayes} &= \frac{(\hat{\beta}_{bayes} - \hat{\beta}_{bayes,-i})'(X'X)(\hat{\beta}_{bayes} - \hat{\beta}_{bayes,-i})}{\hat{\sigma}^2 p} \\ &= \frac{e_{bayes,i}^2 \left(\sum_{j=1}^{n} h_{bayes,ij}^2\right)}{\hat{\sigma}^2 p (1 - h_{bayes,ii})^2} \end{aligned} \tag{7}$$

where $e_{bayes,i}^2 = x_i\hat{\beta}_{bayes} - y_i$ and $h_{bayes,ii} = X(V^{-1} + X'X)^{-1}X'$.

The Welsch-Kuh distance for Bayesian regression ($DFFITS_i^{bayes}$):The impact of ith observation on ith predicted value is measured by scaling change in prediction at x_i when ith observation is omitted. $DFFITS_i^{bayes}$ is proposed using difference between $\hat{\beta}_{bayes}$ and $\hat{\beta}_{bayes,-i}$ as

$$\begin{aligned} DFFITS_i^{bayes} &= \frac{|\hat{y}_{bayes,i} - \hat{y}_{bayes,i,-i}|}{Var(\hat{y}_{bayes,i})} = \frac{\left|x_i'(\hat{\beta}_{bayes} - \hat{\beta}_{bayes,-i})\right|}{Var(\hat{y}_{bayes,i})} \\ &= \frac{1}{Var(\hat{y}_{bayes,i})} \frac{h_{bayes,ii} e_{bayes,i}}{(1 - h_{bayes,ii})} \end{aligned} \tag{8}$$

Hadi Measure for Bayesian regression (H_i^{bayes}): Hadi measure is used to detect the overall potential influence. It can be modified for the Bayesian regression as

$$H_i^{bayes} = \frac{p}{1 - h_{bayes,ii}} \frac{d_{bayes,i}^2}{1 - d_{bayes,i}^2} + \frac{h_{bayes,ii}}{1 - h_{bayes,ii}}, \quad i = 1, 2, ..., n \tag{9}$$

where $d_{bayes,i}^2 = \frac{e_{bayes,i}^2}{\mathbf{e}_{\mathbf{bayes}}^T \mathbf{e}_{\mathbf{bayes}}}$ is square of ith normalized residual.

3 Application on Pavement Data

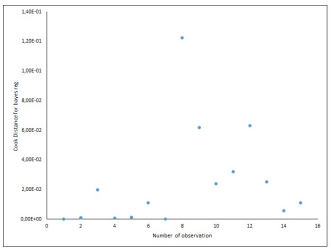
The pavement data is obtained from study that examines effect of several factors on rate at which a machine can rut asphalt pavement. The factors are considered as: viscosity of asphalt, transformed by logarithm function ($X1$), percentage of asphalt in surface ($X2$), percentage of asphalt in base ($X3$), percentage of fines in surface ($X4$), and percentage of voids in surface ($X5$). The response variable (Y) is logarithm of number of inches of change in rut depth per million wheel passes [2]. Based on previous information a mean vector and variance-covariance matrix for the prior distribution of β should be specified. For the mean vector b, the estimate of β obtained from the previous study could be chosen, which is

$$b = (-3.55, -0.44, 0.64, 0.13, 0.041, 0.14)$$

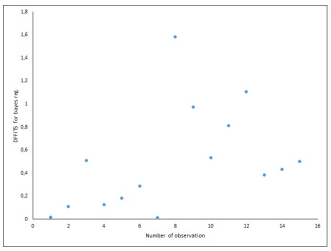
and for the variance-covariance matrix $\sigma^2 V$, the estimated variance-covariance matrix of the estimate of β from the previous study could be chosen, which is

$$V = \begin{bmatrix} 1690 & -38 & -174 & -110 & -10 & -47 \\ -38 & 4.07 & 3.46 & 3.23 & 0.47 & 0.23 \\ -174 & 3.46 & 19.81 & 8.88 & 1.15 & 5.31 \\ -110 & 3.23 & 8.88 & 12.47 & -0.17 & 1.56 \\ -10 & 0.47 & 1.15 & -0.17 & 0.6 & 0.15 \\ -47 & 0.23 & 5.31 & 1.56 & 0.15 & 2.69 \end{bmatrix}$$

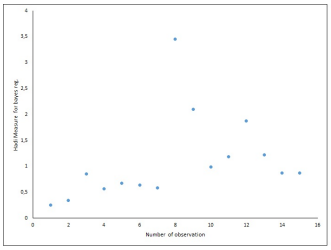
[2]. Using these prior information, values of D_i^{Bayes}, $DFFITS_i^{bayes}$, and H_i^{bayes} are obtained by preparing code in R program. The corresponding index plots of D_i^{Bayes}, $DFFITS_i^{bayes}$, and H_i^{bayes} that are shown in Figure 1, respectively.



(a) Index plot of D_i^{Bayes}



(b) Index plot of $DFFITS_i^{bayes}$



(c) Index plot of H_i^{bayes}

Figure 1: Index Plots of D_i^{Bayes}, $DFFITS_i^{bayes}$ and H_i^{bayes}

The index plots of D_i^{Bayes}, $DFFITS_i^{bayes}$ and H_i^{bayes} show that the observation 8 is found as influential observation.

References

[1] Rossi, P.E., Allenby, G.M., McCulloch, R. (2005). *Bayesian Statistics and Marketing*. 1st edition. John Wiley & Sons. New York.

[2] Birkes, D., Dodge, Y. (1993). *Alternative Methods of Regression*. 1st edition. John Wiley & Sons. New York.

[3] Cook, R.D. (1977). "Detection of influential observations in linear regression". *Technometrics*, **19**, 15-18.

[4] Hadi, A.S. (1992). "A New Measure of Overall Potential Influence in Linear Regression", *Computational Statistics and Data Analysis*, **14**, 1-27.

[5] Turkan, S., Cetin, M.C, Toktamis, O. (2012). "Outlier detection by regression diagnostic based on robust parameter estimates". Hacettepe Journal of Mathematics and Statistics, **41**(1), 147-155.