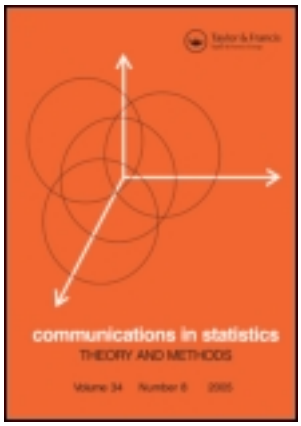

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Predictive Influence of Unavailable Values of Future Explanatory Variables in a Linear Model

S. K. Bhattacharjee [a], Ahmed Shamiri [a], Md. Sabiruzzaman [a] & S. Rao Jammalamadaka [b]
[a] Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur, Malaysia
[b] Department of Statistics and Applied Probability, University of California, Santa Barbara, California, USA

Available online: 10 Nov 2011

To cite this article: S. K. Bhattacharjee, Ahmed Shamiri, Md. Sabiruzzaman & S. Rao Jammalamadaka (2011): Predictive Influence of Unavailable Values of Future Explanatory Variables in a Linear Model, Communications in Statistics - Theory and Methods, 40:24, 4458-4466

To link to this article: http://dx.doi.org/10.1080/03610926.2010.513794

Communications in Statistics—Theory and Methods, 40: 4458–4466, 2011

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2010.513794

Predictive Influence of Unavailable Values of Future Explanatory Variables in a Linear Model

S. K. BHATTACHARJEE[1], AHMED SHAMIRI[1], MD. SABIRUZZAMAN[1], AND S. RAO JAMMALAMADAKA[2]

[1]Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur, Malaysia
[2]Department of Statistics and Applied Probability, University of California, Santa Barbara, California, USA

We consider an approach to prediction in linear model when values of the future explanatory variables are unavailable, we predict a future response y^f at a future sample point x^f when some components of x^f are unavailable. We consider both the cases where x^f are dependent and independent but normally distributed. A Taylor expansion is used to derive an approximation to the predictive density, and the influence of missing future explanatory variables (the loss or discrepancy) is assessed using the Kullback–Leibler measure of divergence. This discrepancy is compared in different scenarios including the situation where the missing variables are dropped entirely.

Keywords Discrepancy; Influence of variables; Kullback–Leibler divergence; Missing variable; Predictive density; Prior density; Taylor expansion.

Mathematics Subject Classification 62F15; 62J05; 62B10; 62M20.

1. Introduction

Predictive inference has been rightly called the central focus of statistical analyses and one may refer to Bjornstad (1990) for a review of various approaches. In recent years, much attention has been given to the influence of variables in both classical and Bayesian predictions. Bhattacharjee and Dunsmore (1991) considered the problem of the influence of variables in a logistic model in Bayesian predictive approach. In the logistic model, Zellner et al. (2004) compared the performance of stepwise selection procedure with a bagging method. The influence of variable selection in Bayesian diagnostic perspective in logistic model is considered by Weiss (1995). Predictive influence of variables in normal linear regression model has been studied by Bhattacharjee and Dunsmore (1995). Mollah and Bhattacharjee (2008)

Received June 8, 2010; Accepted July 27, 2010
Address correspondence to S. K. Bhattacharjee, Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur, Malaysia; E-mail: skbhattacharjee01@yahoo.com

considered the problem of the influence of variables in general linear regression model in Bayesian predictive approach in the presence of perfect multicollinearity.

Our aim here is to detect the influence of missing future explanatory variables in a normal linear model. We consider an approach to prediction analysis in general linear model when the values of some or all of the future explanatory variables are not available. We assume that in the observed data, the past records of all the explanatory variables are available. We want to predict a single future response y^f at a future sample point x^f when some or all components of x^f are unavailable. We assume that the future variables x^f are normally distributed but both the cases are considered where x^f's are dependent or independent. Improper prior densities are considered to derive the predictive density to assess the influence of the missing variables. Since the predictive density is not mathematically tractable for missing future explanatory variables, the Taylor expansion is used to derive the approximate predictive density. We then employ the Kullback–Leibler (K–L) directed measure of divergence (1951) to assess the influence of the missing future variables.

2. Bayes Predictive Density in Linear Model

Let us consider the general linear model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$, where ε's are the random errors normally distributed with mean zero and common variance σ^2. Our aim is to predict a future response y^f when some or all of the future explanatory variables x^f are unavailable. We denote the density of future explanatory variables by $f(x^f)$. Let us suppose that r future explanatory variables, denoted by $x^f_{(r)}$, are not available. For convenience and without loss of generality, we assume that the last r future variables are unavailable.

The density of an observed y is given by $p(y \mid x, \beta, \tau) = N(x\beta, \tau^{-1})$. The density of a future response y^f is $p(y^f \mid x^f, \beta, \tau) = N(x^f\beta, \tau^{-1})$. Then the predictive density of a future response y^f is given by $p(y^f \mid x^f, \Im) = \int p(y^f \mid x^f, \beta, \tau) p(\beta, \tau \mid \Im) d\beta d\tau$, where $p(\beta, \tau \mid \Im)$ is the posterior density of β and τ, and $\Im$ is the observed data.

We assume that the conditional density of $x^f_{(r)}$ given x^{f*} is independent of β and τ, i.e., $p(x^f_{(r)} \mid x^{f*}, \beta, \tau) = p(x^f_{(r)} \mid x^{f*})$, where x^{f*} denotes the future explanatory variables without variables $x^f_{(r)}$. First, we assume that x^f_i's are dependent and the distribution of x^f is k-dimensional multivariate normal, i.e., $f(x^f) = MN_k(\eta, \psi)$. The conditional density of $x^f_{(r)}$ given x^{f*} is given by

$$f(x^f_{(r)} \mid x^{f*}) = MN_r\left(\eta^*_{(r)}, \psi^*_{(r)}\right), \text{ where } \eta = \left(\eta^* \eta_{(r)}\right)', \; x^f = \left(x^{f*} x^f_{(r)}\right)', \; \psi = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix}$$

$$\eta^*_{(r)} = \eta_{(r)} + \psi_{21}\psi_{11}^{-1}\left(x^{f*} - \eta^*\right) \quad \text{and} \quad \psi^*_{(r)} = \psi_{22} - \psi_{21}\psi_{11}^{-1}\psi_{12}.$$

Then the density of future response when $x^f_{(r)}$ is missing is given by

$$p\left(y^f \mid x^{f*}, \beta, \tau\right) = \int p\left(y^f \mid x^f, \beta, \tau\right) f\left(x^f_{(r)} \mid x^{f*}\right) dx^f_{(r)}$$

$$= N\left(\sum_0^{k-r} x^f_i \beta_i + \sum_{k-r+1}^{k} \eta^*_i \beta_i, \; \sum_{k-r+1}^{k} \beta_i \beta_j \psi^*_{ij} + \tau^{-1}\right),$$

where η^*_i is the ith component of $\eta^*_{(r)}$ and ψ^*_{ij} is the (i, j)th component of $\psi^*_{(r)}$.

Using improper prior density for both β and τ, the approximate predictive density of y^f when $x^f_{(r)}$ is missing is given by

$$P_{(r)}\left(y^f \mid x^{f*}, \Im\right) \approx N\left(\sum_0^{k-r} x_i^f b_i + \sum_{k-r+1}^{k} \eta_i^* b_i,\ \sum_{k-r+1}^{k} b_i b_j \psi_{ij}^* + s^2\right) \times \left\{1 + 1/2 \sum_0^k Q_{ij}^*(\beta, \tau)\,\mathrm{cov}\left(\beta_i, \beta_j\right) + 1/2 Q_\tau^2(\beta, \tau)\,\mathrm{var}(\tau)\right\} \quad (1)$$

evaluated at b and s^2, where Q_{ij}^* is the multiplicative factor for the second-order approximation.

If x^f's are independent the corresponding approximate predictive density is

$$P_{(r)}\left(y^f \mid x^{f*}, \Im\right) \approx N\left(\sum_0^{k-r} x_i^f b_i + \sum_{k-r+1}^{k} \eta_i b_i,\ \sum_{k-r+1}^{k} b_i^2 \psi_i^2 + s^2\right) \times \left\{1 + 1/2 \sum_0^k Q_{ij}(\beta, \tau)\,\mathrm{cov}\left(\beta_i, \beta_j\right) + 1/2 Q_\tau^2(\beta, \tau)\,\mathrm{var}(\tau)\right\},$$

evaluated at b and s^2, where η_i and ψ_i^2 are mean and variance of the ith missing variable.

If no observation is missing then the corresponding predictive density based on all explanatory variables is given by

$$P\left(y^f \mid x^f, \Im\right) = St\left(n - p, x^f b, s^2\left(1 + x^f (X'X)^{-1} x^{f'}\right)\right) \quad (2)$$

Remark 2.1. Instead of considering normal linear regression model, the problem may be extended for more general regression model where the errors follow a spherically symmetric distribution. The predictive density is completely unaffected by departures of the normality assumption in the direction of the spherically symmetric family (Jammalamadaka et al., 1987). Therefore, the predictive density (2) based on all explanatory variables and no missing future variables will be unaltered if any spherically symmetric distribution is considered. The predictive density (1) for missing $x^f_{(r)}$ will be changed due to different forms of the distribution and approximation may be required to derive the predictive density.

3. Measure of Influence of the Missing Variables

To assess the influence of the missing future explanatory variables $x^f_{(r)}$, we employ the Kullback–Leibler (K–L) directed measure of divergence D_{KL} between the two predictive densities (1) and (2). The form of the (K–L) divergence used here is given by

$$D_{KL} = \int p\left(y^f \mid x^f, \Im\right) \log\left\{p\left(y^f \mid x^f, \Im\right)/p_{(r)}\left(y^f \mid x^{f*}, \Im\right)\right\} dy^f.$$

There are certainly many other measures of divergence between two distributions, e.g., a very general divergence measure is given by Csiszar (1963). He introduced

the following class of divergence measures, called "h-divergence", between two probability distributions $F_1(\cdot)$ and $F_2(\cdot)$:

$$I_h(F_1, F_2) = \int_R h\left(\frac{dF_1(x)}{dF_2(x)}\right) dF_2(x),$$

where $h : (0, \infty) \to \Re$ is a convex function with $h(1) = 0$. However, we use the special case of the K–L measure (corresponding to $h(x) = -\log(x)$) which is more practical and easy to calculate in our case. Also, we can use information measure $I = \int P(\cdot \mid \cdot) \log P(\cdot \mid \cdot) - \int P_{(r)}(\cdot \mid \cdot) \log P_{(r)}(\cdot \mid \cdot)$, ratio measure $R = D_{KL} \div |I|$, and predictive interval to assess the influence of the missing future explanatory variables, these measures give similar conclusion to K–L directed measure of divergence. Details of these measures may be found in Bhattacharjee (1987), the first author's unpublished Ph.D. thesis.

An explicit expression for D_{KL} with Student distribution is difficult to obtain, so we derive approximate D_{KL} by substituting approximate normal form of (2) as

$$N\left(x^f b, \frac{n-p}{n-p-2} s^2 \left(1 + x^f (X'X)^{-1} x^{f'}\right)\right). \tag{3}$$

Since it is difficult to derive the distributional form of D_{KL}, as in Bhattacharjee and Dunsmore (1995), any discrepancy due to missing variables $x^f_{(r)}$ is less than 1% of the largest discrepancy would be considered as negligible at 1% error. Where largest discrepancy occurs between the predictive density based on all variables and the predictive density based on no variable.

If x^f_i's are not independent then the approximate D_{KL}between the predictive densities (1) and (3) is given by

$$D_{KL} \approx \frac{1}{2}\frac{T^2}{\partial^2_{(r)}} + \frac{1}{2}\left[\frac{\partial^2}{\partial^2_{(r)}} - \log\left\{\frac{\partial^2}{\partial^2_{(r)}}\right\} - 1\right] - \frac{1}{2}\sum_{ij=0}^{k} E\left\{Q^*_{ij}(\beta, \tau)\operatorname{cov}(\beta_i, \beta_j)\right\} + \frac{1}{2}E\left\{Q^2_\tau(\beta, \tau)\operatorname{var}(\tau)\right\}, \tag{4}$$

evaluated at b and s^2, where

$$\partial^2 = \frac{(n-p)s^2}{n-p-2}\left\{1 + x^f (X'X)^{-1} x^{f'}\right\}$$

$$\partial^2_{(r)} = \sum_{i=k-r+1}^{k} b_i^2 \psi^*_{ij} + s^2$$

$$T = \sum_{i=k-r+1}^{k} \left(x^f_i - \eta^*_i\right) b_i.$$

Proof.

$$D_{KL} = \int p\left(y^f \mid x^f, \Im\right) \log\left\{p\left(y^f \mid x^f, \Im\right) / p_{(r)}\left(y^f \mid x^{f*}, \Im\right)\right\} dy^f$$

$$= \int p\left(y^f \mid x^f, \Im\right) \log p\left(y^f \mid x^f, \Im\right) dy^f - \int p\left(y^f \mid x^f, \Im\right) \log p_{(r)}\left(y^f \mid x^f, \Im\right) dy^f$$

$$= \log\left(2\pi\delta^2\right)^{-1/2} - 1/2$$

$$- \int p\left(y^f \mid x^f, \Im\right) \log\left[N\left(\sum_0^{k-r} x_i^f b_i + \sum_{k-r+1}^{k} \eta_i^* b_i, \sum_{k-r+1}^{k} b_i b_j \psi_{ij}^* + s^2 \right) \right] dy^f$$

$$+ \int p\left(y^f \mid x^f, \Im\right) \log\left\{ 1 + \sum_0^k Q_{ij}^*(\beta, \tau) \operatorname{cov}\left(\beta_i, \beta_j\right)/2 + Q_\tau^2(\beta, \tau) \operatorname{var}(\tau)/2 \right\} dy^f$$

$$= \log\left(2\pi\delta^2\right)^{-1/2} - 1/2$$

$$- \left\{ \log\left(2\pi\delta_{(r)}^2\right)^{-1/2} - \delta^2 \Big/ 2\delta_{(r)}^2 - \left[\sum_0^k x_i^f b_i - \sum_0^{k-r} x_i^f b_i - \sum_{k-r+1}^{k} \eta_i b_i, \right]^2 \right.$$

$$\left. - \int p\left(y^f \mid x^f, \Im\right) \left\{ \sum_0^k Q_{ij}(\beta, \tau) \operatorname{cov}\left(\beta_i, \beta_j\right)/2 + Q_\tau^2(\beta, \tau) \operatorname{var}(\tau)/2 \right\} dy^f, \right.$$

neglecting $O(1/n)$ terms in logarithmic series

$$= T^2/2\delta_{(r)}^2 + \left[\delta^2/2\delta_{(r)}^2 - \log\left\{ \delta^2/2\delta_{(r)}^2 \right\} - 1 \right] \Big/ 2$$

$$- \sum_{ij=0}^{k} E\left\{ Q_{ij}^*(\beta, \tau) \operatorname{cov}\left(\beta_i, \beta_j\right)/2 \right\} + E\left\{ Q_\tau^2(\beta, \tau) \operatorname{var}(\tau)/2 \right\}$$

evaluated at b and s^2.

If x_i^f's are independent then the corresponding approximate D_{KL} is same as (4) but replacing ψ_{ij}^* by ψ_i^2, η_i^* by η_i and $E\{Q_{ij}^*(\beta, \tau)\}$ by $E\{Q_{ij}(\beta, \tau)\}$.

4. Illustration

We use the data in a four-variable problem given by Hald (1952). The response y is the amount of heat evolved in calories per gram of cement. The explanatory variables x_1, x_2, x_3, and x_4 are the amount of tricalcium aluminate, tricalcium silicate, calcium aluminium ferrate and dicalcium silicate, respectively. For the analysis of the data, 20 future sample points x_i^f are chosen within the region of previous experience.

We assume that in a natural informative experiment the values of the variables x arise randomly from the same distribution as for the future variables x_i^f. We also assume that the marginal densities of the future variables x_i^f's are normal with means and variances are their sample means and sample variances, respectively. We want to observe here how much information is lost due to missing future variables and to test whether the loss of information is negligible or not. We also want to compare this lose of information with the variable deletion case where missing variables will be deleted from the data set and a reduced model will be constructed. Then the predictive density will be derived based on the reduced model. The discrepancy between the predictive density based on full model and the predictive density based on reduced model is considered as lose of information of the variable deletion case.

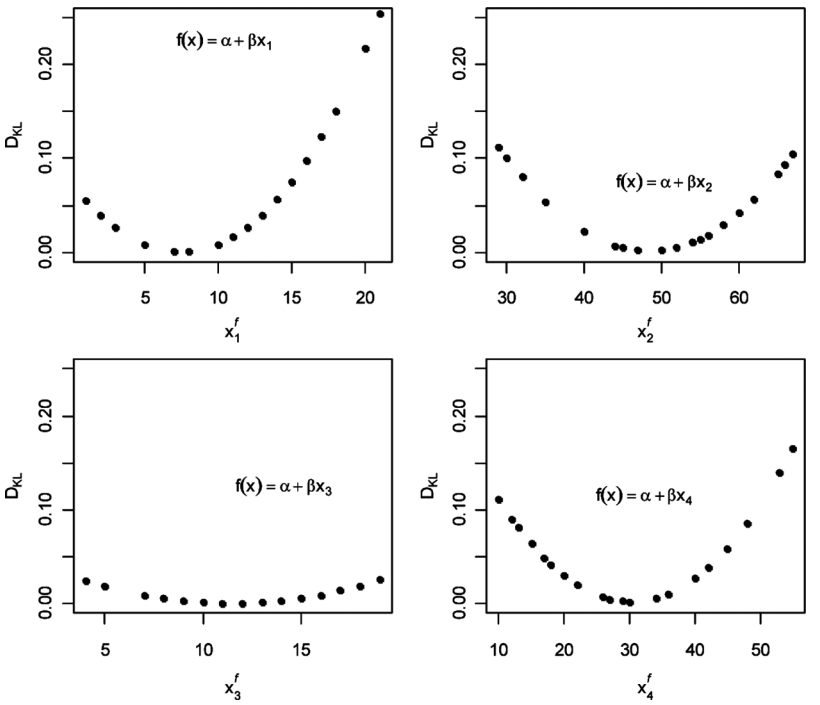

Figure 1. K–L Divergence for single variable case.

In this example, the single most influential variable is x_1 and the least influential variable is x_3. The best pair is (x_1, x_2), and (x_1, x_4) is the second best pair (see Bhattacharjee, 1987; Draper and Smith, 1966). The curves of D_{KL} for missing x_i^f when the predictive density based on any single variable x_i, $i = 1, \ldots, 4$ are given in Fig. 1. We see from the figure that the discrepancies are large for missing x_1^f and negligible for missing x_3^f. We also see that discrepancies are minimum around the mean of the missing variable.

The plots of D_{KL} for missing any single variable when the predictive density based on x_1x_2, x_1x_3, and x_1x_4 are given in Figs. 2(a)–(c). Here, we see that D_{KL} is large due to missing x_1^f in any combination with the other variables and negligible discrepancies occurred when variable x_3^f is missing. We also observe that D_{KL} is minimum at the appropriate combination with the other explanatory variables. The discrepancies are larger toward the ends of the missing variables.

Box plots of D_{KL} for missing a single variable when predictive density is based on all four variables are given in Fig. 3. From this figure it is clear that if a single variable is missing among the four variables, then the summary measures of the discrepancies D_{KL} due to missing x_1^f is larger than the others and summary measures of D_{KL} due to missing x_3^f are negligible.

In Fig. 4, only two variables x_1 and x_2 are considered. Then box plots are shown to compare the discrepancies due to missing variable x_i and due to deleting variable x_i from the data set, $i = 1, 2$. We see that discrepancies for missing variables are smaller than the variable deletion case.

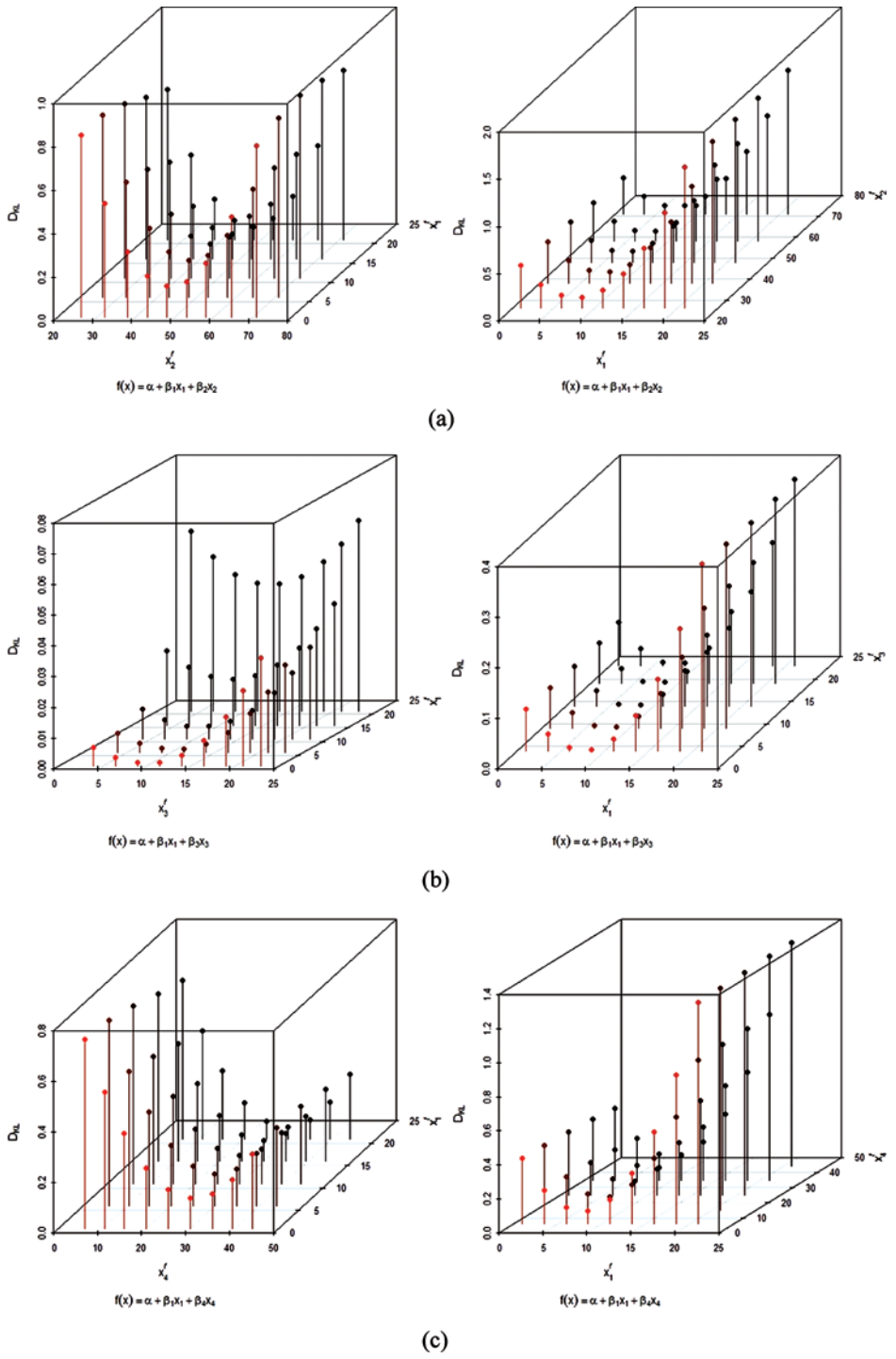

Figure 2. (a) D_{KL} for missing a single variable when predictive density based on x_1 and x_2, (b) D_{KL} for missing a single variable when predictive density based on x_1 and x_3, and (c) D_{KL} for missing a single variable when predictive density based on x_1 and x_4. (color figure available online.)

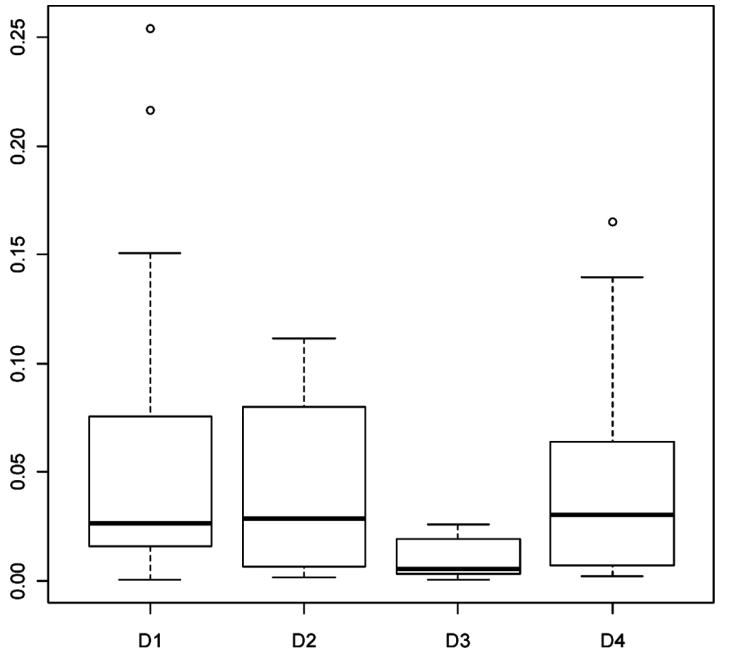

Figure 3. Box plots of D_{KL} for missing any single variable when predictive density based on four variables.

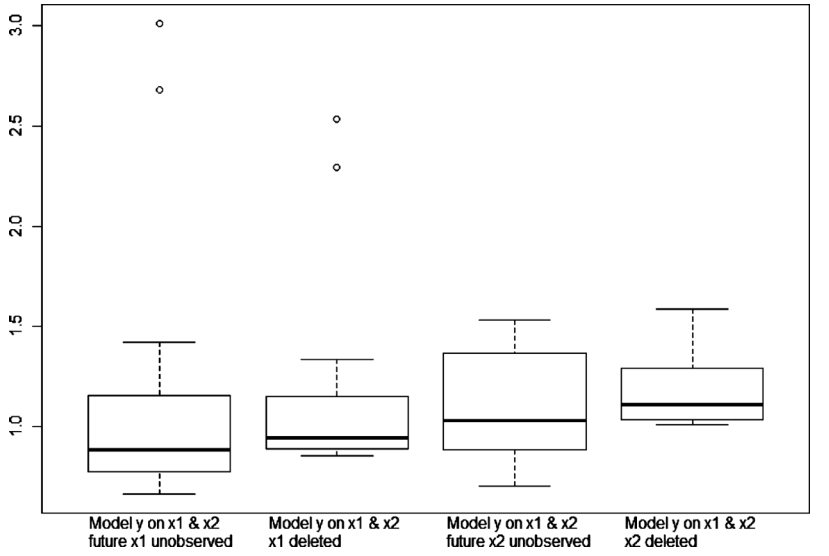

Figure 4. Box plots of D_{KL} for comparison of the two methods.

5. Conclusion

The minimum discrepancies occur around the mean of the missing variable. We also see that the discrepancy depends on the influence of the missing variable. Larger the

influence of the missing variable, more the discrepancy, and less influential variable will produce small discrepancy. Minimum discrepancies occur at the appropriate combination of the explanatory variables. The discrepancies due to missing variables are less than the discrepancies due to deleting the missing variables from the data set. As in Bhattacharjee and Dunsmore (1995), one can test whether any discrepancy is negligible or not.

References

Bhattacharjee, S. K. (1987). Influence of variables in Bayesian prediction. Unpublished Ph.D. Thesis, Department of Probability and Statistics, University of Sheffield, Sheffield, England.

Bhattacharjee, S. K., Dunsmore, I. R. (1995). The predictive influence of variables in a normal regression model. *J. Inform. Optimiz. Sci.* 16(2):327–334.

Bhattacharjee, S. K., Dunsmore, I. R. (1991). The influence of variables in a logistic model. *Biometrika* 78:851–856.

Bjornstad, B. J. F. (1990). Predictive likelihood: a review. *Statist. Sci.* 5:242–265.

Csiszar, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweiz der Ergodizitat von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutato Int. Kozl* 8:85–108.

Draper, N. R., Smith, H. (1966). *Applied Regression Analysis*. New York: John Wiley.

Hald, A. (1952). *Statistical Theory and Engineering Applications*. New York: John Wiley.

Jammalamadaka, S. R., Tiwari, R. C., Chib, S. (1987). Bayes prediction in the linear model with spherically symmetric errors. *Econ. Lett.* 24:39–44.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* 22:79–86.

Mollah, N. H., Bhattacharjee, S. K. (2008). Predictive influence of variables in a multivariate distribution in presence of perfect multicollinearity. *Commun. Statist. Theor. Meth.* 37(1):121–136.

Weiss, R. E. (1995). The influence of variable selection: a Bayesian diagnostic perspective. *J. Amer. Statist. Assoc.* 90(430):619–625.

Zellner, D., Keller, F., Zellner, G. (2004). Variable selection in logistic regression models. *Commun. Statist. Simul. Computat.* 33(3):787–805.