MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Review Article

Open Access

CrossMark

Predictive influence of variables on the odds ratio and in the logistic model

Abstract

We study the influence of explanatory variables in prediction by looking at the distribution of the log-odds ratio. We also consider the predictive influence of a subset of unobserved future variables on the distribution of log-odds ratio as well as in a logistic model, via the Bayesian predictive density of a future observation. This problem is considered for dichotomous, as well as continuous explanatory variables.

AMS subject classification: Primary 62J12, Secondary 62B10, 62F15

Keywords: predictive density/probability, log-odds ratio, logistic model, predictive influence, missing/unobserved variable, kullback-leibler divergence

Volume 5 Issue 1 - 2017

S K Bhattacharjee,[1] Atanu Biswas,[2] Ganesh Dutta,[3] S Rao Jammalamadaka,[4] M Masoom Ali[5]

[1]Indian Statistical Institute, North-East Centre, Tezpur, Assam-784028, India
[2]Indian Statistical Institute, India
[3]Basanti Devi College, India
[4]Department of Statistics and Applied Probability, University of California, USA
[5]Department of Mathematical Sciences, Ball State University, USA

Correspondence: S Rao Jammalamadaka, Department of Statistics and Applied Probability, University of California, USA, Email rao@pstat.ucsb.edu

Received: October 01, 2016 | **Published:** February 01, 2017

Introduction

Odds ratio (OR) is perhaps the most popular measure of treatment difference for binary outcomes and is extensively used in dealing with 2×2 tables in biomedical studies and clinical trials. The distribution of the log of sample OR is often approximated by a normal distribution with true log OR as the mean and with variance estimated by the sum of the reciprocal of the four cell frequencies in the 2×2 table Breslow.[1] Böhning et al.,[2] provide detailed book-length discussion on the OR. For logistic regression, ORs enable one to examine the effect of explanatory variables in that relationship.

Logistic link is perhaps the most popular way to model the success probabilities of a binary variable. Pregibon,[3] Cook and Weisberg[4] and Johnson[5] have considered the problem of the influence of observations for logistic regression models. Several measures have been suggested to identify observations in the data set which are influential relative to the estimation of the vector of regression coefficients, the deviance, the determination of predictive probabilities and the classification of future observations.

Bhattacharjee & Dunsmore[6] considered the effect on the predictive probability of a future observation of the omission of subsets of the explanatory variables. Mercier et al.,[7] used logistic regression to determine whether age and/or gender were a factor influencing severity of injuries suffered in head-on automobile collisions on rural highways. Zellner et al.,[8] considered the problem of variable selection in logistic regression to compare the performance of stepwise selection procedures with a bagging method.

In the present paper, our aim is to measure the predictive influence of a subset of explanatory variables in log-odds ratio of a logistic model using a Bayesian approach. We are also interested in studying the effect of missing future explanatory variables on Bayes prediction, on a logistic model as well as on the log-odds ratio.

In Section 2, we derive the predictive densities of a future log-odds ratio for both the full model and a subset deleted model. We derive the predictive density of log-odds ratio in Section 3, when a subset of future explanatory variables is missing. To derive the predictive densities we assume that the future explanatory variables x^f are distributed as multivariate normal, both when these x^f's are independent or dependent. In Section 4, we discuss the influence of future missing explanatory variables by considering the predictive probability of a future response in a logistic model. This is done by assuming that the future explanatory variables x^f are multivariate normal for the continuous case. Also considered is the dichotomous case. Since the predictive probabilities are not mathematically tractable for the logistic model, we use several approximations.

In Section 2 and 3 we employ Kullback-Leibler[9] directed measure of divergence D_{KL} to assess the influence of variables and also the influence of future missing variables on the log-odds ratio. The form of the Kullback-Leibler[9] measure used here is given by

$$D_{KL} = \int f\left(\mathrm{a}'W^f|.\right)\log\left(\frac{f\left(\mathrm{a}'W^f|.\right)}{f_{(r+s)}\left(a'W^f|.\right)}\right)d\left(a'W^f\right).$$

To assess the influence of missing future variables or to measure the predictive probability in a logistic model we use the absolute difference of the two predictive probabilities.

Influence of variables in log-odds ratio

Consider a phase III clinical trial with two competing treatments, say A and B, having binary responses. Suppose n patients are randomly allocated with n_A and n_B patients to treatments A and B respectively. The patient responses are influenced by a covariate vector $x^{p\times 1}$ where one component of x may be 1 (which covers the constant term). Let $(Y_i; Z_i; x_i)$ be the data corresponding to its patient, where Y_i is the indicator of response (Y_i =1 or 0 for a success or failure), z_i is the indicator of the treatment assignment (z_i = 1 or 0 according as treatment A or B is applied to the its patient), and x is the covariate vector. We assume a logit model for the responses:

$$\Pr\left(Y_i=1|Z_i,x_i\right)=\frac{\exp\left(\Delta Z_i+x_i\beta\right)}{1+\exp\left(\Delta Z_i+x_i\beta\right)}\ i=1,2,....,n. \quad \text{(i)}$$

Then the odds for treatments A and B with covariate vector x_i are respectively

$$O_A=\frac{\Pr\left(Y_i=1|Z_i=1,x_i\right)}{\Pr\left(Y_i=0|Z_i=1,x_i\right)}=\exp\left(\Delta+x_i\beta\right)$$

$$O_B=\frac{\Pr\left(Y_i=1|Z_i=0,x_i\right)}{\Pr\left(Y_i=0|Z_i=0,x_i\right)}=\exp\left(x_i\beta\right)$$

and hence the log-odds ratio is

$$\log OR=\frac{\log O_A}{\log O_B}=\Delta$$

Let us partition

$$x\beta=x_A\beta_A+x_B\beta_B+x_{AB}\beta_{AB}$$

Where x_A indicates the variables used in treatment A only, x_B is for treatment B only, and x_{AB} is for both treatments A and B. Then the model can be partitioned for treatments A and B as:

$$\log O_A=u=\Delta+x_A\beta_A+x_{AB}\beta_{AB}=x_{(A)}\beta_{(A)} \quad \text{(ii)}$$

$$\log O_B=v=x_A\beta_B+x_{AB}\beta_{AB}=x_{(B)}\beta_{(B)} \quad \text{(iii)}$$

The predictive density of future log-odds for A, u^f, for non-informative prior (vague prior) with normal or any spherical symmetric errors is of Student form Jammalamadaka et al. [10] and is given by

$$f\left(u^f\left|x^f_{(A)},data\right.\right)\equiv St\left(n-k,x^f_{(A)}\hat{\beta}_{(A)},s^2_{(A)}\left(1+x^{f'}_{(A)}\left(x'_{(A)}x_A\right)^{-1}x^f_{(A)}\right)\right)$$

where $\hat{\beta}_{(A)}$ is the MLE of $\beta_{(A)}$, $s^2_{(A)}$ is the MLE of σ^2_A and k is the number of parameters in the model (ii). See Bhattacharjee et al. [11] in this context. If the sample size is large then this predictive density can be well approximated by its asymptotic normal form

$$N\left(x^f_{(A)}\hat{\beta}(A),\ s^2_{(A)}\left(1+x^{f'}_{(A)}\left(x'_{(A)}x_{(A)}\right)^{-1}x^f_{(A)}\right)(n-k)/(n-k-2)\right).$$

Similarly one can find the same for treatment B, v^f.

Let us define $w^f=\left(u^f,v^f\right)'$ and $a=(1,-1)'$. Then the predictive density of future log odds ratio $a'w^f$ is given by

$$f\left(a'w^f|x^f_{(A)},x^f_{(B)},data\right)\approx N\left(\theta,\delta^2\right) \quad \text{(iv)}$$

Where

$$\theta=x^f_{(A)}\hat{\beta}_{(A)}-x^f_{(B)}\hat{\beta}_{(B)} \quad \text{and}$$

Our interest is to measure the influence of explanatory variables in the predictive density (iv) for the following cases:

Case 1: Influence of r explanatory variables x^r_A of x_A in treatment A.

Case 2: Influence of r explanatory variables x^r_B of x_B in treatment B.

Case 3: Influence of s explanatory variables x^s_{AB} of x_{AB} in treatment A.

Case 4: Influence of s explanatory variables x^s_{AB} of x_{AB} in treatment B.

Case 5: Joint influence of r explanatory variables x^r_A of x_A and s explanatory variables x^s_{AB} of x_{AB} in treatment A.

Case 6: Joint influence of r explanatory variables x^r_B of x_B and s explanatory variables x^s_{AB} of x_{AB} in treatment B.

To see the influence of explanatory variables in log-odds ratio, we construct a reduced log-odds model deleting a subset of explanatory variables. Then we derive the predictive density of future log-odds ratio for reduced model and compare it with the predictive density (iv) for full model. It is enough to consider Case 5 for illustration. We construct the reduced model by deleting variables x^r_A of x_A and x^s_{AB} of x_{AB} in (ii) as

$$u=\Delta+x^*_A\beta^*_A+x^*_{AB}\beta^*_{AB}=x^*_{(A)}\beta^*_{(A)}$$

Then the predictive density of u^f is given by

$$f\left(u^f|x^{*f}_{(A)},data\right)=St\left(n-k+r+s,x^{*f}_{(A)}\hat{\beta}^*_{(A)},\ S^{*2}_{(A)}\left(1+x^{*f'}_{(A)}\left(x^{*'}_{(A)}x^*_{(A)}\right)^{-1}x^{*f}_{(A)}\right)\right)$$

The normal approximation of the predictive density is

$$N\left(x^{*f}_{(A)}\hat{\beta}^*_{(A)},s^{*2}_{(A)}\left(1+x^{*f'}_{(A)}\left(x^{*'}_{(A)}x^*_{(A)}\right)^{-1}x^{*f}_{(A)}\right)(n-k+r+s)/(n-k+r+s-2)\right)$$

Since no variable is missing in $\upsilon=\log O_B$, the predictive density of υ^f is unaltered along with its normal approximation. Hence the predictive density of log-odds ratio $a'w^f$ under Case 5 is given by

$$f_{(r+s)}\left(a'w^f|x^{*f}_{(A)},x^f_{(B)},data\right)\approx N\left(\theta^*,\delta^{*2}\right) \quad \text{(v)}$$

Where

$$\theta^*=x^{*f}_{(A)}\hat{\beta}^*{}_{(A)}-x^f_{(B)}\hat{\beta}_{(B)}$$

and

$$\delta^{*2}=s^{*2}_{(A)}\left(1+x^{*f'}_{(A)}\left(x^{*'}_{(A)}x^*_{(A)}\right)^{-1}x^{*f}_{(A)}\right)(n-k+r+s)/(n-k+r+s-2)$$

$$+s^2_{(B)}\left(1+x^{f'}_{(B)}\left(x'_{(B)}x_{(B)}\right)^{-1}x^f_{(B)}\right)(n-q)/(n-q-2)$$

To access the influence of the deleted variables we employ the Kullback-Leibler[9] directed measure of divergence D_{KL} between the predictive densities of $a'w^f$ for full model (iv) and reduced model

(v). The form of K-L measure used here is given by

$$D_{KL} = \int f_{(r+s)}\left(a'\omega^{f}|.\right)\log\left(\frac{f_{(r+s)}\left(a'w^{f}|.\right)}{f\left(a'w^{f}|.\right)}\right)da'\omega^{f}$$

The discrepancy measure D_{KL} between the predictive densities (iv) and (v) reduces to

$$D_{KL} = \frac{(\theta-\theta^{*})^{2}}{2\delta^{2}} + \frac{1}{2}\left(\frac{\delta^{*2}}{\delta^{2}} - \log\left(\frac{\delta^{*2}}{\delta^{2}}\right) - 1\right)$$

Here $L = \frac{\left(\theta-\theta^{*}\right)^{2}}{2\delta^{2}}$ is due to difference of location parameters and $S = \frac{1}{2}\left(\frac{\delta^{*2}}{\delta^{2}} - \log\left(\frac{\delta^{*2}}{\delta^{2}}\right) - 1\right)$ due to difference of scale parameters of the two predictive densities (iv) and (v).

Example 1: Here we have considered a flu shot Data Pregibon.[3] A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded Y=1; and a client who did not receive a flu shot was coded Y=0. In addition, data were collected on their age (x_1) and their health awareness (x_2). Also included in the data were client gender (x_3), with males coded $x_3 = 1$ and females coded $x_3 = 0$. Here we have divided whole data set into two groups *A* and *B* on the basis of gender that is group *A* corresponds to the male and group *B* corresponds to the female. We have computed D_{KL} to measure the influence of the deleted variable x_1 in group *A* and *B* separately and the discrepancies are drawn in Figure 1.

Similar figure can be obtained by deleting x_2. From this figure the discrepancy is less around the mean of the deleted variable.

Example 2: This is a simulation exercise. Here we have drawn sample of size 159 from bivariate normal distribution and we have used means, variances and correlation coefficient of x_1 and x_2 of the above flu shot data of size 159 for generating the sample. Now using these x_1 and x_2, we got response that is Y values and thereafter using this whole generated data set we have computed D_{KL}. Now we have repeated whole process 1000 times and computed means of D_{KL}s. The mean discrepancies are shown in Figure 2. Here we get the same conclusion as in the data example.

Influence of missing future explanatory variables in log-odds ratio

Here the aim is to detect the predictive influence of a set of missing future explanatory variables in log-odds ratio of logistic model (i). Our interest is to detect the influence of missing future explanatory variables in the six cases pointed out in Section 2. Let in treatment *A*, r future variables missing from x_A^f and s future variables missing from x_{AB}^f be denoted by $x_{(A)}^{(r+s)f}$. Similarly in treatment *B*, r future missing variables from x_B^f and s future variables missing from x_{AB}^f be denoted by $x_{(B)}^{(r+s)f}$. We assume that the errors of models (ii) and (iii) are normally distributed with zero means and variances $\tau_{(A)}^{-1}$ and $\tau_{(B)}^{-1}$, respectively. We also assume that the conditional density of $x_{(A)}^{(r+s)f}$ given $x_{(A)}^{*f}$ is independent of $\beta_{(A)}$ and $\tau_{(A)}$ and $x_{(B)}^{(r+|s)f}$ given $x_{(B)}^{*f}$ is independent of $\beta_{(B)}$ and $\tau_{(B)}$, i.e.,

$$f\left(x_{(.)}^{(r+s)f}|x_{(.)}^{*f},\beta_{(.)},\tau_{(.)}\right) = f\left(x_{(.)}^{(r+s)f}|x_{(.)}^{*f}\right)$$

where $x_{(.)}^{*f}$ denotes the future explanatory variables $x_{(.)}^{f}$ without $x_{(.)}^{(r+s)f}$.

Explanatory variables are continuous

We assume that x_i^f's are dependent and the distribution of $x_{(A)}^f$ is $(k-1)$-dimensional multivariate normal, i.e. $f\left(x_{(A)}^f\right) \equiv N_{k-1}(\eta, \psi)$.

The conditional density of $x_{(A)}^{(r+s)f}$ given $x_{(A)}^{*f}$ is given by

$$f\left(x_{(A)}^{(r+s)f}|x_{(A)}^{*f}\right) \equiv N_{r+s}\left(\eta_{(r+s)}^{*},\psi_{(r+s)}^{*}\right),$$

Where

$$\eta = \left(\eta^{*},\eta_{r+s}\right), x_{(A)}^{f} = \left(x_{(A)}^{*f},x_{(A)}^{(r+s)f}\right),$$

$$\psi = \begin{pmatrix}\psi_{11} & \psi_{12}\\ \psi_{21} & \psi_{22}\end{pmatrix}, \eta_{r+s}^{*} = \eta_{r+s} + \psi_{21}\psi_{11}^{-1}\left(x_{(A)}^{*f} - \eta^{*}\right)$$

and $\psi_{(r+s)}^{*} = \psi_{22} - \psi_{21}\psi_{11}^{-1}\psi_{12}$.

As earlier it is enough to consider Case 5 to see the joint influence of r missing future explanatory variables x_A^{rf} of x_A^f and s missing future explanatory variables x_{AB}^{sf} of x_{AB}^f in treatment *A*. The density of u^f when $x_{(A)}^{(r+s)f}$ is missing is given by

$$f\left(u^f|x^{*f}_{(A)},\beta(A),\tau(A)\right)=\int f\left(u^f|x^f_{(A)},\beta(A),\tau(A)\right)f\left(x^{(r+s)f}_{(A)}|x^{*f}_{(A)}\right)dx^{(r+s)f}_{(A)}\equiv N\left(\sum_{i=0}^{k-r-s-1}x^f_{(A)i}\hat{\beta}(A)i+\sum_{i=k-r-s}^{k-1}\eta^*_i\hat{\beta}(A)i,\ \sum_{i=k-r-s}^{k-1}\hat{\beta}(A)i\hat{\beta}(A)j\psi^*_{ij}+\tau^{-1}_{(A)}\right)$$

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

See Bhattacharjee et al.,[11] in this context. Using Taylor's expansion and improper prior density for both $\beta_{(A)}$ and $\tau_{(A)}$, the approximate predictive density of u^f when $x^{(r+s)f}_{(A)}$ is missing is given by

$$f_{(r+s)}\left(u^f|x^{*f}_{(A)},\ data\right)\equiv N\left(\sum_{i=0}^{k-r-s-1}x^f_{(A)i}\hat{\beta}_{(A)i}+\sum_{i=k-r-s}^{k-1}\eta^*_i\hat{\beta}_{(A)i},\ \sum_{i,j=k-r-s}^{k-1}\hat{\beta}_{(A)i}\hat{\beta}_{(A)j}\psi^*_{ij}+s^2_{(A)}\gamma^*\right),$$

evaluated at $\hat{\beta}_{(A)}$ and $s^2_{(A)}$ where

$$\gamma^*=\left(1+\frac{1}{2}\sum_0^{k-1}Q^*_{ij}\left(\beta_{(A)},\tau_{(A)}\right)Cov\left(\beta_{(A)i},\beta_{(A)j}\right)+\frac{1}{2}Q^2_{\tau_{(A)}}\left(\beta_{(A)},\tau_{(A)}\right)Var\left(\tau_{(A)}\right)\right)$$

is the multiplicative factor for the second order Taylor's approximation. If $x^f_{(A)}$'s are independent the corresponding approximate predictive density of u^f is

$$f_{(r+s)}\left(u^f|x^{*f}_{(A)},\ data\right)\equiv N\left(\sum_{i=0}^{k-r-s-1}x^f_{(A)i}\hat{\beta}_{(A)i}+\sum_{i=k-r-s}^{k-1}\eta_i\hat{\beta}_{(A)i},\ \sum_{i,j=k-r-s}^{k-1}\hat{\beta}^2_{(A)i}\psi^2_i+s^2_{(A)}\gamma\right)$$

evaluated at $\hat{\beta}_{(A)}$ and $s^2_{(A)}$, where η_i and ψ^2_i are mean and variance of the ith missing variable and

$$\gamma=\left(1+\frac{1}{2}\sum_0^{k-1}Q_{ij}\left(\beta_{(A)},\tau_{(A)}\right)Cov\left(\beta_{(A)i},\beta_{(A)j}\right)+\frac{1}{2}Q^2_{\tau_{(A)}}\left(\beta_{(A)},\tau_{(A)}\right)Var\left(\tau_{(A)}\right)\right)$$

. Since no future variable is missing in υ, the approximate predictive density of υ^f is same as obtained in Section 2. Thus when $x^f_{(A)}$'s are dependent the approximate predictive density of log-odds ratio $a'w^f$ for $x^{(r+s)f}_{(A)}$ missing is given by

$$f_{(r+s)}\left(a'w^f|x^{*f}_{(A)},x^f_{(B)};\ data\right)\equiv\gamma^*N\left(\xi,\omega^2\right),\qquad\text{(vi)}$$

Where

$$\xi=\sum_{i=0}^{k-r-s-1}x^f_{(A)i}\hat{\beta}_{(A)i}+\sum_{i=k-r-s}^{k-1}\eta^*_i\hat{\beta}_{(A)i}-x^f_{(B)}\hat{\beta}_{(B)}$$

and

$$\omega^2=\left(\sum_{i,j=k-r-s}^{k-1}\hat{\beta}_{(A)i}\hat{\beta}_{(A)j}\psi^*_{ij}+s^2_{(A)}\right)+s^2_{(B)}\left(1+x^f_{(B)}\left(X'_{(B)}X_{(B)}\right)^{-1}x^{f'}_{(B)}\right)\frac{n-q}{n-q+2}$$

The Kullback-Leibler [9] directed measure of divergence between the predictive densities (iv) when no variable is missing and the predictive density (vi) when $r+s$ future variables are missing is given by

$$D_{KL}=\int f\left(a'w^f|x^f_{(A)},x^f_{(B)},\ data\right)\log\left(\frac{f\left(a'w^f|x^f_{(A)},x^f_{(B)},\ data\right)}{f_{(r+s)}\left(a'w^f|x^{*f}_{(A)},x^f_{(B)},\ data\right)}\right)da'w^f=\frac{1}{2\omega^2}(\theta-\xi)^2+\frac{1}{2}\left(\frac{\delta^2}{\omega^2}-\log\left(\frac{\delta^2}{\omega^2}\right)-1\right)$$

$$-\frac{1}{2}\sum_{i,j=0}^{k-1}E\left(Q^*_{ij}\left(\beta_{(A)},\tau_{(A)}\right)Cov\left(\tau_{(A)i},\tau_{(A)j}\right)\right)-\frac{1}{2}E\left(Q^2_{\tau_{(A)}}\left(\beta_{(A)},\tau_{(A)}\right)\mathrm{var}\left(\tau_{(A)}\right)\right)\qquad\text{(vii)}$$

If $x^f_{(A)}$'s are independent the predictive density of $a'w^f$ when $(r+s)$ future variables are missing is same as (vi) and the corresponding Kullback-Leibler [9] measure D_{KL} is same as (vii) but replacing η^*_i by η_i in ξ, $\hat{\beta}_{(A)i}\hat{\beta}_{(A)j}\psi^*_{ij}$ by $\hat{\beta}^2_{(A)i}\psi^2_i$ in ω^2 and $Q^*_{ij}\left(\beta_{(A)},\tau_{(A)}\right)$ by $Q_{ij}\left(\beta_{(A)},\tau_{(A)}\right)$ in γ^*, where η_i and ψ^2_i are mean and variance of the ith missing variable.

Explanatory variables are dichotomous

Here we assume that all the explanatory variables are dichotomous and independent. We assume that the errors of models (ii) and (iii) are normally distributed with means zero and variances $\tau^{-1}_{(A)}$ and $\tau^{-1}_{(B)}$ respectively. To assess the influence of the missing variables in treatment A, we consider that $x^f_{(A)i}$ is distributed as

$$\Pr\left(X^f_{(A)i}=x^f_{(A)i}\right)=\theta^{x^f_{(A)i}}_{(A)i}\left(1-\theta_{(A)i}\right)^{1-x^f_{(A)i}},\ x^f_{(A)i}=0,1,\quad i=1,2,\ldots,k-1$$

Citation: Bhattacharjee SK, Biswas A, Dutta G, et al. Predictive influence of variables on the odds ratio and in the logistic model. *Biom Biostat Int J.* 2017;5(1):25–37. DOI: 10.15406/bbij.2017.05.00125

The density of a future u^f is

$$f\left(u^f|x^f_{(A)},\beta_{(A)},\tau_{(A)}\right)\equiv N\left(\sum_{i=0}^{k-1}x^f_{(A)i}\beta_{(A)i},\tau^{-1}_{(A)}\right).$$

If $x^{(r)f}_{(A)}$ future variables are missing in treatment A, then the density of a future u^f is given by

$$f\left(u^f|x^{*f}_{(A)},\beta_{(A)},\tau^{-1}_{(A)}\right)=\sum_{x^f_{(A)k-r}=0}^{1}\cdots\sum_{x^f_{(A)k-1}=0}^{1}N\left(\sum_{i=0}^{k-1}x^f_{(A)i}\beta_{(A)i},\tau^{-1}_{(A)}\right)\prod_{i=k-r}^{k-1}\theta_{(A)i}^{x^f_{(A)i}}\left(1-\theta_{(A)i}\right)^{1-x^f_{(A)i}}.$$

The predictive density of u^f when $x^{(r)f}_{(A)}$ is missing is given by

$$f\left(u^f|x^{*f}_{(A)},\ data=\int f(u^f|x^{*f}_{(A)})\beta_{(A)},\tau^{-1}_{(A)}\right)f\left(\beta_{(A)}|data\right)d\beta_{(A)} \qquad \text{(viii)}$$

which is not mathematically tractable. For vague prior densities for $\beta_{(A)}$ and $\tau_{(A)}$ and using Taylor's expansion, the approximate predictive density of (viii) is

$$f\left(u^f|x^{*f}_{(A)},\ data\right)=\sum_{x^f_{(A)k-r}=0}^{1}\cdots\sum_{x^f_{(A)k-1}=0}^{1}N\left(\sum_{i=0}^{k-1}x^f_{(A)i}\hat{\beta}_{(A)i},s^2_{(A)}\right)\prod_{i=k-r}^{k-1}\theta_{(A)i}^{x^f_{(A)i}}\left(1-\theta_{(A)i}\right)^{1-x^f_{(A)i}}$$

$$\left(1+\sum_{i,j=0}^{k-1}Q_{ij}\left(\hat{\beta},s^{-2}_{(A)}\right)\frac{\mathrm{cov}\left(\beta_{(A)i},\beta_{(A)j}\right)}{2}+Q_{\tau^2_{(A)}}\left(\hat{\beta}_{(A)},s^{-2}_{(A)}\right)\frac{\mathrm{var}\left(\tau_{(A)}\right)}{2}\right)$$

Since there are no missing variables in v^f, the density of v^f is same as that can be obtained in Section 2. Then the predictive density of $a'w^f$ is given by

$$f\left(a'w^f|x^{*f}_{(A)},x^f_{(B)},\ data\right)=\sum_{x^f_{(A)k-r}=0}^{1}\cdots\sum_{x^f_{(A)k-1}=0}^{1}N\left(\sum_{i=0}^{k-1}\left(x^f_{(A)i}\hat{\beta}_{(A)i}-x^f_{(B)i}\hat{\beta}_{(B)i}\right),S^2_{(A)}+s^2_{(B)}\left(1+x^f_{(B)}\left(X'_{(B)}X_{(B)}\right)^{-1}x'_{(B)}\right)\right)$$

$$\prod_{i=k-r}^{k-1}\theta_{(A)i}^{x^f_{(A)i}}\left(1-\theta_{(A)i}\right)^{1-x^f_{(A)i}}$$

$$\left(1+\sum_{i,j=0}^{k-1}Q_{ij}\left(\hat{\beta}_{(A)},s^{-2}_{(A)}\right)\frac{\mathrm{cov}\left(\beta_{(A)i},\beta_{(A)j}\right)}{2}+Q_{T^2_{(A)}}\left(\hat{\beta}_{(A)},s^{-2}_{(A)}\right)\frac{\mathrm{var}\left(\tau_{(A)}\right)}{2}\right) \qquad \text{(ix)}$$

Analytical solution of D_{KL} between the predictive densities (iv) and (ix) is very difficult to obtain but numerical solution can be obtained. In Some situations it is seen that among the explanatory variables, some of the variables are dichotomous and some of the variables are continuous. Among the $k-1$-explanatory variables, without loss of generality we assume that the first l are dichotomous and the remaining last $k-l-1$ are continuous variables. We also assume that out of l dichotomous future variables last d variables are missing and out of $(k-l-1)$ continuous future variables last g variables are missing. Then the predictive density of future log-odds ratio $a'w^f$ when d dichotomous and g continuous variables are missing is given by

$$f\left(a'w^f|x^{*f}_{(A)},x^{*f}_{(B)},\ data\right)=\left(\sum_{x^f_{(A)l-d+1}=0}^{1}\cdots\sum_{x^f_{(A)l}=0}^{1}N\left(\sum_{i=0}^{k-g-1}x^f_{(A)i}\hat{\beta}_{(A)i}+\right.\right.$$

$$\left.\sum_{i=k-g}^{k-1}\eta_i\hat{\beta}_{(A)i}-\sum_{i=0}^{k-1}x^f_{(B)i}\hat{\beta}_{(B)i},\sum_{i=k-g}^{k-1}\hat{\beta}^2_{(A)i}\Psi^2_i+S^2_{(A)}+S^2_{(B)}\left(1+x^f_{(B)}\left(X'_{(B)}X_{(B)}\right)^{-1}x'_{(B)}\right)\right)$$

$$\left.\prod_{i=l-d+1}^{l}\theta_i^{x^f_{(A)i}}\left(1-\theta_i\right)^{1-x^f_{(A)i}}\right)\left(1+\sum_{i,j=0}^{k-1}Q_{ij}\left(\hat{\beta}_{(A)},s^{-2}_{(A)}\right)\frac{\mathrm{cov}\left(\beta_{(A)i},\beta_{(A)j}\right)}{2}\right.$$

$$\left.+Q_{T^2_{(A)}}\left(\hat{\beta}_{(A)},s^{-2}_{(A)}\right)\frac{\mathrm{var}\left(T_{(A)}\right)}{2}\right) \qquad \text{(x)}$$

Citation: Bhattacharjee SK, Biswas A, Dutta G, et al. Predictive influence of variables on the odds ratio and in the logistic model. *Biom Biostat Int J.* 2017;5(1):25–37. DOI: 10.15406/bbij.2017.05.00125

Again, analytical solution of D_{KL} between the predictive densities (iv) and (x) is very difficult but we can obtain its numerical solution. In similar way we can derive the predictive density of future log-odds ratio when some future variables are missing in treatment B.

Example 1 revisited: This example is based on the flu shot data of Example 1. From Figure 3 we have observed same as Examples 1 and 2 that the discrepancies are less around the mean of the missing variables. Moreover we have observed from Figures 1 and 3 that the discrepancies of the missing variables are less as compared to the discrepancies of the deleted variables. **Example 2 revisited:** This example is based on the simulation data of Example 2 and here we have also got same conclusion as Example 1 revisited (Figures 2 & 4).

Examples 1 and 2 revisited: In this example, we have used D_{KL} values for real data for drawing box plots for each cases (deleted and missing). From Figure 5, we have observed that x_2 is more in uential than x_1. Moreover the discrepancies are much less in missing case than deleted case. We have got same result in simulation study and are illustrated in Figure 6.

Evaluation of predictive probability of a logistic model

We consider the logistic model as

$$\Pr\left(y=1|x,\beta\right) = \exp\left(x\beta\right) / \left(1+\exp\left(x\beta\right)\right)$$

The probability that a future response y^f will be a success is given by

$$\Pr\left(y^f=1|x^f,\beta\right) = \exp\left(x^f\beta\right) / \left(1+\exp\left(x^f\beta\right)\right)$$

We assume that the conditional density of $x^{*f}_{(r)}$ given x^{*f} is independent of β where x^{*f} denotes the future explanatory variables without variables $x^f_{(r)}$. Then predictive probabilities of y^f will be a success for models are given by

$$\Pr\left(y^f=1|x^f,\ data\right) = \int \Pr\left(y^f=1|x^f,\beta\right) f\left(\beta|\ data\right) d\beta$$

and

$$\Pr\left(y^f=1|x^{*f},\ data\right) = \int \Pr\left(y^f=1|x^{*f},\beta\right) f\left(\beta|\ data\right) d\beta$$

respectively. Simple analytically tractable priors are not available here. Numerical integration techniques might be used for some specified priors to approximate $\Pr\left(y^f=1|x^f,data\right)$ and $\Pr\left(y^f=1|x^{*f},data\right)$, respectively.

Normal approximation for the posterior density

Let us suppose that the sample size is large. Lindley[12] stated that the posterior density $f\left(\beta|data\right)$ may then be well approximated by its asymptotic normal form as

$$f\left(\beta|data\right) \approx N_p\left(\hat{\beta},\Sigma\right)$$

where $\hat{\beta}$ is the maximum likelihood estimate of β, $\sum = (-H)^{-1}$ and H is the Hessian of log L(β) evaluated at $\hat{\beta}$.

For the logistic model (xi), the Hessian H=($h_{ji}(\hat{\beta})$) evaluated at $\hat{\beta}$ is given by

$$h_{jl}\left(\hat{\beta}\right) = -\sum_{i=1}^{n} \frac{x_{ij}x_{il}\exp\left(x_i\hat{\beta}\right)}{\left(1+\exp\left(x_i\hat{\beta}\right)\right)^2},\ j,l = 0,1,...,k,$$

Where x_{ij} is the jth component of x_i with $x_{i0} = 1$. For given x^f, $z = x^f\beta$ will have approximately a posteriori a normal distribution with mean $x^f\hat{\beta} = x^f\hat{\beta}$ and variance $d^2_{x^f} = x^f\Sigma x^{f'}$, and with probability density function $\phi\left(z\left|b_{x^f},\ d^2_{x^f}\right.\right)$. Using the transformation we can approximate $f\left(\beta\left|x^f,\ data\right.\right)$ by

$$\Pr\left(y^f=1|x^f,\ data\right) \approx \int \frac{\exp(z)}{1+\exp(z)}\phi\left(z|b_{x^f},d^2_{x^f}\right)dz.$$

Analytical evaluation of (xi) is very difficult. We can however evaluate then by numerical integration techniques viz Gauss-Hermite Quadrature (Abramowitz and Stegun[13]), Normal approximation (Cox[14]), Laplace's approximation (de Bruijn[15]).

If the sample size is small, the posterior normality assumption may not be accurate. Therefore, we consider Flat prior approximation (Tierney and Kadane[16]) as an alternative approach using the Laplace's method for integrals.

Effect of the variables x^f

Here we assume that the future variables x^f are dependent and the density of x^f is p-dimensional multivariate normal i.e.

$$f\left(x^f\right) \equiv N_p\left(n,\ \psi\right)$$

The conditional density of $x^f_{(r)}$ for given x^{*f} is

$$f\left(x^f_{(r)}|x^{*f}\right) \equiv N_r\left(n^*_{(r)},\ \psi^*_{(r)}\right)$$

The probability of y^f as a success when $x^f_{(r)}$ is missing given by

$$\Pr\left(y^f=1|x^{*f},\ \beta\right)=\int\frac{\exp\left(x^f\beta\right)}{1+\exp\left(x^f\beta\right)}f\left(x^f_{(r)}|x^{*f}\right)dx^f_{(r)}$$

$$\approx\phi\left(\left(\sum_{i=0}^{k-r}x_i^f\beta_i+\sum_{i=k-r+1}^{k}n_i^*\beta_i\right)\Big/\left(k^2+\sum_{ij=k-r+1}^{k}\beta_i\beta_j\Psi\right)^{1/2}\right)$$

$$=g^*(\beta)\ \text{(Say)}$$

Then the predictive probability of y^f as a success when $x^f_{(r)}$ is missing given by

$$pr\left(y^f=1\left|x^{*f},\ data\right.\right)=\int g^*\left(\beta\right)f\left(\beta\left|data\right.\right)d\beta. \qquad \text{(xii)}$$

The integral in (xii) can be evaluated as the integral in (xi) using Taylor's and Laplace's approximations.

If, instead, the future variables $x_1^f,\ldots,\ x_k^f$ are independently and normally distributed with mean η_i and variance ψ_i^2 (i = 1, 2, … , k), then the conditional density of $x^f_{(r)}$ is

$$f\left(x^f_{(r)}|x^{*f}\right)\equiv f\left(x^f_{(r)}\right)$$

Consequently, we get

$$\Pr\left(y^f=1|x^{*f},\ \beta\right)=\int\frac{\exp\left(x^f\beta\right)}{1+\exp\left(x^f\beta\right)}f\left(x^f_{(r)}\right)dx^f_{(r)}$$

$$\approx\phi\left(\left(\sum_{i=0}^{k-r}x_i^f\beta_i+\sum_{i=k-r+1}^{k}n_i\beta_i\right)\Big/\left(k^2+\sum_{i=k-r+1}^{k}\beta_i^2\Psi_i^2\right)^{1/2}\right)$$

$$=g(\beta)\ \text{(Say)}$$

See Aitchison and Begg [17] in this context. Again,

$$\Pr\left(y^f=1|x^f,\ \text{data}\right)=\int g(\beta)f\left(\beta|\text{data}\right)d\beta$$

Variables x^f are dichotomous

Here we assume that the variables x^f are independent and they can take only two values 0 or 1. We also assume that x_i^f is distributed as

$$\Pr\left(x_i^f=x_i^f\right)=\theta_i^{x_i^f}\left(1-\theta_i\right)^{1-x_i^f}$$

If $x^f_{(r)}$ is missing the probability of y^f as a success is given by

$$\Pr\left(y^f=1|x^{*f},\beta\right)=\sum_{x^f_{k-r+1}=0}^{1}\ldots\sum_{x^f_k=0}^{1}\frac{\exp\left(x^f\beta\right)}{1+\exp\left(x^f\beta\right)}\prod_{i=k-r+1}^{k}\theta_i^{x_i^f}\left(1-\theta_i\right)^{1-x_i^f}=h\left(\beta\right)$$

(Say).

The predictive probability of y^f as a success when $x^f_{(r)}$ is missing is given by

$$\Pr\left(y^f=1|x^{*f},\ \text{data}\right)=\int h(\beta)f\left(\beta|\text{data}\right)d\beta. \qquad \text{(xiii)}$$

If the sample size is large, assuming the normality assumption for the posterior density we can approximate (xiii) using Taylor's theorem, Laplace's method and normal approximation.

Example: one variable case

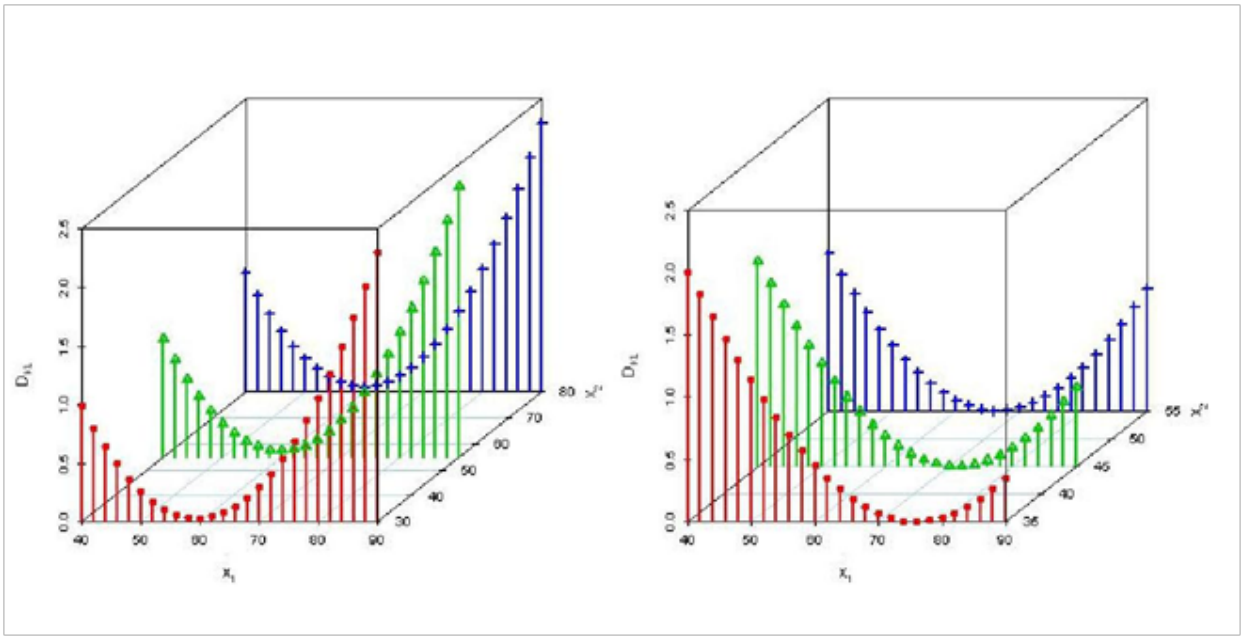
Here we consider two different logistic models based on any single variable either x_1 or x_2. We want to measure the discrepancies between the predictive probability $\hat{p}_i$, based on a single variable x_i when x_i^f is known, and the predictive probability $\hat{p}_0$, based on x_i alone when x_2^f is missing, to assess the influence of the missing variable x_i^f, i = 1, 2. The predictive probability $\hat{p}_i$ is determined using quadrature approximation and the predictive probability $\hat{p}_0$ is determined using second order Taylor's approximation.

We assume that the marginal densities of the future variables x_1^f and x_2^f are normal with means 33.35, 78.24 and variances 65.39, 1827.0 respectively, where means and variances are the estimated sample means and sample variances from the observed data. We employ the absolute difference of probabilities and Kullback-Leibler divergence measure to assess the influence of the missing variable. The discrepancies are drawn in Figure 7. Here we see that the discrepancies due to missing x_1^f in the predictive probability based on x_1 are very large compared to the discrepancies due to missing x_2^f in the predictive probability based on x_2. The discrepancies are less around the mean of the missing variable.

Example: two-variable case

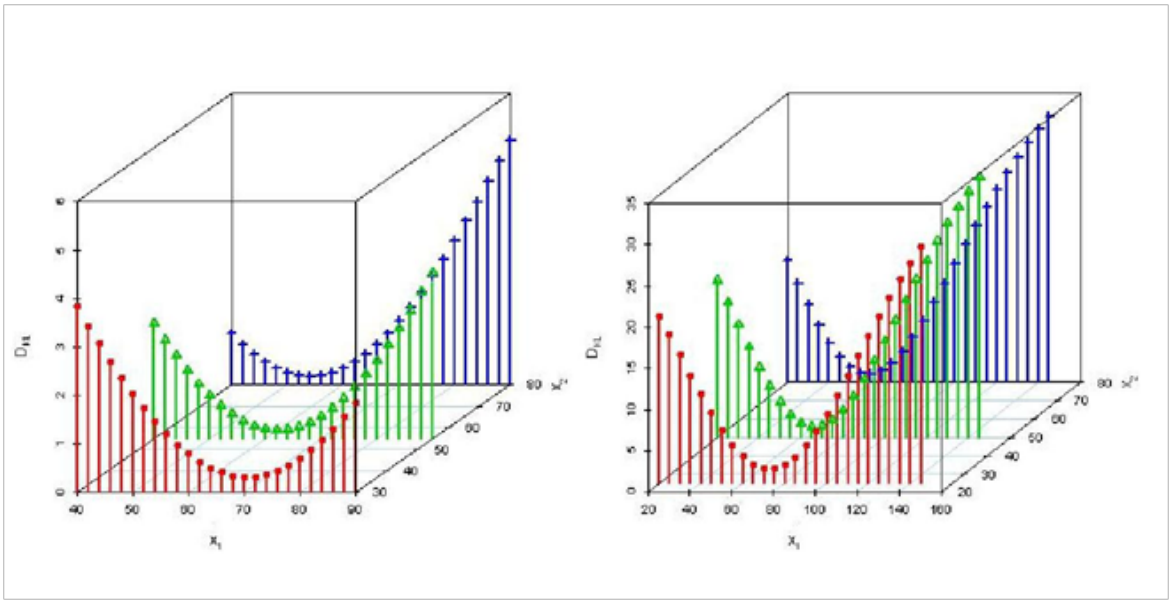
Now we consider that the predictive probability based on two variables x_1^f and x_2^f when both x_1^f and x_2^f are known is denoted by $\hat{p}_{12}$ and the predictive probability $\hat{p}_{ij}$, $i=0,1$, $j=0,2$ and $(i,j)\neq(1,2)$ based on x_1 and x_2 when any future variable is missing. "0" indicates missing variable. Here also the predictive probability $\hat{p}_{12}$ is determined using quadrature approximation and predictive probabilities $\hat{p}_{10}$, $\hat{p}_{02}$ and $\hat{p}_{00}$ are determined using second order Taylor's approximation. Here we assume that the joint density of x_1^f and x_2^f is bivariate normal with correlation coefficient -0.33 which is the estimated sample correlation coefficient from the observed data. The absolute differences of the two predictive probabilities $\hat{p}_{12}$ and $\hat{p}_{02}$ when x_1^f is missing and the absolute differences of the two

predictive probabilities $\hat{p}_{12}$ and $\hat{p}_{10}$ when x_2^f is missing are drawn in Figure 8. Kullback-Leibler directed divergence D_{KL} are drawn in Figure 9. The discrepancies when x_1^f is missing and for different given values of the other variable for both the cases are close together since the correlation between x_1^f and x_2^f are very small. The discrepancies due to missing x_1^f are very large compared to missing x_2^f except near the mean of the missing variable. If both x_1^f and x_2^f are missing the discrepancies are drawn in Figure 10. These discrepancies are very similar to the discrepancies due to missing x_1^f alone in the predictive probability based on x_1 and x_2 since the contribution of x_2 is negligible.



Group A Group B

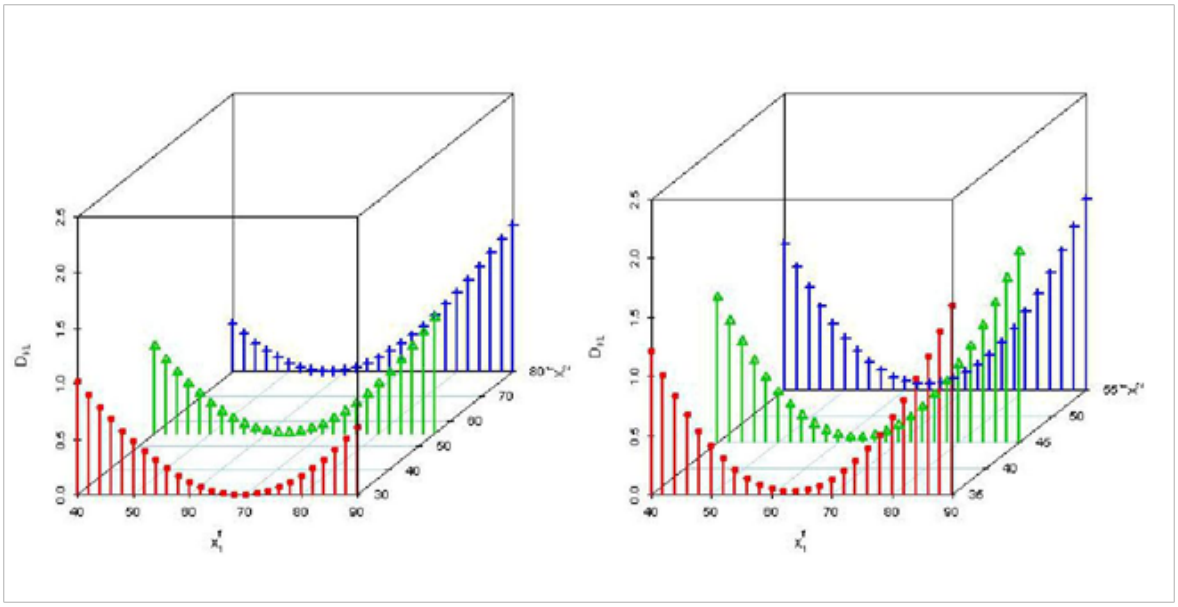
Figure 1 Three dimensional scatter plots based on real data for D_{KL} when x_1 is deleted.



Group A Group B

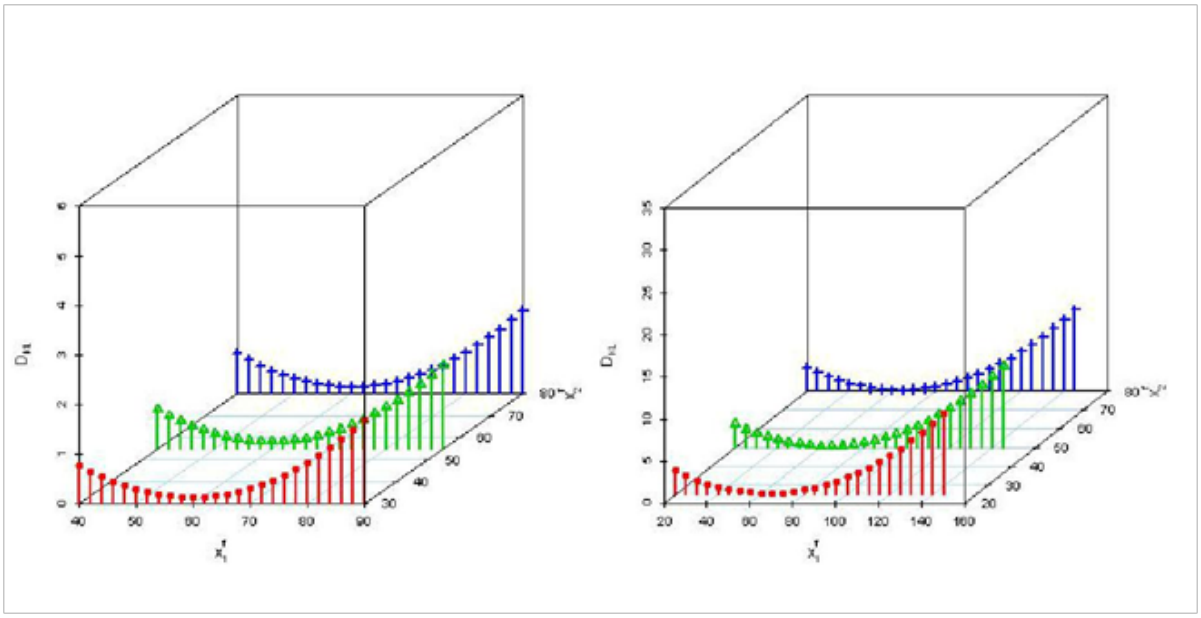
Figure 2 Three dimensional scatter plots based on simulated data for D_{KL} when x_1 is deleted.

Citation: Bhattacharjee SK, Biswas A, Dutta G, et al. Predictive influence of variables on the odds ratio and in the logistic model. *Biom Biostat Int J.* 2017;5(1):25–37. DOI: 10.15406/bbij.2017.05.00125



Group A Group B

Figure 3 Three dimensional scatter plots based on real data for D_{KL} when x^f_1 is missing.



Group A Group B

Figure 4 Three dimensional scatter plots based on simulated data for D_{KL} when x_1 is missing

Citation: Bhattacharjee SK, Biswas A, Dutta G, et al. Predictive influence of variables on the odds ratio and in the logistic model. *Biom Biostat Int J.* 2017;5(1):25–37. DOI: 10.15406/bbij.2017.05.00125

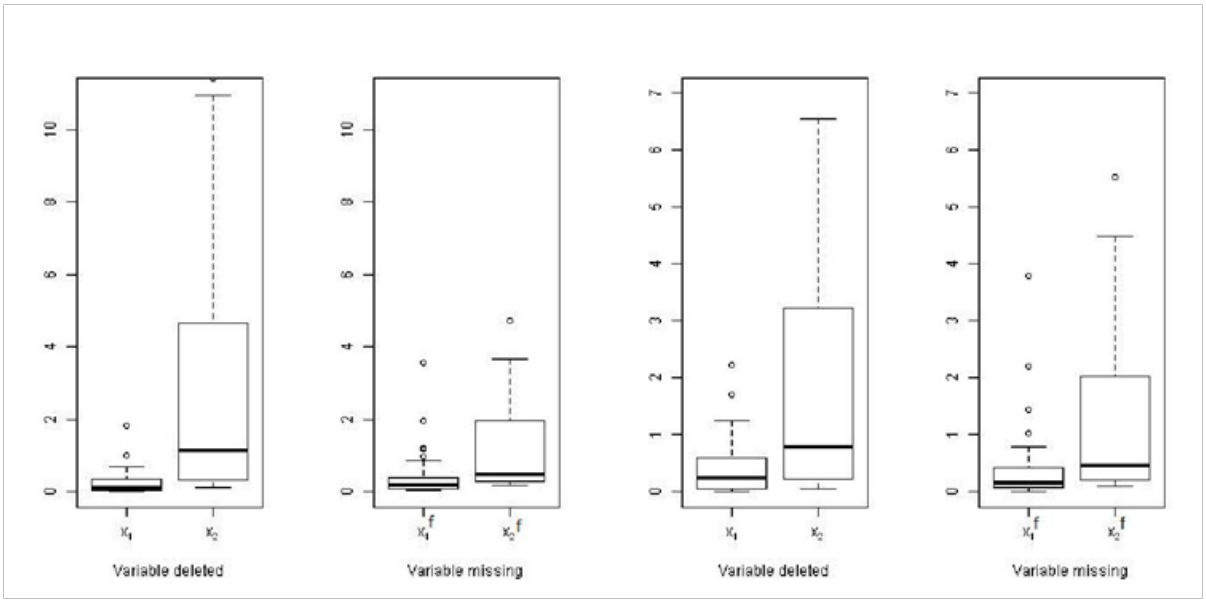


Figure 5 Box plot for D_{KL} based on real data.

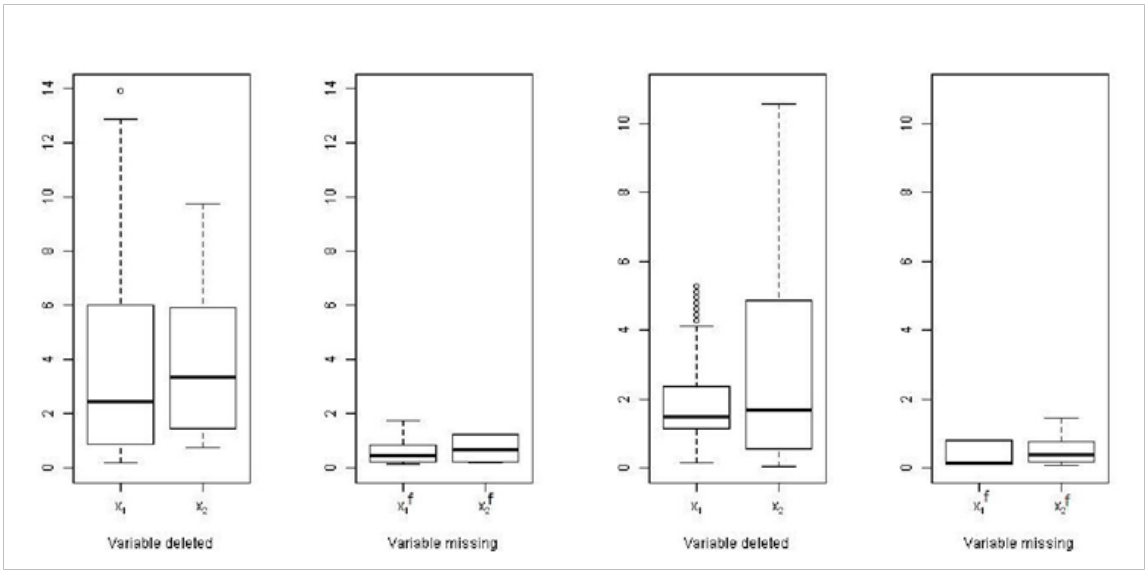
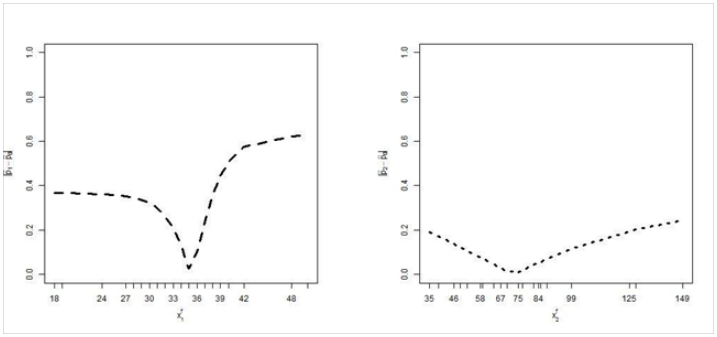


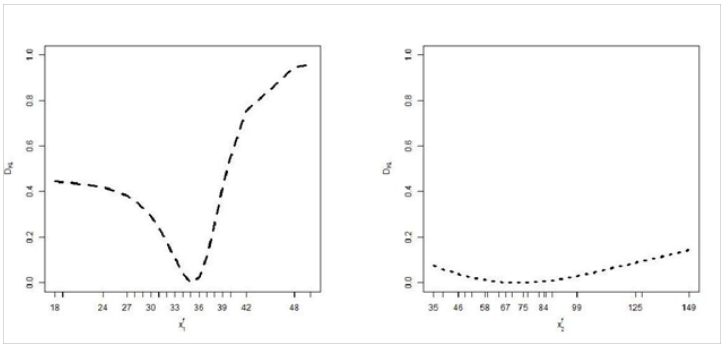
Figure 6 Box plot for D_{KL} based on simulated data.

Citation: Bhattacharjee SK, Biswas A, Dutta G, et al. Predictive influence of variables on the odds ratio and in the logistic model. *Biom Biostat Int J.* 2017;5(1):25–37. DOI: 10.15406/bbij.2017.05.00125



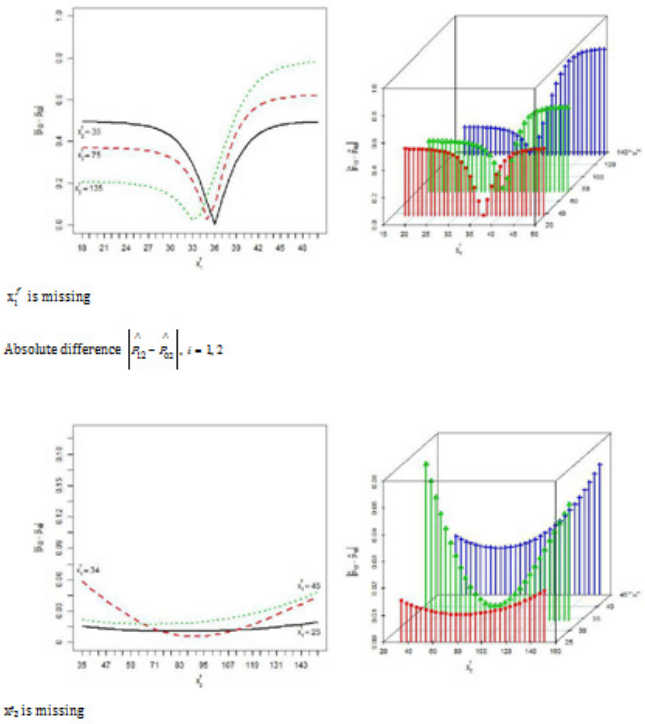
x_i^f is missing $\quad$ x_2^f is missing

Kullback-Leibler directed divergence D_{KL}



x_i^f is missing $\quad$ x_2^f is missing

Figure 7 Absolute difference $\left|\hat{P}_i - \hat{P}_0\right|, i = 1, 2$



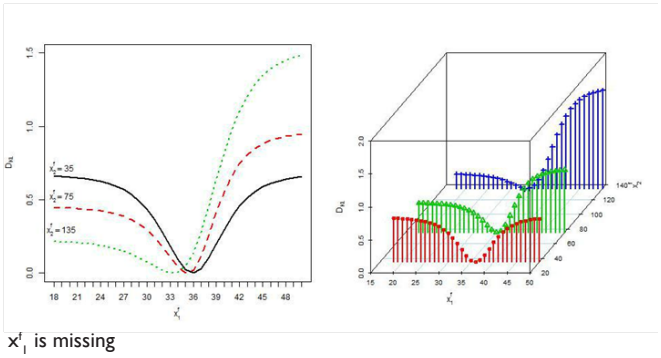
x_1^f is missing

Absolute difference $\left|\hat{p}_{12} - \hat{p}_{02}\right|, i = 1, 2$

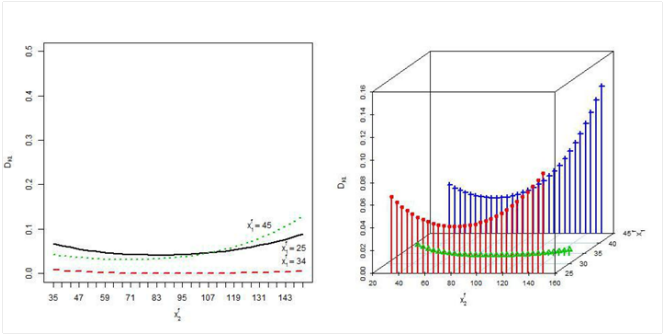
x_2^f is missing

Figure 8 Absolute difference $\left|\hat{p}_{12} - \hat{p}_{10}\right|$

Citation: Bhattacharjee SK, Biswas A, Dutta G, et al. Predictive influence of variables on the odds ratio and in the logistic model. *Biom Biostat Int J.* 2017;5(1):25–37. DOI: 10.15406/bbij.2017.05.00125

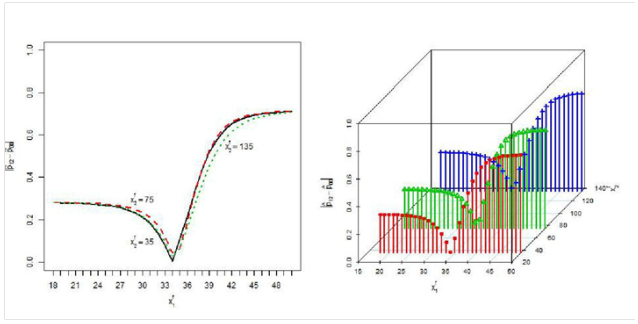


x_1^f is missing

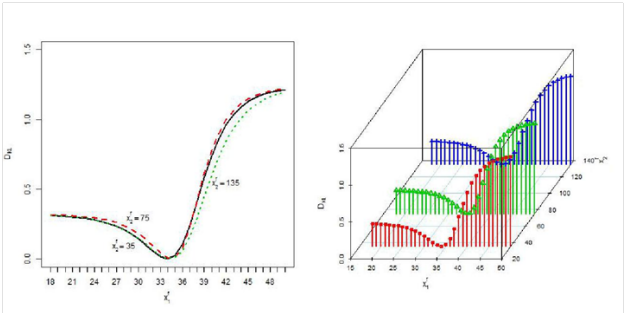


x_2^f is missing

Figure 9 Kullback-Leibler directed divergence D_{KL}



Absolute difference $\left|\hat{P}_{12} - \hat{P}_{00}\right|, \; i = 1, 2$



Kullback-Leibler directed divergence D_{KL}

Figure 10 x_1^f and x_2^f are both missing.

Concluding remarks

In our present study we have observed that the discrepancies are minimum around the mean of the deleted variables as well as the mean of the missing future variables in both the logistic model and the log-odds ratio; the discrepancies are larger if the deleted or missing variables are more influential; the discrepancies in the deleted case are higher than the missing case.

In this present paper we studied the important problem of predictive influence of variables on the log odds ratio under a Bayesian set up. The treatment difference

$$\Pr\left(Y_i=1|Z_i=1,x_i\right)-\Pr\left(Y_i=1|Z_i=0,x_i\right)$$

Or the risk of ratio

$$\Pr\left(Y_i=1|Z_i=1,x_i\right)/\Pr\left(Y_i=1|Z_i=1,x_i\right)$$

can also be studied along the same lines.

We have also considered the influence of missing future explanatory variables in a logistic model. Influence of missing future explanatory variables in a Probit and complementary log-log models can also be studied in similar fashion.

Acknowledgement

None.

Conflict of interest

None.

References

1. Breslow N. Odds ratio estimators when the data are sparse. *Biometrika*. 1981;68:73–84.
2. Bohning D, Kuhnert, Rattanasiri S, et al. Meta–analysis of binary data using pro le likelihood. 1st ed. *A Chapman and Hall CRC Interdisciplinary Statistic*. 2008.
3. Pregibon D. Logistic regression diagnostics. *Annals of Statistics*. 1981;9:705–724.
4. Cook, R Dennis, Weisberg, et al. Residuals and Influence in Regression. USA: New York: *Chapman and Hall*. 1982.
5. Johnson W. Influence measures for logistic regression: Another point of view. *Biometrika*. 1985;72(1):59–65.
6. Bhattacharjee SK, Dunsmore IR. The influence of variables in a logistic model. *Biometrika*. 1991;78(4):851–856.
7. Mercier C, Shelley MC, Rimkus J, et al. Age and Gender as Predictors of Injury Severity in Head–on Highway Vehicular Collisions. The 76th Annual Meeting , Transportation Research Board, Washington, USA. 1997.
8. Zellner D, Keller F, Zellner GE. Variable selection in logistic regression models. *Communications in Statistics*. 2004;33(3):787–805.
9. Kullback S, Leibler R A. On information and sufficiency. *Ann Math Statist*. 1951;22:79–86.
10. S Rao Jammalamadaka, Tiwari RC, Chib Siddhartha. Bayes prediction in the linear model with spherically symmetric errors. *Economics Letters*. 1987;24:39–44.
11. Bhattacharjee SK, Shamiri A, Sabiruzzaman Md, et al. Predictive Influence of Unavailable Values of Future Explanatory Variables in a Linear Model. *Communications in Statistics – Theory and Methods*. 2011;40:4458–4466.
12. Lindley D V. The use of prior probability distributions in statistical inference and decisions. Proc. *4th Berkeley Symp*. 1961;1:453–468.
13. Abramowitz M, Stegun I A . Handbook of Mathematical Functions. USA: *National Bureau of Standards*. 1966.
14. Cox DR. Binary regression. UK: London: *Chapman and Hall*. 1970.
15. De Bruijn N C. Asymptotic Methods in Analysis. Amsterdam, North–Holland. 1961.
16. Tierney L, Kadane, Joseph B, et al. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*. 1986;81(393):82–86.
17. Aitchison J, Bagg CB (1976) Statistical diagnosis when basic cases are not classified with certainty. Biometrika 63: 1-12.
18. Logistic Regression Example with Grouped Data. Regression FluShots, University of North Florida.
19. Bhattacharjee SK, Dunsmore IR. The predictive influence of variables in a normal regression model. *J Inform Optimiz Sci*. 1995;16(2):327–334.