Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/030201

Bayesian Relative Importance Analysis of Logistic Regression Models

*Xiaoyin Wang**

Department of Mathematics, Towson University, Towson, MD 21252, USA

Received: 21 Jul. 2015, Revised: 6 Mar. 2016, Accepted: 8 Mar. 2016
Published online: 1 May 2016

Abstract: The goal of determining the relative importance of predictors is to expose the individual contribution of the predictor in the presence of other predictors within a selected model. In practice, it is often desired to understand the extent to which each predictor variable drives the response variable The purpose of this article is to expand the current research practice to evaluate the relative importance of each predictor in a logistic regression setting by developing a statistical model-based approach in the Bayesian framework. Results from extended simulation studies suggest that the proposed weighted paired comparison model with the two sided power (TSP) link function provides the most effective and reliable measure of the relative importance of predictors.

Keywords: Bayesian Estimate, Markov Chain Monte Carlo Method(MCMC); Paired Comparison; Relative Importance Analysis; Two Sided Power Distribution

1 Introduction

Regression analysis is one of the most important tools used by researchers and practitioners in determining the relationship between a response variable and its predictor variables. Often times there is a desire to further analyze the selected model to determine which predictor or predictors are considered more important the others. Analysis of this type is referred to as the relative importance analysis, and serves as an important supplement to traditional regression analysis. For cases where the predictor variables are uncorrelated, zero-order correlation and regression coefficients nicely provide the answers to these questions. The problem arises when the predictor variables are correlated with each other. Past research has documented how commonly used indices fail to appropriately answer this question when predictors are correlated (Darlington 1968).

Over the years, researchers have suggested various methods for measuring the relative importance of predictors in various subjects such as, statistics, psychology, political economy, organizational research, and medicine; for a comprehensive review see Johnson and LeBreton (2004). Statements have been confused, because there has been a lack of agreement on the definition of the relative importance of a predictor. In an effort to circumvent this issue, the concept of predictor ”dominance” was proposed by Budescu (1993) as a new way to compare predictors in a multiple regression context. Dominance Analysis (Budescu 1993) approaches the problem of relative importance by examining the change in R^2 resulting from adding a predictor to all possible subset regression models. Azen and Traxel (2009) introduces and extends the Dominance Analysis in logistic regression models. The novelty of the approach is that predictors are compared in a pairwise fashion based on a common subset reference model across all possible subset models, and a hierarchy of levels of dominance is established, *i.e.*, complete dominance, conditional dominance and general dominance. Nevertheless, critics have disapproved of Dominance Analysis on the grounds that such analysis is atheoretic. Moreover, the general dominance weight is calculated based on average of conditional dominance weights over difference sizes of subset models, which introduces bias in the aggregation because the additional contribution of a predictor is more substantial in subset models with fewer number of predictors. As the number of predictors increases, the additional contribution of the predictor gets less, for detailed explanation. Wang and Yao (2014) mentions that the value of the general dominance index is influenced to a greater extent by the conditional dominance of the subset model with a smaller number of predictors.

* Corresponding author e-mail: xwang@towson.edu

With regards to these issues, Wang et. al. (2013) proposed an inferential approach to discover the intrinsic dominance ability of predictors from a behavioral study point of view in the Bayesian framework using the Bradley-Terry paired comparison model. Results show that, in most cases, conclusions from the Bayesian approach are consistent with those from the Dominance Analysis methods. However, the basic Bradley-Terry model is weak in identifying the suppressor variable, that is, a predictor that is uncorrelated with the response variable but whose presence improves prediction because of its correlation with other predictors. As a remedy, Wang and Yao (2014) proposed a weighted Bradley-Terry model, which is more efficient and accurate than the basic un-weighted Bradley-Terry model and the Dominance Analysis method. One weak point of those methods is that it is difficult to apply practical meaningful interpretation to the resulting dominance index, which varies from $-\infty$ to ∞. The dominance ability is compared based on the additional R^2, which varies between 0 and 1. It would be desirable to have a dominance index on the 0-1 scale. Wang (2015) applies the cumulative distribution function of the two-sided power (TSP) distribution as the link function in the paired comparison model to ensure the resulting dominance index vary between 0 and 1. The results show that the TSP link function provides similar conclusions as the logit link function but it is more computational efficient.

Linear regression is used to determine the scalar relationship of observed data and its corresponding predicted variable. Logistic regression models are used to predict binary outcomes and the probability of an event occurring (as opposed to not occurring) from a set of predictors. In this article, we extend the current research of Bayesian analysis of predictors' relative importance (Wang et. al, 2013; Wang and Yao 2014; Wang 2015), which initially developed for linear regression, to the logistic regression setting to determine the dominance hierarchy of the predictors. In section 2, we discuss the measures of model adequacy in the logistic regression setting. In section 3, we formulate paired comparison models based on logit and TSP link functions with weighted and un-weighted likelihood functions. In section 4, we introduce the Bayesian inference and the computation procedure. In section 5, we perform simulation studies to evaluate the accuracy of the proposed methods using simple random samples from a known population. In section 6, we report an empirical example of choosing between fixed and adjustable rate mortgages. Finally, we discuss potential future research avenues regarding the relative importance of predictors.

2 Model Adequacy Measures

In linear regression, the measure of goodness-of-fit is defined as the proportion of variation in the response accounted for by the predictors, or the squared correlation between the observed and the predicted responses. This measure is commonly referred as R^2. In logistic regression, however, a R^2 analogues goodness-of-fit measure is required because the classical R^2 measure does not extend to the non-linear case. To extend the dominance analysis to the logistic regression, it is necessary to define what is meant by the additional contribution of a predictor to the prediction model and how to measure this contribution.

Reviews of measures of fit proposed for logistic regression can be found in Amemiya (1981), Menard (2000), Mittlbock and Schemper (1996), and Zheng and Agresti (2000). The following criteria are typically applied for defining appropriate R^2 analogues (Kvalseth, 1985; Vanden Burg and Lewis, 1988, Azen and Traxel, 2009):

– Boundedness: The measure should vary between a minimum of zero, indicating complete lack of fit, and a maximum of one, indicating perfect fit.
– Linear invariance: The measure should be invariant to non-singular linear transformations of the variables (Y's and X's).
– Monotonicity: The measure should not decrease with the addition of a predictor.
– Intuitive interpret-ability: The measure of fit is intuitively interpretable, in that it agrees with the scale of the linear case for intermediate values.

Based on these criteria, the following three R^2 analogues were chosen that satisfied at least three of these four properties.

– McFadden's (1974) measure is defined as
$$R_M^2 = \frac{\ln(L_0) - \ln(L_M)}{\ln(L_0)}$$
– Nagelkerke's (1991) measure is defined as
$$R_N^2 = \frac{1 - (L_0/L_M)^{2/n}}{1 - L_0^{2/n}}$$
– Estrella's (1998) measure is defined as
$$R_E^2 = 1 - \left[\frac{\ln(L_M)}{\ln(L_0)}\right]^{-(2/n)\ln(L_0)}$$

For these measures, L_0 represents the likelihood of the null (intercept-only) model and L_M represents the likelihood of the fitted (intercept and predictors) model. A summary of the properties satisfied by all three measures of fit (R^2 analogues) is presented in Table 1. Using each of these three R^2 analogues, the additional contribution of a given predictor to a specific logistic model can be measured as the change (i.e., increase) in the R^2 analogue when the predictor is added to the model. Azen and Traxel (2009) provided an algebraic proof showing that all three of these measures will produce the same order of dominance analysis results but the magnitude of these measures are different. In this paper, we will show that, applying the proposed Bayesian approach, all these three measure produce the same index of relative importance.

3 Paired Comparison Models

Statistical methods of determining relative dominance abilities based on paired comparisons have long history dated back to Bradley and Terry (1952) in the context of chess tournaments and have been broadly applied in many fields such as statistics, psychometrics, marketing research, preference measurement, sports competition, behavioural study, education, machine learning, and many others. In this paper, we determine the intrinsic dominance ability of predictors by applying the statistical method of paired comparisons based on the additional increase of R^2 analogues amongst all possible subset models.

3.1 Likelihood Function

Suppose there are p predictors in a logistic regression model, $X_1,\cdots,X_p$. Let X^* be a subset of $(X_1,\cdots,X_p)$ and denote $R^2_{Y\cdot X^*}$ as any form of the three R^2 analogues of Y and X^*. Let $\Delta R^2_{Y\cdot X_i|X^*} = R^2_{Y\cdot(X_i,X^*)} - R^2_{Y\cdot X^*}$ be the increase in R^2 analogues by adding X_i to the subset model with X^*. Here, the model with X^* is considered as the baseline reference model. The predictor X_i is said to dominate or "win" X_j if $\Delta R^2_{Y\cdot X_i|X^*} > \Delta R^2_{Y\cdot X_j|X^*}$, that is adding X_i to a model leads to a greater increase in R^2 analogues than would be obtained by adding X_j to the model with the same subsets of other variables.

In a model with p predictors, X_i and X_j are compared amongst all possible 2^{p-2} subset reference models. Let W_{ij} be the number of times the predictor X_i dominating X_j, and θ_{ij} be the probability of the predictor X_i dominating X_j, then W_{ij} has a binomial distribution with parameters $(2^{p-2},\theta_{ij})$. The Bradley-Terry model is based on the assumption that $\theta_{ij} = \xi_i/(\xi_i+\xi_j)$, where ξ_i represents the intrinsic dominance ability of predictor X_i. Let $d_i = ln(\xi_i)$, then d_i can be interpreted as the dominance ability of predictor X_i on the logarithm scale. Based upon this reparameterization, we have $logit(\theta_{ij}) = d_i - d_j$, or $\theta_{ij} = \exp(d_i-d_j)/(1+\exp(d_i-d_j))$, which means that the dominance probability depends only on the difference of dominance indices. Therefore, the Bradley-Terry model assumes the probability that one predictor prevails over another is a logit function (the link function) of the difference in dominance indices between these two predictors. The likelihood function of the results of paired comparisons of the predictors can be written as follows.

$$\prod_{i<j} C_{2^{p-2}}^{w_{ij}} \theta_{ij}^{w_{ij}} (1-\theta_{ij})^{2^{p-2}-w_{ij}} \tag{1}$$

Wang and Yao (2014) points out that the number of times that the predictor X_i and X_j encounter each other depends on the size of the baseline reference model. For a logistic model with p predictors, based on the baseline reference model with k predictors, the number of times that the predictor X_i and X_j encounter is C^k_{p-2}. For example, when $p=4$ and $k=0$, X_1 and X_2 only meet once, that is $R^2_{Y\cdot X_1}$ and $R^2_{Y\cdot X_2}$; when $k=1$, X_1 and X_2 are compared under two baseline reference models, that is $\Delta R^2_{Y\cdot X_1|X_3}$ vs $\Delta R^2_{Y\cdot X_2|X_3}$, and $\Delta R^2_{Y\cdot X_1|X_4}$ vs $\Delta R^2_{Y\cdot X_2|X_4}$; and when $k=2$, X_1 and X_2 only meet once with the same baseline reference model with two predictors X_3 and X_4. Taking this fact into consideration, the weighted Bradley-Terry model is proposed with $C^k_{p-2}/2^{p-2}$ as the weight. Let $w_{ij\cdot k}$ be the number of times that the predictor X_i outweighs X_j with baseline reference model of size k. The likelihood function of the weighted Bradley-Terry model is

$$\sum_{k=0}^{p-2} \frac{C^k_{p-2}}{2^{p-2}} \prod_{i<j} \binom{C^k_{p-2}}{w_{ij\cdot k}} \theta_{ij}^{w_{ij\cdot k}} (1-\theta_{ij})^{C^k_{p-2}-w_{ij\cdot k}} \tag{2}$$

Both weighted and unweighted likelihood functions are based on the number of time one predictor dominating the other. Azen and Traxel (2009) provided an algebraic proof showing that if one measure of the R^2 produces a given dominance relationship, the other two measures will necessarily also produce the same direction of dominance. Hence, the three R^2 analogues will yield the same paired comparison results and the same likelihood function.

3.2 Link Functions

The two most commonly used link functions are the logit link, as assumed by the Bradley-Terry model, and the probit link, which assumes $\theta_{ij} = \Phi(d_i - d_j)$ with Φ being the cumulative function of the standard normal distribution. Wang et al. (2013) concludes that the results from the two link functions are consistent with each other. A more general approach is to assume that $\theta_{ij} = H(d_i - d_j)$, where H is a link function, which maps the difference in the dominance indices to a probability that lies between 0 and 1, inclusively. Also, the link function should have the following characteristics: (1) the larger the difference, the larger the value of θ_{ij}; that is, the more likely one predictor will dominate the other; (2) when two dominance indices are equal, θ_{ij} equals 0.5; that is, each predictor has a 50% chance of prevailing when dominance indices of two predictors are the same.

Since the paired comparisons of predictors are based upon the additional R^2, which is between 0 and 1. It would be ideal if the returning value of d_i is also within 0 and 1 so that d_i can be considered as an indicator of the additional R^2 associated with X_i in the population in the presence of other predictors. vanDorp and Kotz (2002) provide a four-parameter two-sided power (TSP) distribution with the cumulative distribution function that satisfies those characters. This link function is not only sufficiently rich from the mathematical perspective but also allows efficient implementations in practice. The four-parameter TSP distribution is described by parameters denoting the minimum value, a, the maximum value, b, the mode or most likely value, c, and a fourth parameter, η, describing the curvature of the distribution. The parameter η requires some evidence for the relative importance of the most likely value relative to distribution bounds a and b. That is, if $\eta = 2$, c is equally important as a and b; if $\eta > 2$, c has more weight than the bounds; if $\eta < 2$, c is given less emphasis relative to the bounds. The probability density function of $TSP(a,b,c,\eta)$ is demonstrated in equation (3).

$$f(x) = \begin{cases} \frac{\eta}{b-a}\left(\frac{x-a}{c-a}\right)^{\eta-1} & \text{if } a < x \le c \\ \frac{\eta}{b-a}\left(\frac{b-x}{b-c}\right)^{\eta-1} & \text{if } c < x \le b \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The expected value of a TSP distribution is

$$E(X) = \frac{a + (\eta - 1)c + b}{\eta + 1}$$

The cumulative distribution function follows from the expression as equation (4)

$$F(x) = \begin{cases} \frac{c-a}{b-a}\left(\frac{x-a}{c-a}\right)^{\eta} & \text{if } a < x \le c \\ 1 - \frac{b-c}{b-a}\left(\frac{b-x}{b-c}\right)^{\eta} & \text{if } c < x \le b \end{cases} \tag{4}$$

Wang (2015) stated that it is reasonable to choose $a = -1$ and $b = 1$ because the desired relative importance index is between 0 and 1. There is no preference regarding the possible value of d_i's. It is reasonable to use $c = (a+b)/2 = 0$ and $\eta = 2$. The probability density function of this particular TSP distribution has a triangular shape with expected value of 0. Wang (2015) compares the performance between the logit and the TSP link functions under both weighted and unweighted likelihood functions, and concludes that the TSP link function provides similar results as the logit link function with a more efficient computational procedure in a multiple regression setting. In this paper, we would like to ascertain if the same conclusion can be extended to a logistic regression setting.

4 Bayesian Dominance Inference

Algorithms used to produce maximum likelihood estimates of dominance abilities under the Bradley-Terry model fail to converge to finite values, and so cannot be used for many data sets with zero counts. Davidson and Solomon (1973) and Leonard (1977) describe Bayesian versions of the method of paired comparisons. The major benefit of the Bayesian approach is that prior information can be incorporated into the analysis so that the resulting estimates of dominance index are always finite (Leonard 1977).

4.1 Prior Distribution

The prior distribution plays a key role of Bayesian inference. It represents the information about an uncertain parameter and is combined with the probability distribution of observed data to yield the posterior distribution, which in turn, is used for future inferences and decisions. In practice, often there is no or very limited prior information about the dominance ability of predictors. Hierarchical (multilevel) prior and vague prior models are central to the modern Bayesian statistics for both conceptual and practical reasons. On the theoretical side, those prior models allow a more ”objective” approach rather than requiring them to be specified using subjective information (see James and Stein, 1960, Efron and Morris, 1975, and Morris, 1983). At a practical level, they are flexible tools for aggregating information and partial pooling of inferences (see, for example, Kreft and De Leeuw, 1998, Snijders and Bosker, 1999, Carlin and Louis, 2001, Raudenbush and Bryk, 2002, Gelman et al., 2003).

When the parameter varies from $-\infty$ to $+\infty$, the normal distribution is the most commonly used prior distribution because it corresponds to the prior belief that modest values of the parameter are nearly equally likely, and that very large values are somewhat less probable. For the logit link, the values of d_i's range from $-\infty$ to $+\infty$, and it is also reasonable to assume that there is the same symmetrical prior probability distribution for each predictor's dominance index. Therefore, without loss of generality, we assume that $d_i \sim N(\mu_i, \sigma^2)$.

The second stage of the hierarchical prior assume that another normal distribution is applied to the mean of the prior normal distribution, i.e. $\mu_i \sim N(\nu, \tau^2)$, and an Inverse-Gamma distribution is applied to the variance of the prior normal distribution; i.e., $\sigma^2 \sim IG(a,a)$. Here, ν, τ and a are predetermined hyper-parameters, which are chosen using vague information as follows. When the values of d_i vary between $-l$ and l, by applying the 3σ or empirical rule, the expected value of the standard deviation of d_i is around $l/3$, namely the expected value of σ^2 is about $l^2/9$. On the other hand, σ^2 is assumed to have an $IG(a,a)$, which has a mean value of $a/(a-1)$. By letting $a/(a-1) = l^2/9$ and solving for a, we can get the hyper-parameter $a = l^2/(l^2-9)$. When the values of d_i are in $(-4,4)$, the values of $d_i - d_j$ are in $(-8,8)$. As a result, the values of $\theta_{ij} = \exp(d_i-d_j)/(1+\exp(d_i-d_j))$ vary from 0.0003 to 0.9997, which is sufficiently wide for practical purposes. Therefore, it is adequate to presume that d_i's range between ± 4, and choose the value of a to be 2. In addition, it is also reasonable to assume the mean μ_i has a standard normal distribution with $\nu = 0$ and $\tau = 1$.

For the TSP link function, we assume that d_i's range from 0 to 1, and we do not have any preference towards any particular value. It is reasonable to apply the $(0,1)$ Uniform distribution as the prior distribution to each d_i. Therefore, the TSP link function bears a simpler prior distribution.

4.2 Bayesian Computation

Under the square error loss function, the posterior mean is the Bayesian estimator. However, the closed form expressions of the posterior distribution of d_i's are not easy to achieve. Therefore, we use Markov chain Monte Carlo (MCMC) methods to obtain numerical results. In order to implement the MCMC procedure, it is necessary to have the full conditional posterior distributions, which are the conditional distributions of one parameter given all the other unknown parameters and data. In this paper, let $\mathbf{d} = (d_1,\cdots,d_p)$, $\mathbf{d}[-j] = (d_1,\cdots,d_{j-1},d_{j+1},\cdots,d_p)$, $\boldsymbol{\mu} = (\mu_1,\cdots,\mu_p)$, $\boldsymbol{\mu}[-j] = (\mu_1,\cdots,\mu_{j-1},\mu_{j+1},\cdots,\mu_p)$, $\phi(\cdot|\mu,\sigma^2)$ be the probability density function of $N(\mu,\sigma^2)$, $H_1(\cdot)$ be the logit link function, and $H_2(\cdot)$ be the TSP link function. Also, let $(\cdot|\ )$ denote the conditional distribution and $[\cdot|\ ]$ denote the conditional density function.

Proposition 4.2.1. The full conditional posterior distributions of $(d_1,\cdots d_p,\mu_1,\cdots,\mu_p,\sigma^2)$ given data $\mathbf{X}$ under the model with unweighted likelihood and the logit link function, Model I, are as follows.

–For $i = 1,\cdots,p$,

$$[d_i \mid \mathbf{d}_{[-i]},\boldsymbol{\mu},\sigma^2,\mathbf{X}] \propto \prod_{i<j} C_{2^{p-2}}^{w_{ij}} H_1(d_i-d_j)^{w_{ij}}(1-H_1(d_i-d_j))^{2^{p-2}-w_{ij}}\phi(d_i|\mu_i,\sigma^2)$$

–For $i = 1,\cdots,p$, $(\mu_i \mid \mathbf{d},\boldsymbol{\mu}[-i],\sigma^2,\mathbf{X}) \sim N\left(\frac{d_i\tau^2+\sigma^2\nu}{\tau^2+\sigma^2},\ \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}\right)$

–$(\sigma^2 \mid \mathbf{d},\boldsymbol{\mu},\mathbf{X}) \sim IG\left(\frac{p}{2}+a,\ \sum_{i=1}^{k}\frac{(d_i-\mu_i)^2}{2}+a\right)$,

Proposition 4.2.2. The full conditional posterior distributions of $(d_1,\cdots d_p,\mu_1,\cdots,\mu_p,\sigma^2)$ given data $\mathbf{X}$, under the model with weighted likelihood and the logit link function, Model II, are as follows.

–For $i = 1,\cdots,p$,

$$[d_i \mid \mathbf{d}_{[-i]},\boldsymbol{\mu},\sigma^2,\mathbf{X}] \propto \sum_{k=0}^{p-2}\frac{C_{p-2}^k}{2^{p-2}}\prod_{i\neq j}\binom{C_{p-2}^k}{w_{ij\cdot k}} H_1(d_i-d_j)^{w_{ij.k}}(1-H_1(d_i-d_j))^{C_{p-2}^k-w_{ij.k}}\phi(d_i|\mu_i,\sigma^2)$$

–For $i = 1,\cdots,p$, $(\mu_i \mid \mathbf{d},\boldsymbol{\mu}[-i],\sigma^2,\mathbf{X}) \sim N\left(\frac{d_i\tau^2+\sigma^2\nu}{\tau^2+\sigma^2},\ \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}\right)$

–$(\sigma^2 \mid \mathbf{d},\boldsymbol{\mu},\mathbf{X}) \sim IG\left(\frac{p}{2}+a,\ \sum_{i=1}^{k}\frac{(d_i-\mu_i)^2}{2}+a\right)$,

Proposition 4.2.3. The full conditional posterior distributions of $(d_1, \cdots d_p,)$ given data $\mathbf{X}$ under the model with unweighted likelihood and the TSP link function, Model III, are as follows.

$$[d_i \mid \mathbf{d}_{[-i]}, \boldsymbol{\mu}, \sigma^2, \mathbf{X}] \propto \prod_{i<j} C_{2^{p-2}}^{w_{ij}} H_2(d_i - d_j)^{w_{ij}} (1 - H_2(d_i - d_j))^{2^{p-2} - w_{ij}},$$

for $i = 1, \cdots, p$.

Proposition 4.2.4. The full conditional posterior distributions of $(d_1, \cdots d_p)$ given data $\mathbf{X}$ under the model with weighted likelihood and the TSP link function, Model IV, are as follows.

$$[d_i \mid \mathbf{d}_{[-i]}, \boldsymbol{\mu}, \sigma^2, \mathbf{X}] \propto \sum_{k=0}^{p-2} \frac{C_{p-2}^{k}}{2^{p-2}} \prod_{i \neq j} \binom{C_{p-2}^{k}}{w_{ij \cdot k}} H_2(d_i - d_j)^{w_{ij.k}} (1 - H_2(d_i - d_j))^{C_{p-2}^{k} - w_{ij.k}},$$

for $i = 1, \cdots, p$.

Comparing the full conditional distributions, one can observe that the TSP link function simplifies the computational procedure with simpler full conditional posterior distributions.

The Bayesian dominance method provides prosperous information about the relative importance of predictors via posterior distributions of dominance indices d_i's. First of all, the posterior mean of the dominance index d_i can be applied as an overall measure of relative importance/dominance ability of the predictor X_i. Secondly, the posterior distribution of dominance probability θ_{ij} can be obtained by applying the corresponding link function to the MCMC chain of d_i's, and the posterior mean of θ_{ij} can be applied as an estimate of the dominance probability to reveal the conclusion of paired comparisons among predictors in a probabilistic manner. Moreover, the lower and upper 2.5^{th} percentiles of the posterior distributions of d_i can be used to construct 95% confidence intervals of d_i. Last but not least, the posterior probability of one particular order of dominance can be estimated by the proportion of steps occurred in the MCMC chain that is in the same order. Hence, we can obtain the most likely order of dominance directly from the MCMC chain.

5 Simulations

Simulation studies are based on a generated population of size $(N = 500,000)$ with correlation matrices as displayed in Table **??**. First, the correlation matrix is used to solve for a vector of the log odds of the dependent variable for logistic regression (Y vlaues) and the matrix containing predictor values from the multivariate normal population. Second, a dichotomous response variable (Y* values) is generated to be used in the logistic regression analysis by rearranging the logistic transformation $\pi(y) = \exp(y)/(1 + \exp(y))$, which is then compared to random values drawn from a uniform distribution (with a range of 0 to 1), such that if the random value is less than $\pi(y)$ then $y* = 1$ and otherwise $y* = 0$. We believe that a random sample of size $500,000$ is large enough to stand for a population. The general Dominance Analysis indices and the Bayesian dominance indices for all four models, as listed in Table 3, indicate that the population order of dominance is $X_2 > X_3 > X_4 > X_1$.

To better evaluate the performance of the proposed Bayesian approaches in logistic regression settings, one thousand simple random samples of size $n = 500$ are selected from the previously generated population, and the estimated dominance indices based on different models are computed for each sample. The average and standard deviation of the 1,000 estimated dominance index values are calculated for each model. Because the magnitude of population dominance indices from different approaches are quite different, relative bias and relative standard deviation are proposed to better measure the accuracy and reliability of the estimates, as are presented in Table 4. In probability theory and statistics, the relative bias is a commonly used measure of accuracy, and is defined as the ratio of the bias to the population value, where the bias is defined as the the difference between the average of the estimates and the population value. The coefficient of variation (CV) is a standardized measure of dispersion, which is defined as the ratio of the standard deviation to the population value. The absolute value of the CV is sometimes known as relative standard deviation, which is expressed as a percentage.

The Dominance Analysis methods produces the estimated dominance indices with the largest relative biases as far as 2622.34% and the largest relative standard deviation of 3717.91%. Estimates using logit link functions bear less relative biases and relative standard deviations than those from the Dominance Analysis methods but greater than those from the TSP link functions, with the exception of X_4. In the Model III, the dominance index of X_4 has a relative bias 123.96% and relative standard deviation 138.16%, which are greater than those from the other three Bayesian models but are still less than that from the Dominance Analysis. Overall, the Model IV produces the smallest relative biases and relative standard deviation. Thus, we conclude that it is the most accurate and reliable model to determine the relative importance of predictors.

6 Example

Dhillon et. al. (1987) provides an interesting data to explain the choice by home buyers of fixed versus adjustable rate mortgages. They use 78 observations from a bank in Baton Rouge, Louisiana, taken over the period January, 1983 to February 1984. There are 6 explanatory variables used to predict whether an adjustable rate mortgage was chosen $Y = 1$:

– X_1: FIXRATE: fixed interest rate in whole numbers;
– X_2: MARGIN: the variable rate minus the fixed rate;
– X_3: YIELD: The 10-year Treasury rate less the 1-year rate;
– X_4: MATURITY: ratio of maturities on adjustable to fixed rates;
– X_5: POINTS: ratio of points paid or an adjustable mortgage to those paid on a fixed rate mortgage;
– X_6: NETWORTH: borrower's net worth

The sample correlation matrix of the variables is displayed in Table **??**. There exist minor correlation amongst the explanatory variables such as the sample correlation between X_1 and X_2 is about 0.468. Paired comparisons of predictors based on the additional increase in R^2 are conducted among $2^{6-2} = 16$ possible subset regression models. The total numbers of times that the variable X_i prevails over X_j in all possible subset reference models are presented in Table 6. Predictors X_1, X_2 and X_6 completely dominate X_4 and X_5; X_3 prevails over X_5 in all subset models except when $k = 1$, and prevails X_4 most of the time (12 out of 16); X_4 prevails over X_5 most of the time (13 out of 16 times); X_1 prevails over X_2 and X_3 most of the time (12 out of 16 times); X_2 prevails over X_3 most of the time (13 out of 16 times); and X_6 prevails over X_1, X_2, X_3 a majority of the time (10, 12, 9 out of 16 times, respectively). In summary, the observed order of dominance from paired comparison is $X_6 > X_1 > X_2 > X_3 > X_4 > X_5$. The results of Dominance Analysis based on the general dominance indices show that the hierarchical order of dominance is $X_1 > X_6 > X_2 > X_3 > X_4 > X_5$, see Table 7. The dominance indices of X_1 and X_6 are very close to each other. The Bayesian point estimates and 95% confidence intervals of dominance indices d_i's under four different models are listed in Table 7. The resulting order of dominance indices from both un-weighted models is $X_1 > X_6 > X_2 > X_3 > X_4 > X_5$, which is the same as the results from the Dominance Analysis, and the order from both weighted models is $X_2 > X_6 > X_3 > X_1 > X_4 > X_5$.

All models put X_4 (Maturity) and X_5 (Points) at the bottom of the hierarchy when predicting whether an adjustable rate mortgage will be used or not. All models agree that X_2 (Margin) is more important than X_3 (Yield); and X_6 (Net Worth) is more important than X_3 (Yield). There exists discrepancy regarding the order of dominance among X_2, X_6, and X_1. From theoretical point of view, there exists Simpson's paradox in the paired comparison results in Table 6. The Simpson's's paradox often occurs when subgroups are combined together, and the data are examined in aggregate form, the conclusion made from the subgroups may reverse itself when using the combined group. For example, the total paired comparison table shows that X_1 prevails X_2 12 out of 16 times (75%), which implies that X_1 dominates X_2. However, when investigating the subgroups, we notice that 5 of the 12 "winning" occur when $k = 1$ where X_1 and X_2 are compared total of 6 times. In the subgroups, where $k = 4$, X_2 prevails over X_1 (100%) but they are only compared once. Wang and Yao (2014) states that the results from the unweighted model is less accurate than the weighted mode because the former one does not take into account the possible number of times that predictors encountering under different size of baseline reference models. From practical point of view, when two subjects are compared, generally speaking, the difference between the two subjects would carry more weights than the actual value of the subjects. Also when examine the data, we found that there is very less variation in X_1 with an average of 13.25%. Therefore, the margin between variable rate and fixed rate (X_2) would matter more when one determines if an adjustable rate will be used or not, and the actual fixed interest is a less important factor. As a summary, we conclude that, for the given data set, the order of dominance $X_2 > X_6 > X_3 > X_1 > X_4 > X_5$ is more reasonable from both practical and theoretical perspective.

When using the lower and upper 2.5^{th} percentiles of the MCMC chain, the 95% confidence intervals of dominance indices are obtained and are shown in Table 7. Because the 95% confidence intervals of dominance indices overlap each other, it is challenging to determine which predictor definitely dominates the other predictors. Fortunately, the Bayesian approach provides a probabilistic solution to this difficulty. By applying the link functions to the dominance indices within the MCMC, we obtain the estimate of dominance probabilities, $\theta_{i,j}$'s, which are listed in Table 8, with $i < j$. For $i > j$, $\theta_{i,j} = 1 - \theta_{j,i}$. The results are consistent with the order of dominance indices in Table 7, correspondingly and respectively. For example, both weighted models have $\theta_{2,3}$, $\theta_{2,4}$, $\theta_{2,5}$, $\theta_{2,6}$ greater than 0.5 and $\theta_{1,2}$ less than 0.5, which implies that predictor X_2 has more chance to dominate the other predictors. Arranging the dominance probabilities from smallest to largest, one can obtain $\theta_{2,6} < \theta_{2,3} < \theta_{2,1} < \theta_{2,4} < \theta_{2,5}$, which indicate the order of dominance is $X_2 > X_6 > X_3 > X_1 > X_4 > X_5$.

The posterior probability of a particular order of the dominance ranking is estimated by the proportion of this order occurring in the MCMC steps. Among the 720 possible of ranking orders, the order ranking $X_1 > X_6 > X_2 > X_3 > X_4 > X_5$ receives the highest probability under both un-weighted models, and the order ranking $X_2 > X_6 > X_3 > X_1 > X_4 > X_5$ receives the highest probability under both weighted models; Once again, the results are consistent with previous findings using dominance indexes and probabilities within each model.

It is worth mentioning that the MCMC chain produced by the TSP link function converges faster and is less correlated than those by the logit link function. The trace plots of the MCMC chains of the four models, Figure 2, Figure 3, Figure 4 and Figure 5, show that the MCMC chains produced by the TSP link function mixed better and converge faster than those by the logit link function. Moreover, autocorrelation functions of the four models, as presenting in Figure 6, Figure 7, Figure 8 and Figure 9, show that the MCMC chains produced by the TSP link function die down faster than those from the logit link function, which means the former ones are less correlated and more stationary than the latter ones. Thus we conclude that the TSP link provides more reliable results than the logit link function.

As a summary, the most probable order of dominance is consistent among the results from the estimated dominant index, dominant probability within each case, respectively. Although there exist discrepancies between weighted and unweighted models, we believe that the order $X_2 > X_6 > X_3 > X_1 > X_4 > X_5$ is the most preferable from both practical and theoretical perspective.

7 Summary and Concluding Remarks

Relative importance analyses permit a greater understanding of the particular role played by variables in a logistic regression analysis. Crucially, these analyses can reveal the underlying impact of a particular predictor more accurately than standardized regression coefficients or simple correlations. This paper extends the current Baysian approach of measuring the relative importance of predictors in linear regression models (Wang 2015) to a logistic regression scenario by applying the generalized R^2 measures.

The advantage of the Bayesian approach is that it allows the use of genuine prior information in addition to the information that is available in the observed data to produce better results. In general, Bayesian methods provide a better approximation to the level of uncertainty than other approaches which use only information provided by the model and the observed data. In addition to providing useful statistics, such as, the mean and percentiles of the posterior distribution of the unknown parameters, Bayesian methods give more reliable results for small samples (Dunson 2000; Lee and Song 2004; Scheines, Hoijtink and Boomsma 1999).

The Bayesian approach offers several advantages over the current methods in determining the relative importance of predictors in a linear regression model. First, this probabilistic model based approach provides more comprehensive inference about the population relative dominance ability of predictors than the current Dominance Analysis. Secondly, the Bayesian approach provides more information about the relative importance of the predictor by making straightforward statements about the dominance ability of the predictors, the dominance probability of each possible pair of predictors, and the probability of each possible order of dominance. Thirdly, the main advantage of the TSP link function over the logit link function is that the TSP link function is more flexible in modelling the results of paired comparisons based on indices, which are between 0 and 1. As a result, the TSP link provides more useful and meaningful estimation of the dominance index between 0 and 1, which can be treated as a measure of association between the predictor and the response variable in the presence of other predictors. More to point that the TSP link function simplifies the Bayesian computational process. Last but not least, the Bayesian method is not limited to a simple model, but offers a rich potential to incorporate more complex models. Both simulation studies and empirical example support that these benefits extend very well into the logistic regression setting. The model with weighted likelihood function and the TSP link function provides the most accurate estimate with the smallest relative bias and standard deviation.

One of the central questions in a multivariate analysis of variance (MANOVA) considers identifying the dependent variables that are driving the significant multivariate F-test. Unfortunately, the correlations among the various dependent variables often make it difficult to accurately identify the role being played by the various dependent variables. Although both Dominance Analysis and Bayesian approach are developed for use with OLS regression, Azen et.al. (2006) presented modifications of these respective analysis to the multivariate regression models. Thus, we will continue our work about questions of the relative contribution of each of the variables in terms of predicting the multivariate predictors, and how the proposed method can be examined in this context as well.

Table 1: Summary of Properties of R^2 Analogues for Logistic Regression Model

Boundedness	Invariance	Monotonicity	Interpretability	
R_M^2	Yes	Yes	Yes	Yes
R_N^2	Yes	Yes	Yes	No
R_E^2	Yes	Yes	Yes	Yes

Table 2: Population Correlation Matrix for Simulation

Simulation I					
	Y	X_1	X_2	X_3	X_4
Y	1	0.06	-0.42	0.19	-0.02
X_1	0.06	1	0.01	-0.01	0.07
X_2	-0.42	0.01	1	-0.13	0.11
X_3	0.19	-0.01	-0.13	1	0.01
X_4	-0.02	0.07	0.11	0.01	1

Table 3: Population Dominance Indices of Predictors

	Model I	Model II	Model III	Model IV	DA
X_1	-1.00424	-0.45630	0.34411	0.41491	0.00045
X_2	3.08734	1.63116	0.87559	0.73775	0.02095
X_3	0.91236	0.71019	0.66948	0.59835	0.00322
X_4	-3.06442	-1.49385	0.12919	0.26067	0.00006

Table 4: Relative Biases and Standard Deviations of Dominance Indices Estimates

	Model I	Model II	Model III	Model IV	DA
X_1	40.98% (155.70%)	69.46% (184.31%)	-9.51% (56.56%)	-7.16% (30.23%)	327.54% (612.71%)
X_2	-3.60% (22.86%)	0.41% (21.54%)	-1.72% (5.57%)	0.15% (4.93%)	6.00% (52.48%)
X_3	-103.42% (162.52%)	-103.98% (118.11%)	-25.93% (31.47%)	-16.26% (21.22%)	43.68% (138.06%)
X_4	-49.59% (47.29%)	-43.33% (52.32%)	123.96% (138.16%)	42.72% (44.13%)	2622.34% (3717.91%)

Table 5: Sample Correlation Matrix in Example

	Y	X_1	X_2	X_3	X_4	X_5	X_6
Y	1.000	−0.213	−0.069	0.076	0.265	0.381	0.399
X_1	−0.213	1.000	−0.468	0.277	−0.271	0.093	−0.236
X_2	−0.069	−0.468	1.000	−0.405	−0.011	−0.254	−0.176
X_3	0.076	0.277	−0.405	1.000	0.005	0.285	−0.011
X_4	0.265	−0.271	−0.011	0.005	1.000	0.022	0.135
X_5	0.381	0.093	−0.254	0.285	0.022	1.000	0.365
X_6	0.399	−0.236	−0.176	−0.011	0.135	0.365	1.000

Table 6: Paired Comparisons of Predictors in the Example

Null k=0						
	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0	1	1	1	1	1
X_2	0	0	1	1	1	0
X_3	0	0	0	1	1	0
X_4	0	0	0	0	0	0
X_5	0	0	0	1	0	0
X_6	0	1	1	1	1	0

k=1						
	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0	4	4	4	4	3
X_2	0	0	4	4	4	1
X_3	0	0	0	3	3	0
X_4	0	0	1	0	2	0
X_5	0	0	1	2	0	0
X_6	1	3	4	4	4	0

k=2						
	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0	5	5	6	6	2
X_2	1	0	4	6	6	2
X_3	1	2	0	4	6	1
X_4	0	0	2	0	6	0
X_5	0	0	0	0	0	0
X_6	4	4	5	6	6	0

k=3						
	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0	2	2	4	4	1
X_2	2	0	3	4	4	2
X_3	2	1	0	3	4	2
X_4	0	0	1	0	4	0
X_5	0	0	0	0	0	0
X_6	3	2	2	4	4	0

k=4						
	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0	0	0	1	1	0
X_2	1	0	1	1	1	1
X_3	1	0	0	1	1	1
X_4	0	0	0	0	1	0
X_5	0	0	0	0	0	0
X_6	1	0	0	1	1	0

Total						
	X_1	X_2	X_3	X_4	X_5	X_6
X_1	0	12	12	16	16	7
X_2	4	0	13	16	16	6
X_3	4	3	0	12	15	4
X_4	0	0	4	0	13	0
X_5	0	0	1	3	0	0
X_6	9	10	12	16	16	0

Table 7: Summary of Posterior Statistics of Dominance Index of Predictors in Example

	Model I	Model II	Model III	Model IV	DA
X_1	1.3160 (-0.0170 , 2.9450)	1.0182 (-2.0485 , 5.0925)	0.9307 (0.8136 , 0.9966)	0.5705 (0.0683 , 0.9800)	0.0784
X_2	0.8112 (-0.6637 , 2.3337)	2.2989 (-0.5725 , 6.0324)	0.8380 (0.6793 , 0.9749)	0.7041 (0.1832 , 0.9861)	0.0540
X_3	-0.3221 (-1.7305 , 1.1992)	1.3224 (-1.5519 , 4.5752)	0.5833 (0.4290 , 0.7345)	0.6031 (0.0986 , 0.9721)	0.0384
X_4	-1.9394 (-3.3713 , -0.4106)	-0.9500 (-3.9111 , 1.5068)	0.1996 (0.0449 , 0.3683)	0.2377 (0.0135 , 0.6506)	0.0091
X_5	-3.2350 (-4.8809 , -1.6467)	-1.7622 (-5.2753 , 0.9015)	0.0473 (0.0014 , 0.1555)	0.2230 (0.0116 , 0.6173)	0.0051
X_6	1.2831 (-0.0961 , 2.9226)	1.4435 (-0.6937 , 4.5667)	0.9252 (0.7877 , 0.9939)	0.6190 (0.1356 , 0.9690)	0.0758

Table 8: Summary of Posterior Statistics of Dominance Probability of Predictors in Example

	Model I	Model II	Model III	Model IV
$\theta_{1,2}$	0.6201	0.3387	0.5856	0.4074
$\theta_{1,3}$	0.8307	0.4324	0.7839	0.4742
$\theta_{1,4}$	0.9583	0.7280	0.9601	0.7254
$\theta_{1,5}$	0.9874	0.8584	0.9912	0.7495
$\theta_{1,6}$	0.5079	0.4272	0.5052	0.4633
$\theta_{2,3}$	0.7499	0.6638	0.7183	0.5791
$\theta_{2,4}$	0.9337	0.9108	0.9304	0.8270
$\theta_{2,5}$	0.9796	0.9223	0.9748	0.8284
$\theta_{2,6}$	0.3879	0.6276	0.4194	0.5639
$\theta_{3,4}$	0.8253	0.8318	0.8063	0.7666
$\theta_{3,5}$	0.9414	0.8677	0.8889	0.7669
$\theta_{3,6}$	0.1742	0.5008	0.2200	0.4903
$\theta_{4,5}$	0.7741	0.6276	0.6371	0.5118
$\theta_{4,6}$	0.0432	0.1763	0.0414	0.2301
$\theta_{5,6}$	0.0131	0.0940	0.0096	0.2137

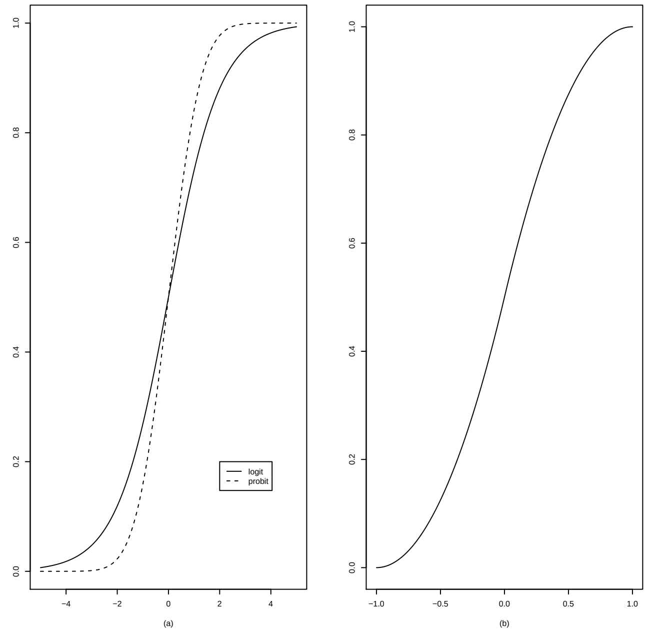


Fig. 1: Link Functions

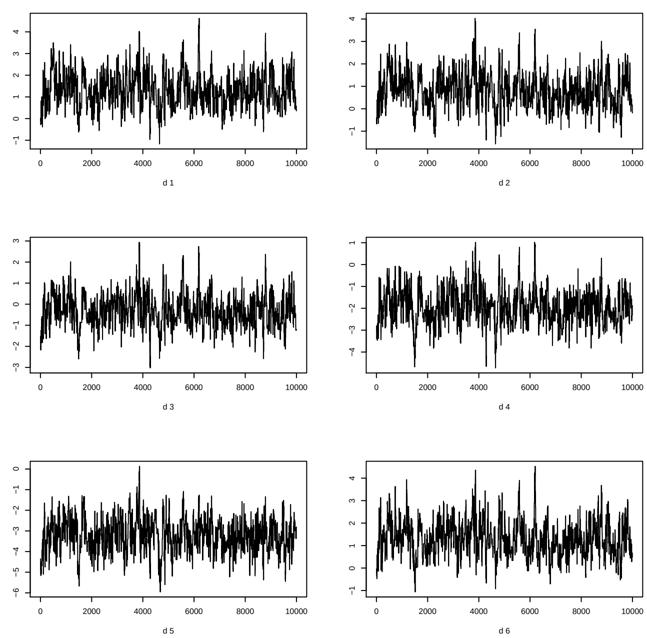


Fig. 2: MCMC Trace Plots of Model I

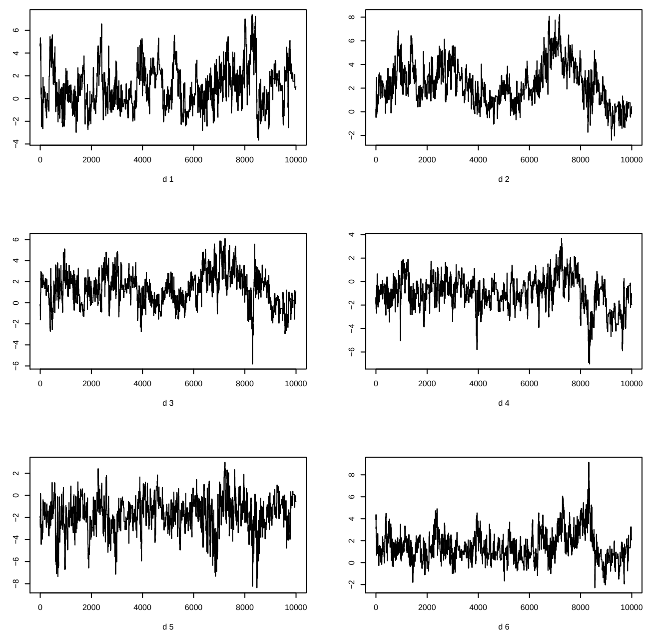


Fig. 3: MCMC Trace Plots of Model II

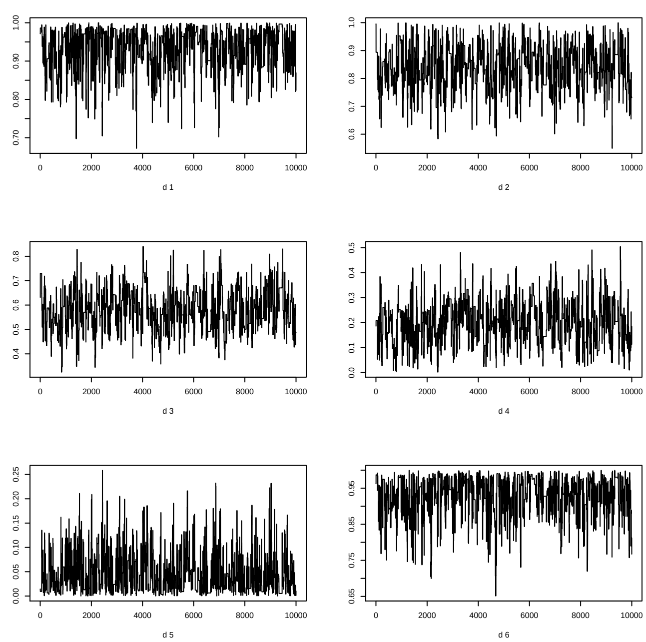


Fig. 4: MCMC Trace Plots of Model III

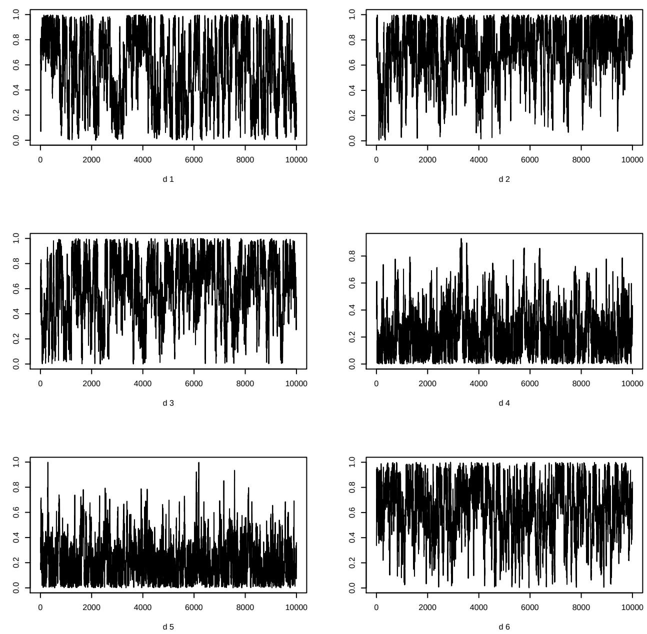


Fig. 5: MCMC Trace Plots of Model IV

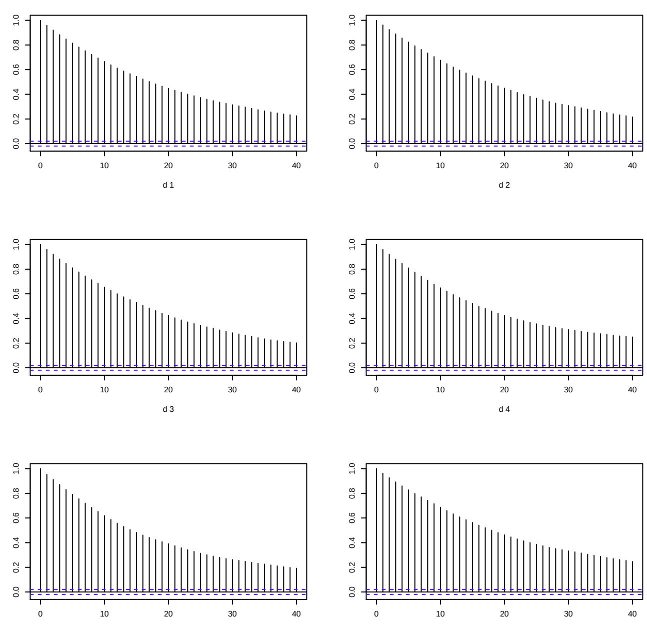


Fig. 6: ACF Plots of Model I

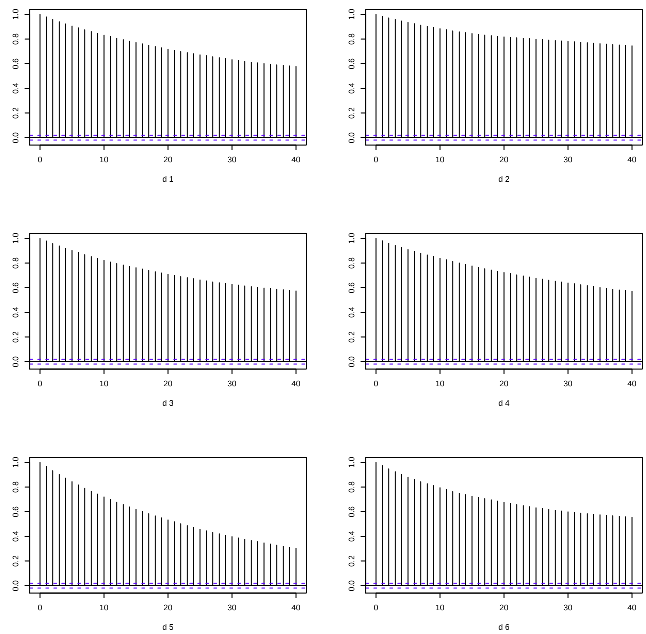


Fig. 7: ACF Plots of Model II

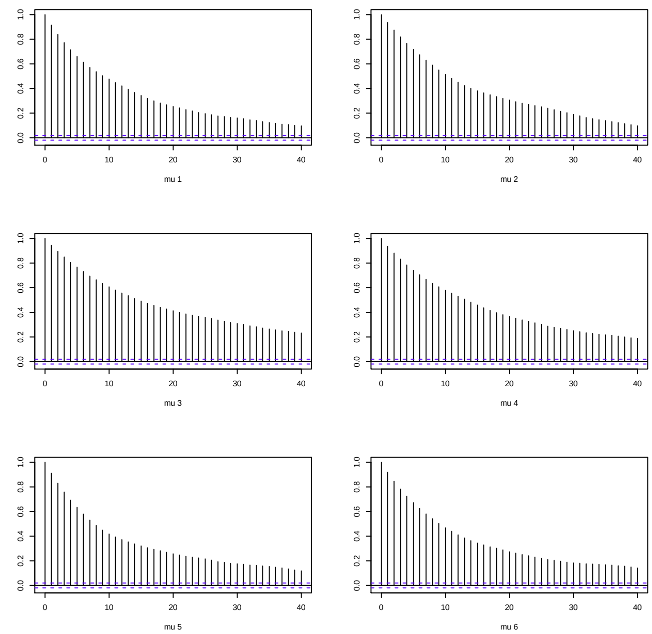


Fig. 8: ACF Plots of Model III

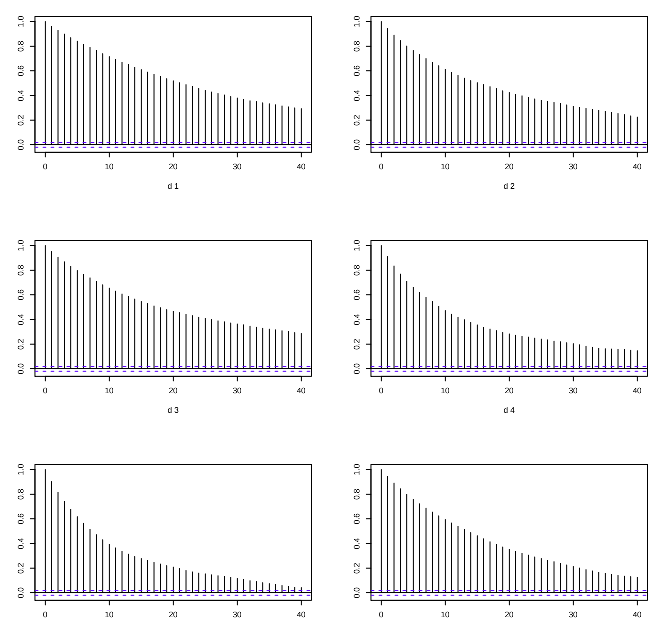


Fig. 9: ACF Plots of Model IV

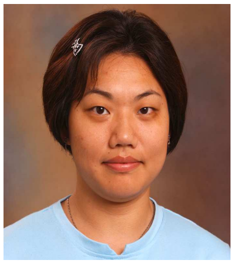
Acknowledgement

I would like to thank Mr. Dan Korpon for helpful comments on the manuscript.
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Amemiya, T. (1981). Qualitative response models: A survey. Journal of Economic Literature, 19, 14831536.
[2] Azen, Razia, and David V. Budescu. Comparing predictors in multivariate regression models: An extension of dominance analysis. Journal of Educational and Behavioral Statistics 31.2 (2006): 157-180.
[3] Azen, R., and Traxel, N. M. (2009). Using dominance analysis to determine predictor importance in logistic regression. Journal of Educational and Behavioral Statistics, 34, 319-347
[4] Bradley, R.A.; and Terry, M.E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika 39(3/4):324-34.
[5] Budescu, David V. (1993), "Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression, Psychological Bulletin, 114, 542-551.
[6] Carlin, B. P., and Louis, T. A. (2001). Bayes and Empirical Bayes Methods for Data Analysis, second edition. London: Chapman and Hall.
[7] Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69, 161-182.
[8] David, H. A. (1988) The Method of Paired Comparisons, 2nd edn. London: Chapman and Hall.
[9] Davidson, R.R.; and Solomon, D.L. (1973). A Bayesian Approach to Paired Comparison Experimentation. Biometrika 60(3):477-487.
[10] Dhillon,; and Shilling,; and Sirmans,. (1987). Choosing Between Fixed and Adjustable Rate Mortgages. Journal of Money, Credit and Banking 19(1):260-267
[11] Dunson, D.B. (2000). Bayesian Latent Variable Models for Clustered Mixed Outcomes. Journal of the Royal Statistical Society. Series B (Statistical Methodology). 62(2): 355-366.
[12] Efron, Bradley, and Carl Morris. Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association 70.350 (1975): 311-319.
[13] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). Bayesian Data Analysis, second edition. London: Chapman and Hall.
[14] James, W., and Stein, C. (1960). Estimation with quadratic loss. In Proceedings of the Fourth Berkeley Symposium 1, ed. J. Neyman, 361-380. Berkeley: University of California Press.
[15] Johnson, J.W. and LeBreton, J.M. (2004). History and Use of Relative Importance Indices in Organizational Research. Organizational Research Methods, 7 (3): 238-257.
[16] Kreft, I., and De Leeuw, J. (1998). Introducing Multilevel Modeling. London: Sage.
[17] Kvalseth, T. O. (1985). Cautionary notes about R2. The American Statistician, 39, 279285.
[18] Lahiri, P. (2003). On the Impact of Bootstrap in Survey Sampling and Small-Area Estimation. Statistical Science, 18(2): 199-210.
[19] Lee, S.Y. and Song X.Y. (2004). Bayesian model comparison of nonlinear structural equation models with missing continuous and ordinal categorical data. British Journal of Mathematical and Statistical Psychology, 57: 131-150.
[20] Leonard, T. (1977). An Alternative Bayesian Approach to the Bradley-Terry Model for Paired Comparisons. Biometrics 33(1):121-132.
[21] Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. The American Statistician, 54, 1724.
[22] Mittlbock, M., and Schemper, M. (1996). Explained variation for logistic regression. Statistics in Medicine, 15, 19871997.
[23] Morris, C.N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. Journal of the American Statistical Association, 78(381): 47-55.
[24] Raudenbush, S. W., and Bryk, A. S. (2002). Hierarchical Linear Models, second edition. Thousand Oaks, Calif.: Sage.
[25] Rossi, P.E. and Allenby, G.M. (2003). Bayesian Statistics and Marketing. Marketing Science, 22(3): 304-328.
[26] Scheines, R.; Hoijtink, H.; and Boomsma, A. (1999). Bayesian Estimation and Testing of Structural Equation Models. Psychometrika, 64(1): 37-52.
[27] Snijders, T. A. B., and Bosker, R. J. (1999). Multilevel Analysis. London: Sage.
[28] Van den Burg, W., and Lewis, C. (1988). Some properties of two measures of multivariate association. Psychometrika, 53, 109122.
[29] Van Dorp, J. R., and Kotz, S. (2002). The standard two-sided power distribution and its properties: with applications in financial engineering. The American Statistician, 56(2), 90-99.
[30] Wang, X.; Duverger, P.; and Bansal, H. S. (2013). Bayesian Inference of Predictors Relative Importance in Linear Regression Model Using Dominance Hierachies. International Journal of Pure and Applied Mathematics, 88(3), 321-339.
[31] Wang, X., and Yao, D. (2014). Bayesian Inference of Predictors Relative Importance in Multiple Regression Model. Advances and Applications in Statistics, 41(1), 69.
[32] Wang, X. (2015). An Application of Two Sided Power Distribution in Bayesian Analysis of Paired Comparison of Relative Importance of Predictors in Linear Regression Models. International Journal of Advanced Statistics and Probability (Accepted)

[33] Zheng, B., and Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. Statistics in Medicine, 19, 17711781.



Xiaoyin Wang is Associate Professor of Mathematics at Towson Universit (USA). She received the PhD degree in Statistics at University of Missouri-Columbia. Her main research interests are: Applied Statistics, Bayesian Statistics, Capture-Recapture Model, Time Series Analysis, and Regression Analysis. She has published many research articles in reputed international journals In mathematics and statistics area.