ABSTRACT

Title of Dissertation: THE IMPACT OF MODEL SELECTION ON LOGLINEAR ANALYSIS OF CONTINGENCY TABLES

Jing Gao, Doctor of Philosophy, 2009

Dissertation directed by: Professor C. Mitchell Dayton
Department of Measurement, Statistics and Evaluation

It is common practice for researchers in the social sciences and education to use model selection techniques to search for best fitting models and to carry out inference as if these models were given *a priori*. This study examined the effect of model selection on inference in the framework of loglinear modeling. The purposes were to (i) examine the consequences when the behavior of model selection is ignored; and (ii) investigate the performance of the estimator provided by the Bayesian model averaging method and evaluate the usefulness of the multi-model inference as opposed to the single model inference.

The basic finding of this study was that inference based on a single "best fit" model chosen from a set of candidate models leads to underestimation of the sampling variability of the parameters estimates and induces additional bias in the estimates. The results of the simulation study showed that due to model uncertainty the post-model-selection parameter estimator has larger bias, standard error, and mean square error than the estimator under the true model assumption. The same results applied to the conditional odds ratio estimators. The primary reason for these results is that the sampling distribution of the post-model-selection estimator is, in actuality, a mixture of distributions from a set of candidate models. Thus, the variability of the post-model-

selection estimator has a large component from selection bias. While these problems were alleviated with the increase of sample size, the interpretation of the p-value of the Z-statistic of the parameters was misleading even when sample size was quite large. To avoid the problem of inference based on a single best model, Bayesian model averaging adopts a multi-model inference method, treating the weighted mean of the estimates from each model in the set as a point estimator, where the weights are derived using Bayes' theorem. Generally speaking, the simulation results confirmed the efficacy of the BMA method as compared with data-driven single "best-fit" model inference.

THE IMPACT OF MODEL SELECTION ON LOGLINEAR ANALYSIS OF CONTINGENCY TABLES

By

Jing Gao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:

Professor C. Mitchell Dayton, Chair
Professor George B. Macready
Professor Gregory R. Hancock
Professor Jeffrey R. Harring
Professor Partha Lahiri

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee members, Dr. Hancock, Dr. Macready, Dr. Harring, and Dr. Lahiri for providing me with constructive suggestions on this paper. I especially would like to thank Dr. Dayton for his guidance during my graduate career.

I also would like to thank my family for their support. To my husband, Yong, thank you for loving me and supporting me. To my daughter, Gabrielle, your smile is the biggest reward for me during this journey. To my parents, thank you for your unconditional love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1 Introduction to the Dilemma of Model Selection

Model selection is the process of choosing an appropriate model from a set of candidate models, given the data. Model selection is thought to be "important and unavoidable." (Raftery, 1995). In practice, it is not possible to specify the correct model for relations among variables in a data set. In the absence of knowledge about the true model, it is natural for investigators to fit several candidate models to the data and choose the one with, in some sense, the best fit. If one of the research questions is to find out the most convincing theory among several competing ones, which are represented by different statistical models, then model selection is an essential part of the research.

When the model selection procedure is applied, one question is asked implicitly: among the set of models, which one most likely to have produced the data. Various model selection techniques have been proposed to address this question. Once a model is selected, the model is used to estimate model parameters as if the model were the true model. In other words, the data are analyzed as if they were generated by the selected model. A major criticism of this approach is that the model is selected based on a data-driven criterion and the same data are used to select the model and estimate the parameters (Hurvich & Tsai, 1990). When one first chooses a best model and then bases the inferences on that model, one ignores the uncertainty involved in the choice of this best model which may have a large effect on estimates of parameters of interest. This problem is well known and has been widely discussed in the research literature (Chatfield, 1995).

The reasoning underlying the modeling process in conventional inferential statistics is shown in Figure 1. First, we impose a model on a specified population and assume that the population data are generated by this model. Then an observed sample drawn (in theory) from the population is used to estimate the parameters of the model. An assumption in using the sample data to estimate the parameters of the model is that the model is given *a priori*. Of course, the a priori model may be incorrect; that is, misspecified. If the model is selected based on the data, then the assumption is violated and the parameter estimation is dependent on the outcome of a model selection process. Therefore, the estimates may be biased, the standard errors may be biased, the coverage of the confidence interval may not be at nominal levels, and inferences derived from these estimators may be misleading. Hurvich and Tsai (1990) pointed out "Conditionally on the event of having selected a particular model, the distribution of the data may be substantially different from their unconditional distribution." Zhang (1992) commented on this process as "logically unsound and practically misleading."

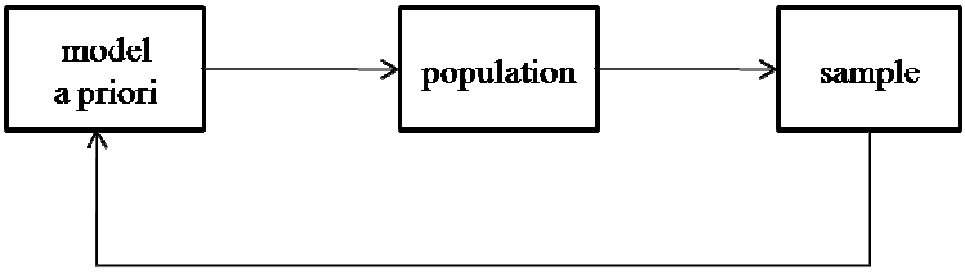

FIGURE 1. The "model *a priori*" assumption in conventional inferential statistics.

1.2 An Illustrative Example

As an illustration of the problem, consider a hypothetical example of linear regression. Suppose, in an empirical study, the researcher has one response variable Y,

and two possible explanatory variables X_1 and X_2. Assume that the researcher is primarily interested in the effect associated with variable X_1 and that variable X_2 is treated as an auxiliary variable. Whether or not X_2 enters into the regression equation depends on analysis of the sample data. Thus, assuming that the "true model" is unknown to the researcher, two candidate models exist:

Unrestricted model: $$y_u = \beta_{0u} + \beta_{1u} x_1 + \beta_{2u} x_2 + \varepsilon_u \quad (1)$$

Restricted model: $$y_r = \beta_{0r} + \beta_{1r} x_1 + \varepsilon_r \quad (2)$$

Suppose model (1) is the "true model." The researcher uses a-data-driven method to select between the two models and then estimates the parameter of interest, β_{1u}. This leads to what has been called the "pretest estimator" (Judge and Bock, 1978). The procedure to obtain the "pretest estimator," β_1^* , say, is:

(1) conduct a two-sided t test for the regression coefficient β_2 where the test hypotheses are:

H_0: β_2=0 vs. H_1: $\beta_2 \neq 0$.

(2) if $|t_{\beta_2}| \geq$ c, where c is the critical value of student's t with α=.05, say, reject H_0 and use the unrestricted model to estimate β_1.

(3) otherwise retain H_0 and use the restricted model to estimate β_1.

In summary,

$$\beta_1^* = \delta \hat{\beta}_{1u} + (1-\delta)\hat{\beta}_{1r} \quad (3)$$

where $\delta = 1$ if $|t_{\beta_2}| \geq$ c or δ= 0 otherwise; note that $\hat{\beta}_{1r}$ is the OLS estimator from the restricted model and $\hat{\beta}_{1u}$ is the OLS estimator from the unrestricted model. For a linear

regression model $y = X\beta + \varepsilon$, the OLS of β is obtained from $\hat{\beta} = (x^T x)^{-1} x^T y$, the variance and standard error are $\text{var}\ \hat{\beta} = (x^T x)^{-1}\sigma^2$, and $\text{se}\left(\hat{\beta}_i\right) = \sqrt{\left(x^T x\right)^{-1}_{ii}}\hat{\sigma}$, and the t statistic is obtained from $t_{\beta_i} = \hat{\beta}_i / se\left(\hat{\beta}_i\right)$.

When this two-stage process is performed over repeated samples, the empirical sampling distribution of the "pretest estimator," β_1^* can be obtained. Clearly, it is a mixture of the distributions of β_{1r} and β_{1u}. Their mixing proportion is the proportion of times when the test hypothesis, H_0: $\beta_2 = 0$, is rejected. In general, the empirical sampling distribution of β_1^* is different from that of β_{1u}, which suggests that the means and standard errors of these distributions might differ to a non-trivial degree.

Figure 2 shows the empirical distributions of these empirical distributions for two estimators over 10,000 repeated samples of sizes 25, 50 and 200 in a simulation study (Gao & Dayton, 2008). For our illustrative purposes, we take a very severe case when the correlation between X_1 and X_2 is -.9, the correlation between Y and X_2 is -.8, and the correlation between Y and X_1 is .8. Plots (a) and (b) show that when sample size is 25 and 50, the distributions are substantially different in terms of mean and variance, even the shapes of distributions are different. While β_{1u} is unimodal, β_1^* becomes bimodal. Also the density β_1^* is more variable than the corresponding density of β_{1u}. When the sample size increases to 200, the difference between the two estimators becomes very small as shown in plot (c). This is a special instance of a more general problem that commonly occurs in regression research.

A natural question to ask is how different the distributions of these estimators are under other combinations of correlation coefficients between variables X_1 and X_2, Y and

X_1, Y and X_2. In a simulation study (Gao & Dayton, 2008), variance ratios of the two estimators were computed from the empirical distributions under different combinations of correlation coefficients from -.9 to .9 in increments of .1 over 10000 samples of size 20, 50, 100 and 200. The relationship among variance ratio and variable correlations is complex and cannot be captured by any simple linear functions (see Figure 3).

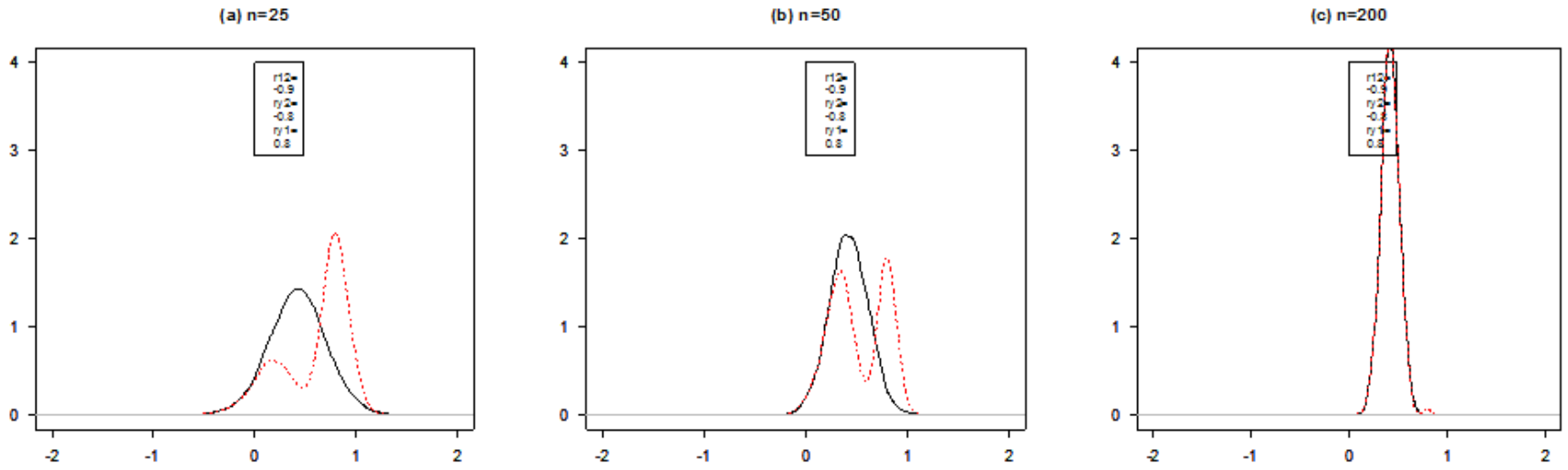

FIGURE 2. Empirical distributions of estimator β_{1u} (solid line) and β_1^* (dashed line)

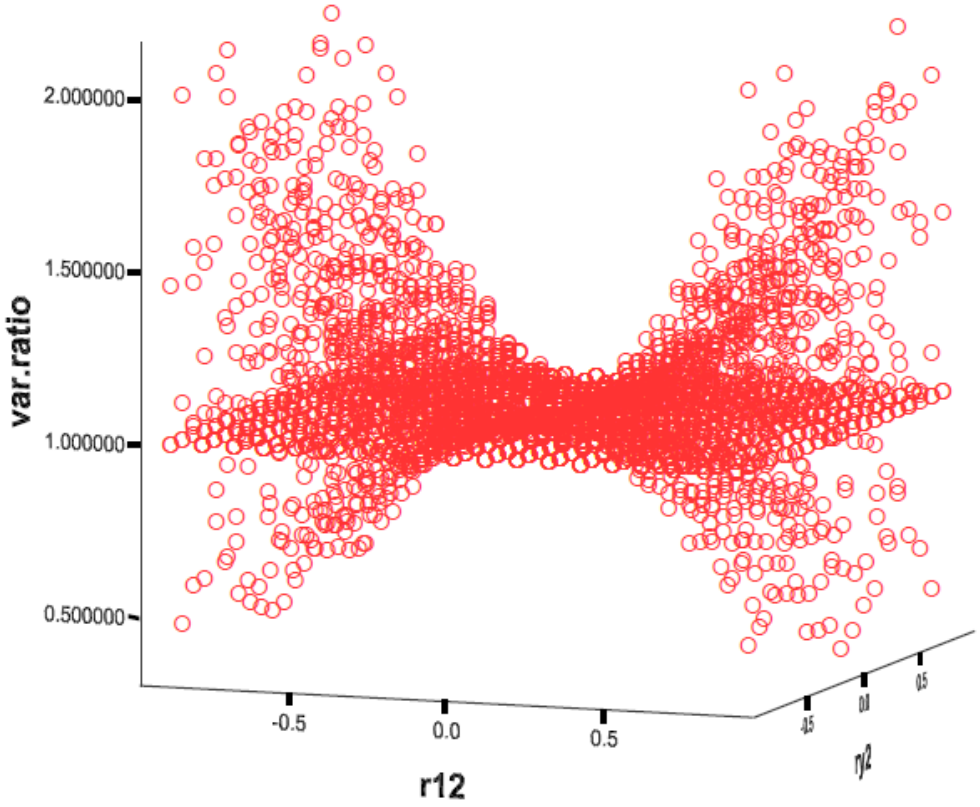

FIGURE 3. Ratio of empirical variance of the post-model-selection estimator to that of estimator under true model assumption, $r_{x_1x_2}$ and r_{yx_2} (n=20)

What makes the model selection problem more serious and hard to analyze is the large number of independent variables in real-life regression analysis. In practical research, with the increase in the computing power, and incorporation of automatic model selection procedure into various software, the problem is exacerbated. Consider the "best subset" model selection procedure in regression analysis, 5 explanatory variables will generate $2^5 = 32$ possible models, not including the interaction terms. When multiple model comparison/selection steps are executed, the bias and difference in standard error might be magnified, but is hard to quantify in finite samples.

1.3 The Purpose of the Current Study

In loglinear analysis of multi-way contingency table, the typical research questions are whether associations of certain factors exist and if they do exist, at which level: i.e., two-factor-effects level and/or three-factor-effects level, etc. To study these questions, a typical strategy is to compare hierarchical models and to apply a selection procedure. Unsaturated (i.e., restricted) models are systematically compared to the unrestricted, saturated model to determine whether the variables interact. An example of this practice is a criminology study of drug use by Rosay, Najaka, and Herz (2007), for which the data comprise a 2^6 contingency table. Hierarchical loglinear models were compared and the contingency table was collapsed so as to include only statistically significant main effects and interactions. They stated "Models are compared using differences in Chi-Square statistics to determine whether the six-way interaction is significant, all five-way, all four-way, all three-way, all two-way, and all main effects are significant. Furthermore, models are compared to determine whether all six and five-way interactions are jointly significant…(p. 51)" Given the nature of the two-step estimation

process, it is apparent that model selection in loglinear modeling is vulnerable to the same criticism as when these techniques are used in linear regression.

Although the topic of the effect of model selection raises a core issue in practical research, it has not been widely investigated. In particular, little research has been conducted on the consequences of model selection in categorical data analysis. As shown in section 2.1, most research papers on the impact of model selection investigate the properties of pretest estimators in multiple regression settings that deal with continuous data. As shown above in the illustrative example, the central issue is that a post-model-selection estimator is in effect a mixture of many potential estimators. Bayesian model averaging (BMA) is a method of incorporating model uncertainty in inference (Hoeting, Madigan, Raftery, & Volinsky, 1999). With the BMA methodology, relatively little effort has been devoted to investigating its actual performance, such as estimator precision. Thus, this study is among the first empirical work to show that the effect of model selection extends to the models for categorical data and to investigate the precision of the BMA method in that setting.

The goal of this study was to (i) investigate the empirical properties of estimators after model selection and compare them with those under the true model assumption in the context of log-linear analysis of cross-classified categorical data; and (ii) evaluate the performance of Bayesian model averaging estimator in the above mentioned context. The magnitude of the bias, relative bias, MSE, and relative efficiency of parameter estimators were calculated under a variety of conditions. The distributions of the z-statistics of the parameters were compared to the reference distribution. Another goal was

to examine the relationship of the model selection problem and sample size as the problem could worsen or alleviate as the sample size grows.

The remainder of this document is organized as follows. Chapter 2 reviews the literature on four topics: the difficulty of model selection, loglinear model theory, model selection in the context of loglinear modeling and Bayesian model averaging. A Monte Carlo simulation study is described in chapter 3, with results from the experiments presented and examined in chapter 4. Discussion of the findings, description of limitation of the study and possibility for future research are given in chapter 5, followed by appendices which provide a case study of BMA method, details of the results of the simulation study, and R code used in the computations and simulations.

Chapter 2: Literature Review

2.1 The Dilemma of Model Selection

In practice, hypothesis tests and other model selection procedures are used to assess the fit of candidate models and a specific model is chosen based on these results. Subsequently, the selected model is interpreted on the basis of the same parameter estimates used in selecting the model. Econometricians call the initial estimators "pretest estimators" (Judge & Bock, 1978). The bias in estimation due to the use of model selection procedures was first investigated by Bancroft (1944). He studied pretest bias in the regression coefficient, β_1 , estimated for the model:

$$y = \beta_1 x_1 + \beta_2 x_2 + e \tag{4}$$

An F test was carried out to decide if regressor x_2 should be retained in the model after the regression model $\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ had been fitted. He derived the mathematical form of the bias of $\hat{\beta}_1$ obtained after the preliminary F test:

$$\text{Bias} = \rho\beta_2 \left[1 - \sum_{j=0}^{\infty} \frac{a^j e^{-a}}{j!} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + j \right) \right], \tag{5}$$

where $x_0 = \dfrac{1}{\dfrac{c}{n-3}+1}$, $a = \dfrac{(1-\rho^2)(n-1)\beta_2^2}{2}$, I_{x_0} is the incomplete beta function, c is the value of F-distribution corresponding to some assigned significance level for 1 and (n-3) degrees of freedom, and ρ is the sample correlation coefficient between the fixed variates x_1 and x_2. This formula suggests that: (1) keeping other factors fixed, bias

decreases as sample size increases; and, (2) bias is proportional to the sample correlation coefficient of the two regressors x_1 and x_2. Bancroft pointed out that when ρ, n, and c are fixed, for increasing β_2, the bias increases at first and then decreases, which means there is no linear relationship between the magnitude of β_2 and the bias of $\hat{\beta}_1$.

Since Bancroft's seminal paper on the effect of model selection, this topic has received considerable attention. Mosteller (1948) assessed the impact of an initial significance test on the pooled estimates of error in analysis of covariance. He gave the closed form for the mean squared error of the pretest estimator under that situation. Huntsberger (1955) generalized the expression of pretest estimator in the area of pooling data. He expressed it as a weighted average of two estimators. These early research studies derived analytical solutions for the properties of the pre-test estimators under specific conditions. Unfortunately, the usefulness of theses results for practitioners is limited because only special conditions of linear regression were considered.

Sclove, Morris and Radhakrishanan (1972) studied the loss functions of the pretest estimators and derived their properties. They found that one undesirable property of pretest estimators is that no pretest estimators can satisfy the minimax criterion. Bankroft and Han (1977) reviewed the early literature on this topic and summarized the difficulties induced by variable selection.

An early simulation study on the effect of model selection can be found in Freedman (1983). He generated a matrix of 100 rows and 51 columns. The numbers in each column were drawn from standard normal distribution independently. Columns 1-50 were treated as independent variables $x_1,...,x_{50}$, and column 51 was taken as dependent variable y. The data were analyzed in two rounds: (1) y was regressed on all 50 of the

$x's$ and (2) only the variables whose coefficients were significant at 0.25 level were kept in the equation and the equation was refitted. Freeman found that "The results from the second pass are misleading indeed, for they appear to demonstrate a definite relationship between Y and the X's, that is, between noise and noise." In addition, he noted that the use of p-values of F test in multiple regression after model selection can be dramatically misleading.

Hurvich and Tsai (1990) conducted a simulation study on the coverage rates of confidence regions for the pretest estimators in linear regression settings. Their result showed that the coverage rates are much smaller for the pretest estimator than the estimator when the model is known as *a prior*. They suggested splitting the data to do "exploratory" and "confirmatory" data analyses. They quoted Tukey (1980, p.821) as asserting "often, confirmation requires a new unexplored set of data." In practice, cross validation is expensive and how to split the data remains unsettled. Danilov and Magnus (2004) found the unconditional first and second moments of the pretest estimator for the linear regression. They showed that the error in the moments varies for different model selection procedures. They also studied the relationship between error and the number of auxiliary regressors. E(UR), the expectation of underreporting ratio, was expressed as a function of sample size, number of variables of interest, and number of auxiliary variables.

Asymptotic properties of the preliminary-test estimators have also been investigated (e.g. Sen ,1979). Pötscher (1991) derived the large sample limit distribution of the preliminary-test estimator in a general setting including linear and nonlinear models. One of his results is that the bias problem vanishes asymptotically when the

model selection criteria are consistent and the variance increases as might be expected from the uncertainty due to the model selection process. He also pointed out that the shape of the distribution may also change. Zhang (1992) studied the asymptotic results for inference on linear regression models when the final prediction error criterion is used to select a model and showed that asymptotic estimates of error variance are satisfactory but that the asymptotic confidence regions for unknown parameters are generally too small in that coverage probabilities are less than nominal probabilities. Pötscher and Novák (1998) conducted a simulation study in which they examined the small sample distribution and compared the small sample distribution to the asymptotic distribution as derived in Pötscher (1991). In the context of linear regression, Leeb and Pötscher (2005) obtained the k-dimensional cumulative distribution function. They stated "a large sample size does not guarantee a small estimation error with high probability when estimating the conditional distribution function of a post-model-selection estimator." The asymptotical results are useful theoretically but not very useful in finite samples. Kabaila (2005) derived a new computationally intense Monte Carlo simulation estimator of the coverage probability, which used conditioning for variance reduction. He also investigated the coverage probability of the 95% confidence interval for post-model-selection estimator. For the real-life data presented in his article, the probability is .79 using AIC and .70 using BIC in as the model selection criterion, confirming that the confidence intervals after model selection are inadequate.

Although these research studies have demonstrated the innate difficulty of model selection, their results are usually not taken into consideration in practice. The reason

may be that no one has derived a correction for estimates and for standard errors due to the complexity of the model uncertainty in common research setting.

2.2 Loglinear Models

Loglinear models are used for modeling cell frequencies in a contingency table as a loglinear combination of effects (model parameters) due to each classification factor singly and in combination. For a three-way table, a loglinear models expresses the expected frequency of cases, F_{ij}, in the cell defined by category *i* of the row dimension, category *j* of the column dimension, and category *k* of the layer dimension (see Bishop, Fienberg, & Holland, 1975; Everitt, 1977; Knoke & Burke, 1980):

$$F_{ijk} = \eta_{ijk} \tau_i^X \tau_j^Y \tau_k^Z \tau_{ij}^{XY} \tau_{ik}^{XZ} \tau_{jk}^{YZ} \tau_{ijk}^{XYZ} \qquad (6)$$

where the "base rate" is η, τ_i^X is the effect of being in category *i* of the row dimension, τ_j^Y is the effect of being in category *j* of the column dimension, τ_k^Z is the effect of being in category k of the layer dimension, τ_{ij}^{XY} is the effect of being in the categories *i* and *j* simultaneously (over and above the effects due to each category separately), etc. Thus, deviations from the "base rate" cell frequency are attributed to the effects of different dimensions and their interactions. Taking logarithms yields a linear relation

$$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \qquad (7)$$

To avoid over-parameterization, necessary constraints are:

$$\lambda_I^X = \lambda_J^Y = \lambda_K^Z = \lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = \lambda_{Ik}^{XZ} = \lambda_{iK}^{XZ} = \lambda_{Jk}^{YZ} = \lambda_{jK}^{YZ} = \lambda_{IJK}^{XYZ} = 0 \qquad (8)$$

With these constraints, parameters such as λ_i^X are the coefficients of dummy variables for the first $(I-1)$ categories of X. The parameters are interpreted in a multiplicative sense.

For example, $\lambda_i^X = 9/10$ means that the frequency of cells in category *i* of dimension X is 146% higher than the base rate since $e^{0.9} = 2.46$.

In loglinear modeling, the "hierarchy principle" (Reynolds, 1977) is often observed. The hierarchy principle prescribes that if a model contains two-factor interactions then it must contain the main effects of the two factors. The same rule extends to higher order interactions that contain main effects as well as lower order interactions. For example, in a three-way table of factors X, Y, and Z, if a model contains interaction XYZ, then it must contain interaction XY, XZ, and YZ. Mathematically, it is possible to include interaction XYZ without including XY. In practice, it makes the model difficult to interpret. The hierarchy principle limits the permissible models and simplifies notation. Thus, for a three-way table, model (XY,Z) implies a model having λ, λ_i^X, λ_j^Y, λ_k^Z and λ_{ij}^{XY}. In this document, we follow the "principle of hierarchy." Table 1 shows the short-handed notation for models of a three-way contingency table.

TABLE 1
Notation for selected log-linear models in three-way tables

Model	Symbol
$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$	(XYZ)
$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	(XY, XZ, YZ)
$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$	(XY, XZ)
$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	(XY, Z)
$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	(X, Y, Z)

In a three-way table, each possible model has an interpretation in terms of association and independence. There are three types of independence relationship in the variables of three-way tables, namely, mutual independence, jointly independence and conditional independence. Table 2 lists these relationships and specific examples. See

Gilbert (1981) and Fienberg (2007) for an extensive discussion of interpretation of independence and association.

TABLE 2
Summary of three types of independence relationship in log-linear models

Model	Interpretation
(X, Y, Z)	Variables X, Y and Z are mutually independent
(XY,Z)	Variable Z is jointly independent of X and Y
(XY, XZ)	Y and Z are conditionally independent, given X

2.3 Model Selection in Loglinear Analysis

In log-linear analysis of three-way contingency tables, model selection is an integral part of the analysis. Popular textbooks on loglinear modeling (e.g., Agresti, 2002; Christensen,1990, 1997; Fienberg, 2007) all suggest the inspection of a series of models. Agresti (2002) suggested the use of likelihood-ratio tests in selecting hierarchical models. However, Raftery (1986) pointed out that the use of likelihood-ratio tests in model selection gives unsatisfactory results when the sample size is large. McCullagh and Nelder (1989) discussed variable selection in the generalized linear model context based on variants of the Akaike information criterion (AIC) (Akaike, 1974):

$$AIC\ (M) = -2\log L\,(M) + 2 \times p \qquad (9)$$

where $\log L\,(M)$ is the maximized log-likelihood and p is the number of independent parameters in model M. The selection procedure is to choose the model with minimum AIC. In general, AIC penalizes the more highly parameterized model. The mathematical reasoning behind AIC is related to the Kullback-Leibler information function (Akaike, 1974). If one replaces $2 \times p$ by $\log(n) \times p$ in equation (9), one obtains the Schwarz (1978) Bayesian Information Criterion (BIC). BIC incorporates a stronger penalty for model complexity (if $n \geq 8$). Useful references on the general topic of model selection

include Linhart and Zucchini (1986), McQuarrie and Tsai (1998), Lahiri (2001), and Miller (2002) as well as special journal issues such as *Journal of Mathematical Psychology* in 2006.

Christensen (1997) suggested a two-step general model selection approach for loglinear models: (i) first choose an initial model from the ones that have all effects of a certain level, such as all main effects, all two-way-effects, and all three-way-effects, etc. (ii) once the initial model is chosen, one can consider removing terms. For three way contingency table, Christensen claims that his inclination is to directly use AIC as the model selection criterion.

To approach the problem of model uncertainty in the model selection process, Good and Crook (1987) and Albert (1989) developed Bayes factors for two-way and three-way contingency tables based on product-multinomial sampling and multinomial sampling schemes. Prior distributions for the cell probabilities are assumed to be normal and the conditional probability of each model is obtained given the observed data. Model selection can be performed by examining the values of the posterior probabilities, Prob(Model M | data). More recently, Madigan and Raftery (1994) used graphical methods for Bayesian model selection in high-dimensional contingency tables. These methods seem attractive, but the choice of prior distributions is arbitrary and computation of the integrals for high dimensional tables is difficult, which may decrease the accuracy of the results and make the methods difficult to apply in practice.

2.4 Bayesian Model Averaging

Due to the dilemma of model selection, an alternative of the traditional approaches of model selection emerged as model averaging. In performing model

averaging, the analyst assumes that each of the candidate models provides a valid estimate of the parameters in its own right. Thus, each model is assigned a weight that reflects the degree to which the data support the model. From the mid-1990s, the "Seattle school of model uncertainty" has studied the use of Bayesian model averaging (BMA) as a method of incorporating model uncertainty in inference (Hoeting, Madigan, Raftery, & Volinsky, 1999). The use of BMA has now been utilized in fields of research such as public health (Morales *et al*., 2006), economics (Fernandez *et al*., 2001) and political science (Geer & Lau, 2006), epidemiology (Viallefont *et al.,* 2001) and air pollution (Shaddick *et al*., 1998; Clyde, 2000).

Let Q denote the quantity of interest (*e.g*., a model parameter) and let $\mathbf{M} = \left\{ \mathrm{M}^{(0)}, \mathrm{M}^{(1)}, \mathrm{M}^{(2)}, ... \mathrm{M}^{(K)} \right\}$ denote the set of candidate models. The law of total probability implies that the posterior probability distribution of Q is

$$\mathrm{pr}(\mathrm{Q} \mid \mathrm{data}) = \sum_{k=0}^{K} \mathrm{pr}\,(\mathrm{Q} \mid \mathrm{M}^{(\mathrm{K})}, \mathrm{data})\ \mathrm{pr}\,(\mathrm{M}^{(\mathrm{K})} \mid \mathrm{data}) \tag{10}$$

The posterior model probability $\mathrm{pr}\,(\mathrm{M}^{(\mathrm{K})} \mid \mathrm{data})$ can be thought of as weights and the quantity $\mathrm{pr}\,(\mathrm{Q} \mid \mathrm{M}^{(\mathrm{K})}, \mathrm{data})$ is the posterior distribution under a specific model. According to the Bayes theorem, the posterior probability of any given model $\mathrm{M}^{(\mathrm{K})}$ is given by

$$\mathrm{pr}\,(\mathrm{M}^{(\mathrm{K})} \mid \mathrm{data}) = \frac{\mathrm{pr}\,(\mathrm{data} \mid \mathrm{M}^{(\mathrm{K})})\,\mathrm{pr}\,(\mathrm{M}^{(\mathrm{K})})}{\sum_{i=0}^{K} \mathrm{pr}(\mathrm{data} \mid \mathrm{M}^{(\mathrm{i})})\,\mathrm{pr}\,(\mathrm{M}^{(\mathrm{i})})} \tag{11}$$

where (i) $\mathrm{pr}\left(\mathrm{data} \mid \mathrm{M}^{(\mathrm{K})}\right) = \int \mathrm{pr}\left(\mathrm{data} \mid \boldsymbol{\theta}_{\mathbf{k}}, \mathrm{M}^{(\mathrm{K})}\right) \mathrm{pr}\left(\boldsymbol{\theta}_{\mathbf{k}} \mid \mathrm{M}^{(\mathrm{K})}\right) \mathrm{d}\boldsymbol{\theta}_{\mathbf{k}}$ (12)

(ii) $\mathrm{pr}\left(\mathrm{data} \mid \mathrm{M}^{(\mathrm{K})}\right)$ is the likelihood of model $\mathrm{M}^{(\mathrm{K})}$,

(iii) $\boldsymbol{\theta}_{\mathbf{k}}$ is the parameter set for model $\mathrm{M}^{(\mathrm{K})}$,

(iv) $\mathrm{pr}(\mathrm{M}^{(\mathrm{K})})$ is the prior probability for model $\mathrm{M}^{(\mathrm{K})}$ being the true model.

The Bayesian point estimate of θ_1 is its posterior mean, as in

$$E(\theta_1 \mid \text{data}) = \sum_{i=0}^{K} \hat{\theta}_1^K \, pr(M^{(K)} \mid \text{data}) \tag{13}$$

where $\hat{\theta}_1^K = E(\theta_1 \mid \text{data}, \mathrm{M}^{(\mathrm{K})})$, i.e., $\hat{\theta}_1^K$ is the posterior mean of θ_1 under model $\mathrm{M}^{(\mathrm{K})}$. When θ_1 is not included in a particular model, this term is zero. Note that, $E(\theta_1 \mid \text{data})$ is a weighted average of the model-specific point estimates and the weights are posterior model probabilities. The Bayesian standard error, the posterior standard deviation of θ_1, is the square root of

its variance:

$$\mathrm{var}(\theta_1 \mid \text{data}) = \sum_{i=0}^{K} \left\{ (\mathrm{var}(\theta_1 \mid \text{data}, \mathrm{M}^{(\mathrm{K})}) + (\hat{\theta}_1^K)^2) \, \mathrm{pr}(\mathrm{M}^{(\mathrm{K})} \mid \text{data}) - E\left[\theta_1^K \mid \text{data}\right]^2 \right\} \tag{14}$$

where $\mathrm{var}(\theta_1 \mid \text{data}, \mathrm{M}^{(\mathrm{K})})$ is the variance of θ_1 under particular model given the data.

From a Bayesian point of view, hypothesis testing is replaced by the question "what is the posterior probability that θ_1 is not equal to zero?" This is given by the sum of the posterior probabilities of the models that include θ_1:

$$\Pr(\theta_1 \neq 0 \mid \text{data}) = \sum_{\mathrm{M}^{(\mathrm{k})}: \theta_1 \in M^{(\mathrm{K})}} \mathrm{pr}(\mathrm{M}^{(\mathrm{K})} \mid \text{data}) \tag{15}$$

The quantity $\Pr(\theta_1 \neq 0 \mid \text{data})$ indicates that the probability that θ_1 is included and estimated in at least one model. The conventional rule of thumb (Viallefont *et al*., 2001). for interpreting this quantity in terms of evidence for the existence of θ_1; is: values less than 0.5 suggest no evidence; values between 0.5 and 0.75 suggest weak evidence; values

between 0.75 and 0.95 suggest positive evidence; between 0.95 and 0.99 suggest strong evidence; and beyond 0.99 suggest very strong evidence.

For generalized linear models such as loglinear models, Raftery (1996) proposed the using of a Bayes factor to compute the posterior probability of a specific model. Suppose the models $M^{(1)}$ and $M^{(2)}$ are parameterized by vectors of parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Thus the Bayes factor K_{12} is given by (Raftery, 1996):

$$K_{12} = \frac{\text{pr}(\text{data} \mid \text{M}^{(1)})}{\text{pr}(\text{data} \mid \text{M}^{(2)})} = \frac{\int \text{pr}(\boldsymbol{\theta}_1 \mid \text{M}^{(1)})\ \text{pr}(\text{data} \mid \boldsymbol{\theta}_1, \text{M}^{(1)})\ \text{d}\boldsymbol{\theta}_1}{\int \text{pr}(\boldsymbol{\theta}_2 \mid \text{M}^{(2)})\ \text{pr}(\text{data} \mid \boldsymbol{\theta}_2, \text{M}^{(2)})\ \text{d}\boldsymbol{\theta}_2} \qquad (16)$$

The Bayes factor is a piece of evidence given by data for $\text{M}^{(1)}$ over $\text{M}^{(2)}$. For the model space $\mathbf{M}$, let model $\text{M}^{(1)}, \text{M}^{(2)}, \ldots, \text{M}^{(\text{K})}$ all compare with $M^{(0)}$. This generates $K+1$ Bayes factors. Then the posterior probability of model $\text{M}^{(\text{k})}$ is given by (Raftery, 1996)

$$\text{pr}(\text{M}^{(\text{K})} \mid \text{data}) = \frac{\dfrac{\text{pr}(\text{M}^{(\text{K})})}{\text{pr}(\text{M}^{(0)})} K_{10}}{\displaystyle\sum_{i=0}^{K} \left(\frac{\text{pr}(\text{M}^{(\text{i})})}{\text{pr}(\text{M}^{(0)})} K_{i0} \right)} \qquad (17)$$

where $\dfrac{\text{pr}(\text{M}^{(\text{K})})}{\text{pr}(\text{M}^{(0)})}$ is the prior odds for model $\text{M}^{(\text{K})}$ against $\text{M}^{(0)}$ which is often assumed to be one in computation.

The BMA approach requires the specification of two types of prior distributions: (i) the prior probabilities of the models $\text{pr}(\text{M}^{(\text{K})})$, and (ii) the prior distributions of the parameters $\boldsymbol{\theta}$ given model $\text{M}^{(\text{K})}$. To make the computation easier, the variables are assumed to be standardized. The prior distribution is assumed to be normal, and distributed as $\left(\boldsymbol{\theta} \mid \text{M}^{(\text{k})}\right) \sim N(\mathbf{v}, \mathbf{U})$, where $\mathbf{v} = (v_1, 0, \ldots, 0)$ and $\mathbf{U} = \mathbf{diag}\left\{\psi^2, \phi^2, \ldots, \phi^2\right\}$.

Raftery (1996) suggested let $\nu_1 = 0$ and $\phi = 1$. For the specification of the parameter ϕ, Raftery (1996) suggested "it is usually better to report, or at least to consider, the result from a range of reasonable values of ϕ." (p. 259) The range he recommended is $1 \leq \phi \leq 5$, with 1.65 as a "central" value.

Chapter 3: Research Design

3.1 Purpose of the Simulation Study

We performed a large-scale simulation study in which several factors were varied in order to:

(1) investigate the performance of the model parameter estimator under (a) the true model assumption, (b) under model selection by the AIC criterion (i.e., the post-model-selection estimators) and (c) under Bayesian model averaging estimators using three different prior distributions;

(2) compare the empirical distribution of the Z-statistics for λ under model selection and normal distribution, which is the reference distribution;

(3) examine the performance of the estimators of the conditional odds ratio under the true model assumption, under model selection, and under BMA using one set of prior distributions.

3.2 Sampling Schemes

There are three common sampling schemes in contingency tables: multinomial, Poisson and product-multinomial.

(1) multinomial sampling scheme: When the sample size is fixed at N, and there are *m* cells (cross-classification conditions), for each trial of N, it must be classified as one of the *m* conditions. Suppose $n_1, n_2, ..., n_m$ are the number of events happens under the *m* conditions and let p_i, i=1,2,…, m be the probability that the i^{th} event occurs on that occasion. The multinomial distribution is written as:$(n_1, n_2, ..., n_m) \sim MN(N, p_1, p_2, \ldots, p_m)$.

The probability mass function of multinomial distribution is

$$\Pr(n_1 = b_1,..., n_m = b_m) = \frac{N!}{b_1!...b_m!} p_1^{b_1} ... p_m^{b_m}.$$

(2) Poisson sampling scheme: The Poisson distribution is used to model counts of event that occur randomly over time or space. The probability mass function of Poisson distribution is

$$Pr(n = x) = \lambda^x e^{-\lambda} / x! \qquad for\ x = 0, 1, 2,...$$

If the sample size is fixed, then independent counts in the *m* cells of the Poisson process follow the multinomial distribution.

(3) product-multinomial sampling scheme: If for each level of one variable, or the combination of levels of two variable, a multinomial sample of size n_{i+} is sampled, then the resulting distribution is product-multinomial. One reason for the product-multinomial sampling design may be the proportions of the sample size, n_{i+}, reflect the proportions in the population. In product-multinomial modeling, any model that has the term λ_{ij}^{XY} automatically has these margins fixed (see Christensen, 1997).

3.3 Fixed Factors

The following aspects were fixed in the simulation study: (1) number of replications, (2) estimation methods, (3) true model for data generation. The number of replications was set at 10000.

Two computational algorithms for maximum-likelihood estimation are commonly used in loglinear modeling: Newton-Raphson and iterative proportional fitting (IPF). IPF (Deming & Stephan,1940) is a simple method for calculating MLEs of cell frequencies that does not involve matrix inversion. IPF calculates the expected cell frequencies, but

does not produce estimates for model parameters or associated standard errors. The Newton-Raphson method does involve matrix inversion to produce model parameter estimates and is more efficient numerically than IPF since the rate of convergence is quadratic rather than linear (Agresti, 2002). In general, both methods yield the same results. In this study, the Newton-Raphson method was used because estimates and standard errors were desired. With respect to sampling schemes, Birch (1963) showed that the MLEs are the same for independent Poisson sampling, simple multinomial sampling, and product- multinomial sampling.

The true model that generates data is (XY, XZ) under all conditions. This model was chosen so that both the effects of over-fitting [e.g., model (XY,XZ,YZ)] and under-fitting [e.g, model (XY,Z)] could be studied.

3.4 Manipulated Factors

The following aspects were systematically manipulated in the simulation study: (1) sample size, (2) sampling schemes, (3) level of factors in a three-way contingency table and (4) magnitude of main and association effects.

A pilot study, summarized in Appendix C, was conducted that showed that sample size was an important factor that influenced the magnitude of relative bias and variance ratios. Thus, five sample sizes were selected, 50, 100, 200, 500, and 1000. When sample size is 50, the average cell frequency is 6 for 2×2×2 tables, which creates no problem for parameter estimation. However, with 2×2×3 tables, the average cell frequency is 4 which is slightly deficient judged by the rule of thumb for average cell size of 5. A sample size of 1000 was chosen to check if the problems induced by model selection were only small sample problems.

When the sample size is fixed, the Poisson distribution and the multinomial distribution are equivalent. We took account of the nature of the product-multinomial sampling design by restricting our attention to models that include the term λ_{ij}^{XY} since this restriction reduces the number of possible models to a manageable level as listed in Table 3. Under multinomial sampling, there are 8 models that include all the main effects as shown in Table 4.

TABLE 3
Unique models with the product- multinomial sampling design, fixing $\mu_{ij.}$

Model	Degree of Freedom
$M^{(1)}$: (XYZ)	*0*
$M^{(2)}$: (XY, YZ, XZ)	*(I-1)(J-1)(K-1)*
$M^{(3)}$: (XY, XZ)	*I(J-1)(K-1)*
$M^{(4)}$: (XY, YZ)	*J(I-1)(K-1)*
$M^{(5)}$: (XY, Z)	*(K-1)(IJ-1)*

TABLE 4
Unique models with the multinomial sampling assuming all main effects exist

Model	Degree of Freedom
$M^{(1)}$: (XYZ)	*0*
$M^{(2)}$: (XY, YZ, XZ)	*(I-1)(J-1)(K-1)*
$M^{(3)}$: (XY, XZ)	*I(J-1)(K-1)*
$M^{(4)}$: (XY, YZ)	*J(I-1)(K-1)*
$M^{(5)}$: (XZ, YZ)	*K(I-1)(J-1)*
$M^{(6)}$: (XY, Z)	*(K-1)(IJ-1)*
$M^{(7)}$: (XZ, Y)	*(J-1)(IK-1)*
$M^{(8)}$: (YZ, X)	*(I-1)(JK-1)*

In this simulation study, two types of three way table are investigated, namely, $2\times 2\times 2$ and $2\times 2\times 3$. The number of parameters of the true model, i.e., the λ's, is six in the former table and ten in the latter table. The true model that generates data was assumed to be (XY,XZ).

The data were generated for different values of λ's. For the $2\times2\times2$ table, based on Pardo and Pardo's study (2003), two different sets of value were chosen: the first set implies small values for the main effects and interactions $\lambda_1^X = \lambda_1^Y = \lambda_1^Z = -.3$, $\lambda_{11}^{XY} = .3$ and $\lambda_{11}^{XZ} = -.3$. The second set implies big values for the main effects and interactions $\lambda_1^X = \lambda_1^Y = \lambda_1^Z = 1$, $\lambda_{11}^{XY} = 1$ and $\lambda_{11}^{XZ} = -2$. For the $2\times3\times3$ table, the third set of parameters was defined as "small": $\lambda_1^X = \lambda_1^Y = \lambda_1^Z = \lambda_2^Y = \lambda_2^Z = -.3$, $\lambda_{11}^{XY} = -.3, \lambda_{11}^{XZ} = -.3$ and $\lambda_{12}^{XZ} = -.3$ and the fourth parameter set implies large effects, $\lambda_1^X = \lambda_1^Y = \lambda_1^Z = \lambda_2^Y = \lambda_2^Z = 1$, $\lambda_{11}^{XY} = 1$, $\lambda_{11}^{XZ} = -2$, and $\lambda_{12}^{XZ} = -2$. The specifications of the parameters are summarized in Table 5.

TABLE 5
Simulated main and interaction effect parameters

$2\times2\times2$ table		$2\times2\times3$ table	
Set One	Set Two	Set Three	Set Four
$\lambda_1^X = -.3$	$\lambda_1^X = 1$	$\lambda_1^X = -.3$	$\lambda_1^X = 1$
$\lambda_1^Y = -.3$	$\lambda_1^Y = 1$	$\lambda_1^Y = -.3$	$\lambda_1^Y = 1$
$\lambda_1^Z = -.3$	$\lambda_1^Z = 1$	$\lambda_1^Z = -.3$	$\lambda_1^Z = 1$
$\lambda_{11}^{XY} = -.3$	$\lambda_{11}^{XY} = 1$	$\lambda_2^Z = -.3$	$\lambda_2^Z = 1$
$\lambda_{11}^{XZ} = -.3$	$\lambda_{11}^{XZ} = -2$	$\lambda_{11}^{XY} = -.3$	$\lambda_{11}^{XY} = 1$
		$\lambda_{11}^{XZ} = -.3$	$\lambda_{11}^{XZ} = -2$
		$\lambda_{12}^{XZ} = -.3$	$\lambda_{12}^{XZ} = -2$

For each generated dataset, three types of estimates were obtained: (1) parameter estimates under the true model (XY, XZ), (2) parameter estimates using BMA (with ϕ=1, 1.65, 5); and (3) the post-model-selection estimator using AIC as the model selection criterion. The choice of the prior parameters ϕ is based on the results of Raftery

(1996), where he mentioned that at $e^{1/2}(=1.65)$, ϕ balances two criteria. He also stated "the log-Bayes factor changes rapidly as a function of ϕ for $\phi < 1$, and then changes much more slowly over this preferred range of value of $\phi\,[1 \le \phi \le 5]$." Under the product-multinomial sampling schemes, the "best" model was chosen from five models as shown in Table 3 and under multinomial sampling schemes, the "best" model was chosen from eight models as shown in Table 4. The empirical distributions of these estimators were studied and compared. Table 6 summarizes the cases for this simulation study.

TABLE 6
Cases for the Simulation Study

Case #	Design
1	Product-multinomial sampling: parameter set 1 selected from 5 models in Table 4
2	Product multinomial sampling: parameter set 2 selected from 5 models in Table 4
3	Product multinomial sampling: parameter set 3 selected from 5 models in Table 4
4	Product multinomial sampling: parameter set 4 selected from 5 models in Table 4
5	Multinomial sampling: parameter set 1 selected from 8 models in Table 5
6	Multinomial sampling: parameter set 2 selected from 8 models in Table 5
7	Multinomial sampling: parameter set 3 selected from 8 models in Table 5
8	Multinomial sampling: parameter set 4 selected from 8 models in Table 5

3.5 Data Generation

The simulation was done in R. The data were generated based on the true model (XY, XZ): The log-linear representation of the model is:

$$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}\,. \qquad (18)$$

Before data can be generated, the value of λ (constant) must be calculated such that the sum of the μ_{ijk}'s is equal to the sample size. For example, when sample size is 50, the data generation proceeds as follows:

Step 1. Calculate λ by solving the following equation:

$$\exp(\lambda)\exp(\lambda_1^X)\exp(\lambda_1^Y)\exp(\lambda_1^Z)\exp(\lambda_{11}^{XY})\exp(\lambda_{11}^{XZ})+$$
$$\exp(\lambda)\exp(\lambda_1^X)\exp(\lambda_1^Z)\exp(\lambda_{11}^{XZ})+$$
$$\exp(\lambda)\ \exp(\lambda_1^Y)\exp(\lambda_1^Z)\ +$$
$$\exp(\lambda)\ \exp(\lambda_1^Z)\ +$$
$$\exp(\lambda)\exp(\lambda_1^X)\exp(\lambda_1^Y)\ \exp(\lambda_{11}^{XY})+$$
$$\exp(\lambda)\exp(\lambda_1^X)\ +$$
$$\exp(\lambda)\ \exp(\lambda_1^Y)\ +$$
$$\exp(\lambda)=50$$

Step 2. Calculate $\mu_{111}, \mu_{121}, ..., \mu_{222}$ from equation (18).

Step 3. Calculate the cell probabilities π_{ijk}, as $\pi_{ijk} = \mu_{ijk}/50$.

Step 4. Generate the data according to the multinomial distribution, $(n_1, n_2, ..., n_8) \sim MN(50, \pi_{111}, ...\pi_{222})$. The R function "rmultinom(n, size, prob)" was used in generating random multinomial distributions.

Step 5. Since the cell probabilities π_{ijk} remain the same for the model (XY, XZ) even under different sample sizes, the only change when generating other data with other sample size is to specify different "size" values in the R function "rmultinom(n, size, prob)". That is, one does not need to go through step 1 through 4 to obtain the same cell probability $(\pi_{111}, ..., \pi_{222})$. The parameters were estimated via the R function glm, which utilizes the Newton-Raphson algorithm (Thompson, 2009).

3.6 Evaluation Criteria

3.6.1 Assessing Model Parameter Estimators

The performance of (i) the post-model-selection estimator, (ii) the estimator under the "true model" assumption, and (iii) three BMA estimators (phi=1, 1.65, 5) were evaluated using criteria such as Empirical Bias, MSE.

Some indicators of the quality of an estimator are the bias, standard error, and mean square error (MSE). Bias is calculated as the expected value of the estimates by the sample average over the Monte Carlo iterations minus the true value of the parameter. For simulation with B replicates,

$$Bias\,(\lambda) = \frac{1}{B}\sum_{\mathrm{i}=1}^{\mathrm{B}} \hat{\lambda}_i - \lambda \tag{19}$$

$$s.e. = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\lambda}_i - \overline{\hat{\lambda}}\right)^2} \tag{20}$$

where $\overline{\hat{\lambda}} = \frac{1}{B}\sum_{i=1}^{B}\hat{\lambda}_i$

$$MSE\,(\lambda) = \frac{1}{\mathrm{B}}\sum_{i=1}^{B}(\hat{\lambda}_i - \lambda)^2 \tag{21}$$

The reported results also include relative efficiency. The Relative Efficiency of estimator λ_1 to estimator λ_2, which is given by

$$\text{REL EFF} = \frac{E\,(\lambda_2 - \lambda)^2}{E\,(\lambda_1 - \lambda)^2} = \frac{MSE\,(\lambda_2)}{MSE\,(\lambda_1)} \tag{22}$$

A REL EFF value above one indicates that estimator λ_1 is more effective in reducing estimation errors than estimator λ_2. This criterion has been used to compare the efficiency of two types of estimators in the literature of simulation study (Yang and Xie, 2003).

3.6.2 Assessing Distribution Assumptions of Z Tests

In the loglinear model, a Z test is used to test the statistical significance of each coefficient (λ's) in the model where Z is given by

$$\frac{(\hat{\lambda}) - \lambda_{null}}{se\,(\hat{\lambda})} \tag{23}$$

where λ_{null} is the parameter value used to generate the data. Although Z statistics are known to be poorly behaved with small sample sizes (Agresti, 2002), they are still routinely used in practice because they are based on the estimates and are easy to implement. The research interest was in whether or not the empirical Z statistics distribution under model selection deviated in a dramatic fashion from its theoretical distribution, the normal distribution. The .01, .025, .05, .1, .5, .9, .95, .975, .99 percentiles of the calculated quantity (23) were reported for model selection conditions and for the true model condition. The percentiles were compared with the critical values from the theoretical Z distribution. In this way, any strong deviations from the Z distribution could be detected. Quantile-quantile (Q-Q) plots of the theoretical Z distribution versus the distribution of the computed Z statistic were graphed. The skewness of the computed Z, if any, was investigated. The interest was to find out if the Z-based inference leads to conservative or liberal conclusion under the model selection conditions.

3.6.3 Assessing Estimated Conditional Odds Ratio in Two-by-two-by-two Table

For a two-by-two table, odds ratio is

$$\theta = \frac{\pi_1 / (1-\pi_1)}{\pi_2 / (1-\pi_2)} = \frac{\pi_{11} / \pi_{12}}{\pi_{21} / \pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{24}$$

If $\theta = 1$, this suggests the independence of the two classification variables. If θ is different than 1, some degree of association between the two variables can be inferred. However, Feinstein (1973) strongly criticized using odds ratios as a measure of association, pointing out that rates of the marginal variables are lost. Since this happens

only when the event is very rare, such as certain diseases, we do not face this problem in the general setting.

Odds ratios are the same under the Poisson, multinomial, and product-multinomial sampling schemes. The sample odds ratio is estimated as

$$\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21}) \tag{25}$$

The problem with this estimator is that if any of the cell counts equals zero, $\hat{\theta}$ would be estimated as zero or infinity and the estimator is undefined. Gart and Zweifel (1967) and Haldane (1956) recommended an adjusted odds ratio which adds .5 to each cell frequency. And Gart (1966) showed that this estimator behaves well.

$$\tilde{\theta} = \frac{(n_{11}+.5)(n_{22}+.5)}{(n_{12}+.5)(n_{21}+.5)} \tag{26}$$

If a loglinear model is fitted to the data, the odds ratio can be estimated from the fitted value of the estimated cell counts generated by the model. In this simulation study, due to the possibility of cell counts of zero in the generated data, we used the adjusted fitted value of cell count to estimate odds ratios:

$$\hat{\theta} = \frac{(u_{11}+.5)(u_{22}+.5)}{(u_{12}+.5)(u_{21}+.5)} \tag{27}$$

Odds ratios can only express the probability relationship in a two-by-two table. However, for loglinear models based on a $2\times 2\times 2$ table, odds ratios are still useful in describing the dependence relationship among the variables.

For three-way tables, conditional odds ratios between two variables are the odds ratios computed at a fixed level of the third variable and marginal odds ratios between two variables are the odds ratio computed based on cell counts collapsed over all the levels of the third variable. If a model, such as model (XY, XZ), has no three-way

association term, then conditional odds ratios between any two variables are the same at each level of the third variable. Therefore, among the five models to be selected in Table 3, conditional association (conditional odds ratio) between variables X and Y is the same for level 1 or Z as for level 2 of Z. In model (XYZ), conditional association (conditional odds ratio) between X and Y has to be estimated at different levels of variable Z. In this simulation study, the conditional odds ratio between any two variables is estimated at level 1 of the third variable.

The means and standard errors of the conditional odds ratios and marginal odds ratios are reported under three conditions: (i) true model assumption, (ii) after model selection, and (iii) using the BMA method.

Chapter 4: Results

Detailed results for cases 1 through 8 are presented in Appendix B, Tables B1 through B37. Although each case has some distinct characteristics, general patterns of the performance of the five estimators are similar across cases. Therefore, in section 4.1, case 1 is presented in detail to illustrate the performance of the five estimators in terms of bias, standard error, MSE, and relative efficiency. Also, the distribution of the Z-statistics of the model parameters in comparison with the reference distribution, under model selection and true model assumption is examined. Finally, the conditional odds ratios of the two-by-two-by-two table under the five estimation methods are investigated. Section 4.2 presents the general results in all eight cases of the simulation study and compares them with case 1 results.

4.1 Case 1 as an Example

4.1.1 Bias, Standard Error, MSE, and Relative Efficiency of Parameter Estimators

Figure 4 and Table 7 present the percentage of time that each candidate model was selected by the AIC model selection criteria for case 1. The candidate models are classified into four categories: the underspecified/underfitted models, the overspecified/overfitted models, the misspecified/misfitted model, and the true model. When the true model is itself a candidate model, overfitting refers to choosing a model with additional variables and underfitting refers to choosing a model with fewer variables. The term “misfit” refers to choosing a model with one or more wrong variables. Both overfitting and underfitting have been known to reduce efficiency and decrease the predictive abilities of a model. (McQuarrie & Tsai,1998). Underfitted

models suffer from the lack of details since it reduces the complexity of the true model. With overfitted models, variations in the extra variables tend to result in larger variances in the predictions. At smaller sample sizes 50, 100, and 200, AIC favors the underspecified model (XY, Z) (over 50%), and the true model is chosen with relative small percentages of 16%, 20%, and 27%, respectively. However, when sample size increases, the percentage of time that the true model is chosen also increases. When sample size reaches 1000, the percentage of times of choosing the true model has increased to 62%; the percentage of overfitting by model (XY, XZ, YZ) or (XYZ) has increased to around 18%; and the percentage of unfitting has reduced to 17%. As expected with AIC, the penalty function of the AIC criteria is playing a more important role when the sample size is smaller, causing more underfitting in smaller samples than in larger samples.

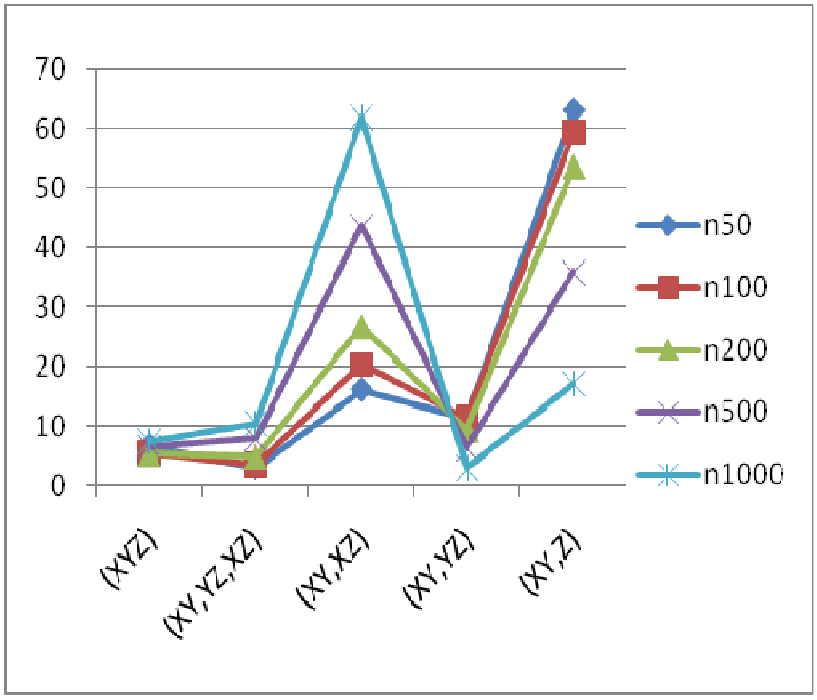

FIGURE 4. Percentage of time each model was selected by AIC

TABLE 7
Percentage of time each candidate model was selected by AIC (case 1)

Models		%
n=50		
$M^{(1)}$: (XYZ)	Overspecified	6.61
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	3.06
$M^{(3)}$: (XY, XZ)	**True model**	16.06
$M^{(4)}$: (XY, YZ)	Misspecified	11.30
$M^{(5)}$: (XY, Z)	Underspecified	62.97
n=100		
$M^{(1)}$: (XYZ)	Overspecified	5.47
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	3.47
$M^{(3)}$: (XY, XZ)	**True model**	20.30
$M^{(4)}$: (XY, YZ)	Misspecified	11.38
$M^{(5)}$: (XY, Z)	Underspecified	59.38
n=200		
$M^{(1)}$: (XYZ)	Overspecified	5.38
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	4.90
$M^{(3)}$: (XY, XZ)	**True model**	26.76
$M^{(4)}$: (XY, YZ)	Misspecified	9.41
$M^{(5)}$: (XY, Z)	Underspecified	53.55
n=500		
$M^{(1)}$: (XYZ)	Overspecified	6.40
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	7.87
$M^{(3)}$: (XY, XZ)	**True model**	43.67
$M^{(4)}$: (XY, YZ)	Misspecified	6.23
$M^{(5)}$: (XY, Z)	Underspecified	35.83
n=1000		
$M^{(1)}$: (XYZ)	Overspecified	7.46
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	10.49
$M^{(3)}$: (XY, XZ)	**True model**	61.89
$M^{(4)}$: (XY, YZ)	Misspecified	2.97
$M^{(5)}$: (XY, Z)	Underspecified	17.19

Tables 8 and 9 present the bias, standard error, MSE, and relative efficiency of the five estimators using product-multinomial sampling with varying sample sizes. In these tables, "True Model Assumption" designated cases with no model selection process; "Model Selection" refers to model selection using AIC criterion, and BMA refers to the estimator obtained using Bayesian Model Averaging method. As expected, for parameter λ_{11}^{XY} and λ_{11}^{XZ}, bias goes down for all five estimators as the sample size increases from 50 to 1000. The maximum likelihood estimator is known to be biased in finite samples (Schaefer, 1983; Firth, 1993) but converges to the true value as sample size approaches infinity. This pattern is seen in the estimator under the

true model assumption. However, it is not true for the post-model-selection estimator or the BMA estimator. For example, when sample size is 1000, the bias for $\hat{\lambda}_{11}^{XZ}$ becomes negligible (-.002) under the true model assumption, while the bias is .022 under model selection (in absolute value, about 10 times larger than -.002). Under some sample sizes, BMA estimators exhibit larger bias than the post-model-selection estimator. However, the standard error of the post-model-selection estimator is always greater than the standard error of the estimator under true model assumption. The BMA estimators generally have smaller standard error than the post-model-selection estimator with no exceptions for either $\hat{\lambda}_{11}^{XY}$ or $\hat{\lambda}_{11}^{XZ}$ in case 1.

TABLE 8

Simulation results of the estimates of $\hat{\lambda}_{11}^{XY}$ in case 1

Sample size	Estimate	Method				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.061	-0.154	-0.120	-0.110	-0.084
	SE	0.847	1.866	1.088	1.010	0.846
	MSE	0.721	3.506	1.199	1.032	0.723
	Rel EFF		4.863	1.663	1.431	1.003
100	Bias	-0.013	-0.022	-0.018	-0.017	-0.015
	SE	0.440	0.643	0.488	0.483	0.469
	MSE	0.193	0.414	0.238	0.233	0.220
	Rel EFF		2.145	1.233	1.207	1.140
200	Bias	-0.009	-0.011	-0.010	-0.010	-0.009
	SE	0.308	0.334	0.311	0.310	0.308
	MSE	0.095	0.112	0.097	0.096	0.095
	Rel EFF		1.179	1.021	1.011	1.000
500	Bias	-0.004	-0.005	-0.004	-0.004	-0.004
	SE	0.191	0.210	0.193	0.192	0.191
	MSE	0.037	0.044	0.037	0.037	0.037
	Rel EFF		1.189	1.000	1.000	1.000
1000	Bias	0.001	0.001	0.001	0.001	0.001
	SE	0.135	0.149	0.136	0.136	0.135
	MSE	0.018	0.022	0.019	0.018	0.018
	Rel EFF		1.222	1.056	1.000	1.000

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE 9

Simulation results of the estimates of $\hat{\lambda}_{11}^{XZ}$ in case 1

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.048	-0.049	-0.010	0.036	0.139
	SE	0.790	2.040	1.064	0.960	0.701
	MSE	0.627	4.165	1.132	0.923	0.511
	Rel EFF		6.643	1.805	1.472	0.815
100	Bias	-0.015	0.081	0.094	0.130	0.200
	SE	0.441	0.576	0.368	0.338	0.262
	MSE	0.194	0.338	0.145	0.131	0.109
	Rel EFF		1.742	0.747	0.675	0.562
200	Bias	-0.008	0.080	0.107	0.141	0.203
	SE	0.308	0.358	0.261	0.243	0.197
	MSE	0.095	0.134	0.080	0.079	0.080
	Rel EFF		1.411	0.597	0.988	1.013
500	Bias	0.000	0.052	0.106	0.135	0.190
	SE	0.195	0.250	0.200	0.193	0.171
	MSE	0.038	0.065	0.051	0.056	0.066
	Rel EFF		1.711	1.342	1.474	1.737
1000	Bias	-0.002	0.022	0.079	0.103	0.151
	SE	0.136	0.181	0.173	0.175	0.173
	MSE	0.018	0.033	0.036	0.041	0.053
	Rel EFF		1.833	2.000	2.278	2.944

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

Figure 5 reports the MSE of $\hat{\lambda}_{11}^{XY}$ and $\hat{\lambda}_{11}^{XZ}$ in case 1. In each panel of the plot, the MSE is denoted by symbols connected by the solid line, and different types of points denote different sample size results. As expected, the estimator under the true model assumption always outperforms the post-model-selection estimator. For the parameter, λ_{11}^{XY}, the relative efficiency (REL EFF) of the post-model-selection estimator versus the estimator under true model assumption is 4 when sample size is 50, goes down to around 2 when sample size increases to 100, and stabilizes at about 1.2 when the sample size is 200, 500, or 1000. The relative efficiencies are smaller for the three BMA estimators; all of them are lower than 1.7 at all sample sizes. For the parameter, λ_{12}^{XY}, the relative efficiencies for the three BMA estimators are smaller than those for the post-model-selection estimator under all sample sizes with the exception of sample size of 1000. At sample size of 1000, the relative efficiency is 1.8, 2.0, 2.3, and 2.9 for post-model-selection estimator, BMA^1, BMA^2, BMA^3, respectively. On the whole, the poster-model-selection estimator is less efficient (in terms of MSE) than the estimator under true model assumption. When sample size increases, the magnitude of inefficiency decreases. Generally speaking, the three BMA estimators outperform the post-model-selection estimators in terms of MSE, although the existence of several exceptions leads to no general conclusion.

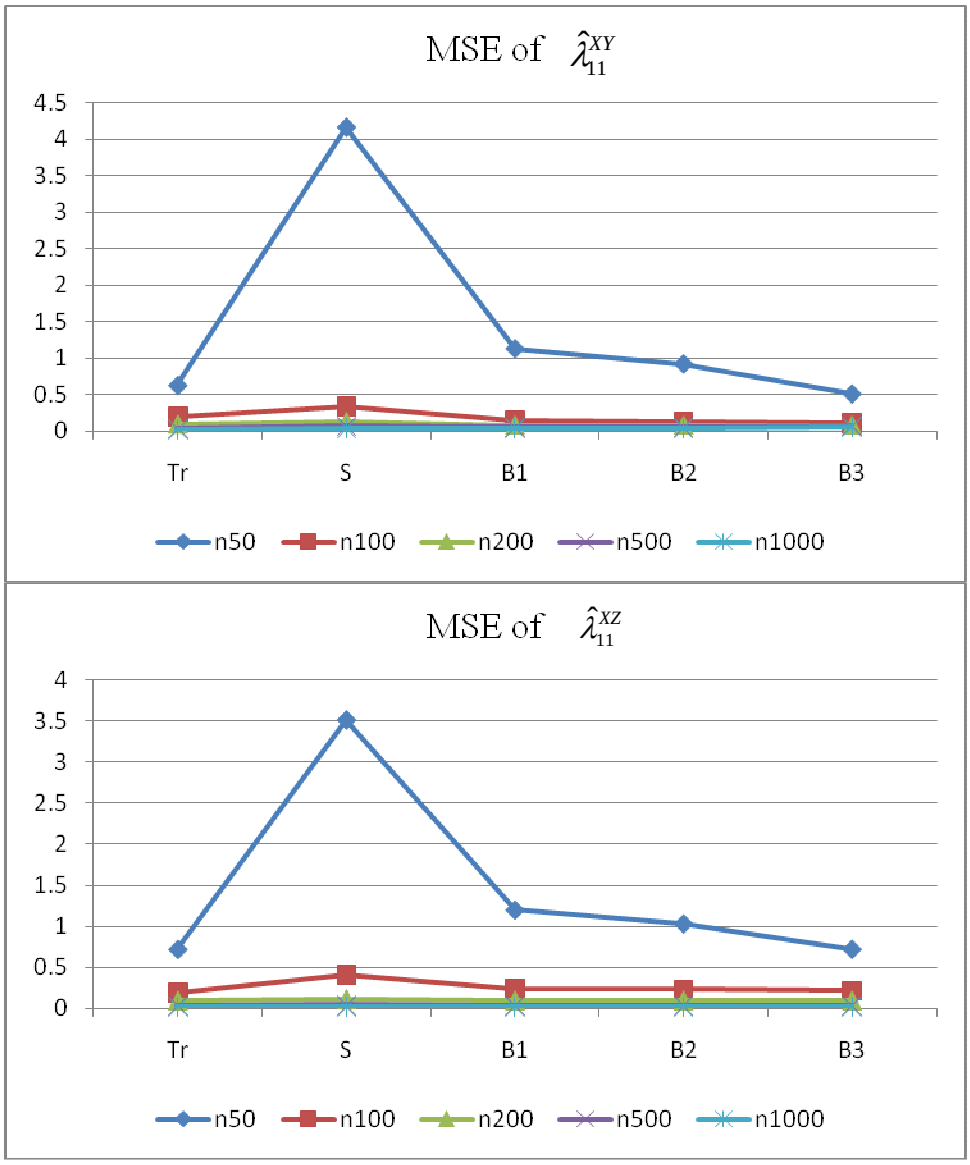

FIGURE 5. MSE of $\hat{\lambda}_{11}^{XY}$ and $\hat{\lambda}_{11}^{XZ}$ in case 1

Figure 6 presents boxplots for the empirical distributions of the estimates for the two parameters of interest, λ_{11}^{XY} and λ_{11}^{XZ}, at the sample size of 500. The labels "Tr", "S", "B1", "B2", "B3" correspond to the estimator under the true model assumption, under model selection, BMA (phi=1), BMA (phi=1.65), BMA (phi=5), respectively. These figures suggest four major findings: (i) The estimator under the true model assumption

outperforms the other four estimators in terms of bias and variability; (ii) Estimates for parameter, λ_{11}^{XY} and λ_{11}^{XZ}, show very different distributional patterns; (iii) For λ_{11}^{XY}, all five estimators show relatively little bias, with the estimator under model selection having the largest variability. The reduction of variability of the three BMA estimators is clearly identified in the first sub-plot; (iv) For λ_{11}^{XZ}, the estimator under model selection shows the largest standard error. Although the BMA estimators reduce the standard error, they show larger bias than the post-model-selection estimator. In general, Bayesian Model Averaging estimators tend to trade bias for variance.

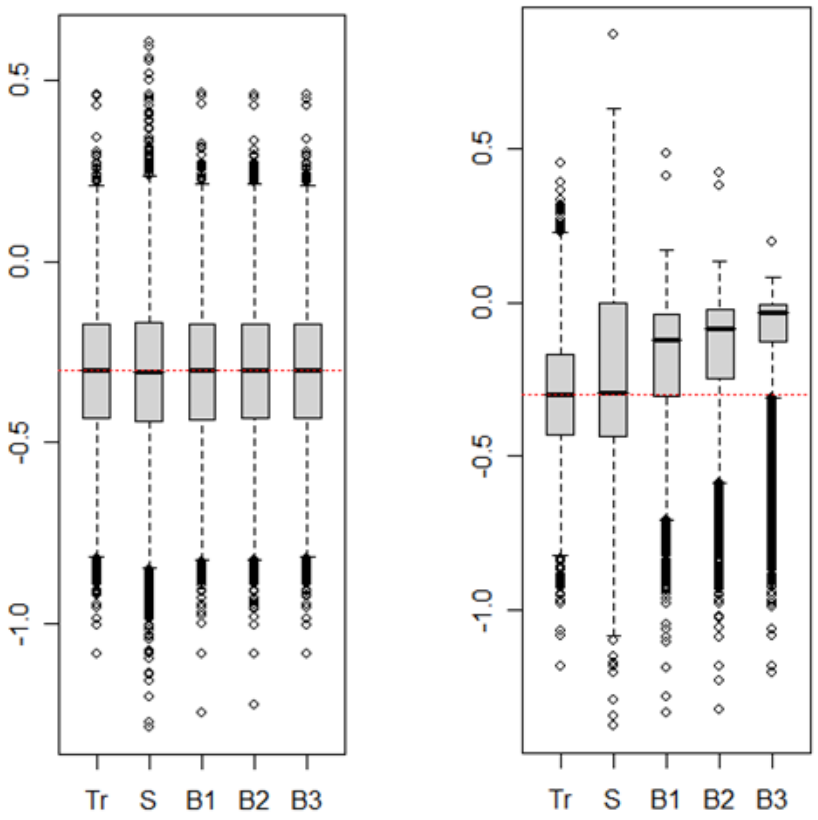

FIGURE 6. Boxplots of the empirical distributions of parameter estimators under true model assumptions (Tr), under model selection (S), BMA(phi=1) (B1), BMA (phi=1.65) (B2), BMA (phi=5) (B3) for λ_{11}^{XY} (plot a) and λ_{11}^{XZ} (plot b) for case 1, n=500.

There are two reasons that the results for $\hat{\lambda}_{11}^{XY}$ and $\hat{\lambda}_{11}^{XZ}$ are quite different. First, model (XY,XZ) and (XY,Z) are contained in the set of candidate models, but not model (XZ,Y). In that sense, $\hat{\lambda}_{11}^{XY}$ and $\hat{\lambda}_{11}^{XZ}$ are not symmetric. Second, the underfitted model (XY, Z) is favored by AIC for small sample sizes. Since the empirical distribution of post-model-selection estimator is a mixture, a large proportion of the mixture comes from the estimate $\hat{\lambda}_{11}^{XZ}$ from model (XY,Z). Therefore, these two parameters, although having the same true value of -.3 are showing different patterns.

4.1.2 Distribution of the Z-statistic

Z-statistics of the form, $(\hat{\lambda} - \lambda_{null}) / \hat{\sigma}_{\lambda}$ were computed for each replication where λ_{null} is the true value and $\hat{\lambda}$ is the estimate. The percentiles of the Z-statistic under model selection and under the true model assumption for case 1 are reported in Table 10. The critical values indicate that strong deviations from the normal distribution are found in the model selection condition while under the true model assumption condition, the Z-statistic distribution is approximately normal. The apparent left skewness of the empirical distribution of $(\hat{\lambda}_{11}^{XZ} - \lambda_{null}) / \hat{\sigma}_{\lambda}$ under model selection condition will tend to lead to erroneous inferences if normality is assumed. Normal-based inference will be conservative if Z-statistic is greater than zero because one will be more likely to accept the null hypothesis than if one had the empirical distribution of the test statistic from Table 10. For instance, consider the scenario of sample size of 1000, two-tail test, and Type I error rate set at .05, a Z-statistic of 1.0 will lead to a decision of “failing to reject the null hypothesis” if one is using the normal distribution, but will lead to a decision of “reject the null hypothesis” if one is using the empirical distribution. If Z-statistic is less

than zero, normal-based inference will be slightly more liberal than if one had the empirical distribution of the test statistic. For instance, consider the scenario of sample size of 1000, two-tail test, and Type I error rate set at .05, a Z-statistic of -2.0 will lead to a decision of "rejecting the null hypothesis" if one uses the normal distribution, but will lead to a decision of "failing to reject the null hypothesis" if one uses the empirical distribution. While under the true model assumption, the empirical distributions of the Z-statistic generally agree with the normal distribution.

TABLE 10
Simulation results: percentiles of the Z-statistic under model-selection for parameter λ_{11}^{XY} and λ_{11}^{XZ} for case 1

n	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
Under model-selection condition									
λ_{11}^{XY}									
50	-2.233	-1.959	-1.657	-1.316	-0.009	1.360	1.727	2.036	2.338
100	-2.386	-1.997	-1.698	-1.339	-0.001	1.350	1.714	2.015	2.374
200	-2.373	-2.024	-1.707	-1.352	-0.014	1.340	1.702	2.051	2.448
500	-2.422	-2.015	-1.720	-1.350	-0.016	1.345	1.724	2.035	2.466
1000	-2.412	-2.034	-1.719	-1.332	0.015	1.362	1.743	2.062	2.443
λ_{11}^{XZ}									
50	-2.206	-1.910	-1.664	-1.307	0.000	0.000	0.403	2.012	2.397
100	-2.296	-2.023	-1.716	-1.335	0.000	0.000	0.000	1.159	2.357
200	-2.388	-2.051	-1.742	-1.377	0.000	0.000	0.000	0.848	2.433
500	-2.410	-2.048	-1.710	-1.340	0.000	0.000	0.092	0.562	1.567
1000	-2.465	-2.051	-1.732	-1.345	0.000	0.573	0.733	0.921	1.727
Under true model assumption									
λ_{11}^{XY}									
50	-2.146	-1.881	-1.627	-1.264	-0.014	1.282	1.660	1.940	2.281
100	-2.295	-1.904	-1.636	-1.307	0.000	1.304	1.658	1.956	2.293
200	-2.275	-1.935	-1.651	-1.282	-0.012	1.274	1.615	1.946	2.318
500	-2.295	-1.921	-1.631	-1.283	-0.015	1.283	1.636	1.935	2.310
1000	-2.262	-1.924	-1.623	-1.262	0.014	1.282	1.672	1.941	2.336
λ_{11}^{XZ}									
50	-2.141	-1.840	-1.623	-1.258	-0.014	1.275	1.613	1.940	2.319
100	-2.239	-1.909	-1.610	-1.270	-0.023	1.265	1.632	1.930	2.281
200	-2.217	-1.939	-1.646	-1.296	0.009	1.301	1.672	1.985	2.341
500	-2.325	-1.954	-1.620	-1.245	0.010	1.304	1.628	1.962	2.320
1000	-2.310	-1.939	-1.626	-1.264	-0.007	1.277	1.636	1.956	2.323
Z-statistic	-2.326	-1.960	-1.645	-1.282	0.000	1.282	1.645	1.960	2.326

The kernel density of the Z-statistic for λ_{11}^{XY} and λ_{11}^{XZ} under model selection condition, under true model assumption and normal densities are plotted in Figure 8. The distribution of Z-statistic for λ_{11}^{XY} is symmetric and approximately normal. The distribution of the Z-statistic for the coefficient λ_{11}^{XZ} is perhaps more interesting, because in the set of the five models listed in Table 3, which the "best fit model" is chosen from, every model includes the term λ_{11}^{XY}. Examination of the percentage of times each model is selected in case 1 reveals that the underspecified model (XY, Z) is selected most of the time for sample sizes at or below 200. The distribution of λ_{11}^{XZ} is essentially a mixture of the estimated λ_{11}^{XZ} from each of the five models and the proportion of each component is the percentage of the time each model is selected as the "best fitting model." In the mixture of the distribution of λ_{11}^{XZ}, the largest component is 0 for sample sizes below 200 in case 1, since the estimated $\hat{\lambda}_{11}^{XZ}$ is 0 in the model (XY,Z).

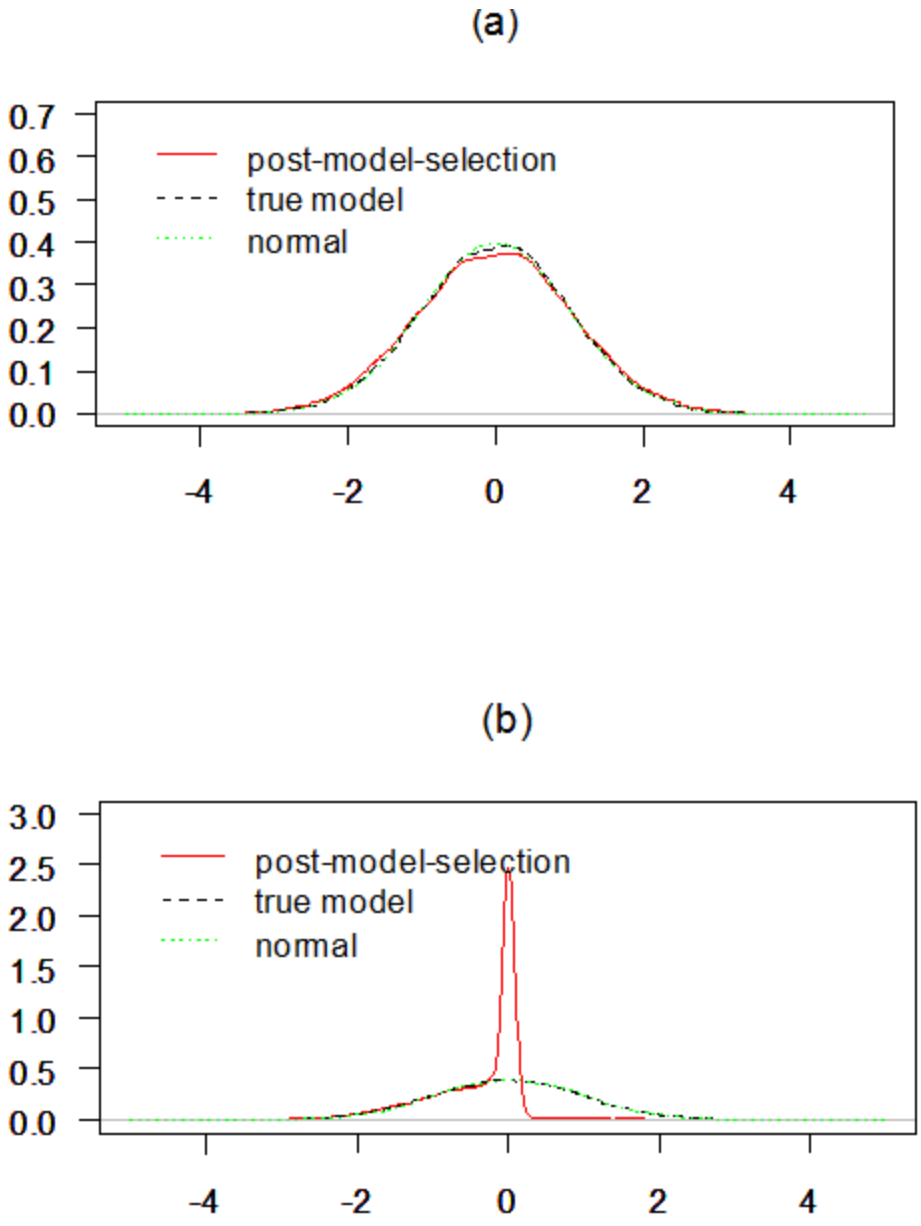

FIGURE 7. Kernel densities for the Z-statistic testing (a) parameter λ_{11}^{XY} and (b) parameter λ_{11}^{XZ} in case 1,n=500.

The QQ-plot is a useful tool to checking if the two data sets come from populations with a common distribution. The empirical distributions of the Z-statistics may be very far from normality, and therefore the corresponding inference of the statistic based on normal distribution may be misleading. In this analysis, QQ-plots serve as a visual aid to determine the model adequacy. Figure 8 shows the QQ-plot of the standard normal distribution versus the distribution of the Z-statistics for testing parameters λ_{11}^{XY} and λ_{11}^{XZ} under the true model assumption and under the model selection condition for case 1 when sample size is 500. For λ_{11}^{XY}, both Z-statistics under true model assumption and under model selection condition both seem to be normal. For λ_{11}^{XZ}, Z-statistics seems normal under true model assumption and is quite different from normal under the model selection condition.

On the whole, the empirical distributions of the Z-statistic deviate in a dramatic fashion from theoretical normal distributions. In particular, the tail percentiles are quite different from those of the normal and this problem does not diminish with large sample size.

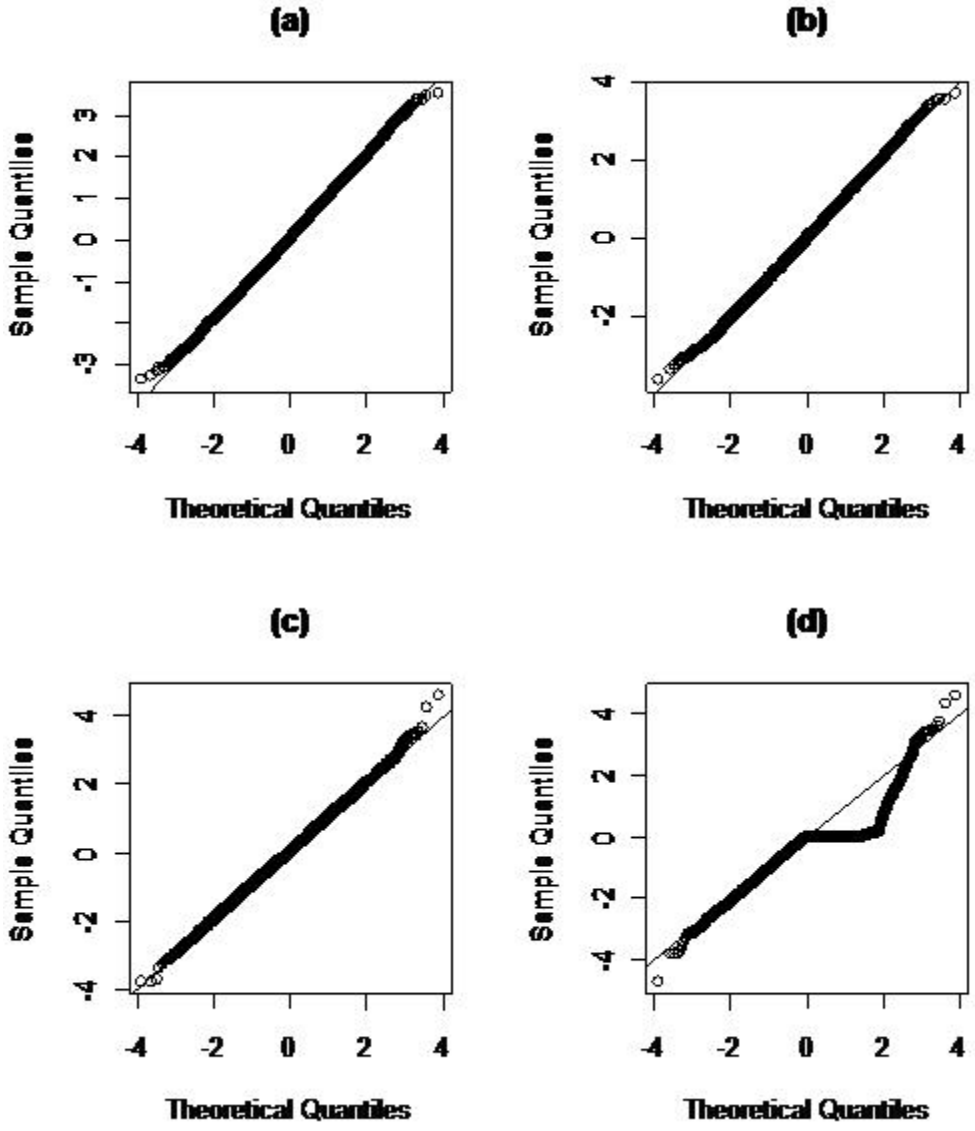

FIGURE 8. QQ-plots of the standard normal distribution versus the distribution of Z-statistics for testing parameters (a) λ_{11}^{XY} under true model assumption; (b) λ_{11}^{XY} under model selection condition; (c) λ_{11}^{XZ} under true model assumption; (d) λ_{11}^{XZ} under model selection condition for case 1, n=500.

4.1.3 Conditional Odds Ratio

Table 11 reports the empirical mean, standard error of the conditional odds ratio estimator under true model assumption, model selection condition, and BMA with phi=1.65. Because of the addition of .5 in both denominator and numerator of the conditional odds ratio estimator, the true value of this estimator cannot be calculated and, consequently, the bias of the estimators cannot be directly computed. There are three major findings: (i) For all the three types of estimators, the standard error goes down as sample size rises. (ii) As expected, $se\ (\hat{\theta}_{TRUE}) < se\ (\hat{\theta}_{BMA}) < se\ (\hat{\theta}_{SELECT})$, with one exception, $\hat{\theta}_{XZ}$ at sample size of 50, where $se(\hat{\theta}_{SELECT}) < se\ (\hat{\theta}_{BMA})$. (iii) The magnitude of $se(\hat{\theta}_{YZ})$ is much smaller than that of $se(\hat{\theta}_{XY})$ and $se(\hat{\theta}_{XZ})$ under the true model assumption, while it is not the same with the other two conditions. For instance, at sample size of 500, $se(\hat{\theta}_{YZ})$ and $se(\hat{\theta}_{XZ})$ under true model assumptions are .002 and .101, respectively, while the $se(\hat{\theta}_{YZ})$ and $se(\hat{\theta}_{XZ})$ under model selection condition are .235 and .155, respectively. One reason might be that under the true model (XY, XZ), there is no λ^{YZ} term. Therefore, the empirical mean of the conditional odds ratio, $E(\hat{\theta}_{YZ})$, is close to 1 under all sample sizes and this conditional odds ratio estimate is quite stable. For example, with $se(\hat{\theta}_{YZ})$, the $se(\hat{\theta}_{SELECT})$ is 118 times $se(\hat{\theta}_{TRUE})$ and 46 times $se(\hat{\theta}_{BMA})$ at sample size of 500. These striking results suggest that the estimate of the conditional odds ratio under post-model-selection condition is far less stable than that under the true model assumption condition, and the inference based on a set of candidate models (BMA) helps to noticeably bring down the variability.

TABLE 11
Empirical means, standard error for the conditional odds ratio estimator under true model assumption, model selection, and BMA (phi=1.65) for n=50, 100, 200, 500 and 1000: case 1.

	TRUE MODEL		MODEL SELECTION		BMA (phi=1.65)	
$\hat{\theta}_{XY}$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$
n=50	0.900	0.533	1.049	1.657	1.005	1.391
n=100	0.826	0.345	0.867	0.603	0.824	0.382
n=200	0.782	0.234	0.801	0.344	0.782	0.254
n=500	0.757	0.143	0.765	0.188	0.758	0.147
n=1000	0.751	0.101	0.754	0.131	0.748	0.103
$\hat{\theta}_{XZ}$						
n=50	0.908	0.542	1.108	1.651	1.050	2.361
n=100	0.825	0.344	0.931	0.560	0.907	0.300
n=200	0.783	0.235	0.870	0.311	0.888	0.196
n=500	0.760	0.147	0.815	0.212	0.867	0.147
n=1000	0.749	0.101	0.774	0.155	0.833	0.138
$\hat{\theta}_{YZ}$						
n=50	1.057	0.121	1.396	2.469	1.307	2.314
n=100	1.024	0.026	1.150	0.972	1.062	0.538
n=200	1.011	0.008	1.068	0.432	1.025	0.205
n=500	1.004	0.002	1.030	0.235	1.010	0.091
n=1000	1.002	7E-04	1.013	0.159	1.006	0.060

4.2 General Results

4.2.1 Parameter Estimate, Bias, Standard Error, MSE, and Relative Efficiency

The percentages of times that each model is selected by AIC as the "best-fit" model for case 1-case 8 are displayed in Figures 9 and 10. For the product-multinomial sampling design (case 1-case 4), the $2 \times 2 \times 2$ table and the $2 \times 2 \times 3$ table show similar patterns. In particular, in the small-valued parameter setting, case 1 and case 3 both show poor selection accuracy when sample size is small or moderate (at or below 500) but have improved selection accuracy (about 69%) when sample size is as large as 1000. At small or moderate sample sizes, the underfitted model (XY, Z) is selected most often as the "best fit" model. In large-valued parameter setting, case 2 and case 4 both have satisfactory (above 70%) selection accuracy for all sample sizes. For the multinomial sampling design (case5-case8), selection accuracy is low (below 50%) even when sample size is as large as 1000 for small-valued parameter scenario but selection accuracy is satisfactory (about 80%) when sample size is large as 1000 for the big-valued parameter scenario.

The bias, standard error, and mean-square error are presented in Tables B1-B37. Figure 11—17 display the box-plots of distributions of the parameters. The results in cases 2-8 confirm the observation made in case 1, that if inference is made conditional on the best model chosen, the bias, the standard error and MSE of the estimator are all greater than those of the estimator under the model *a priori* assumption.

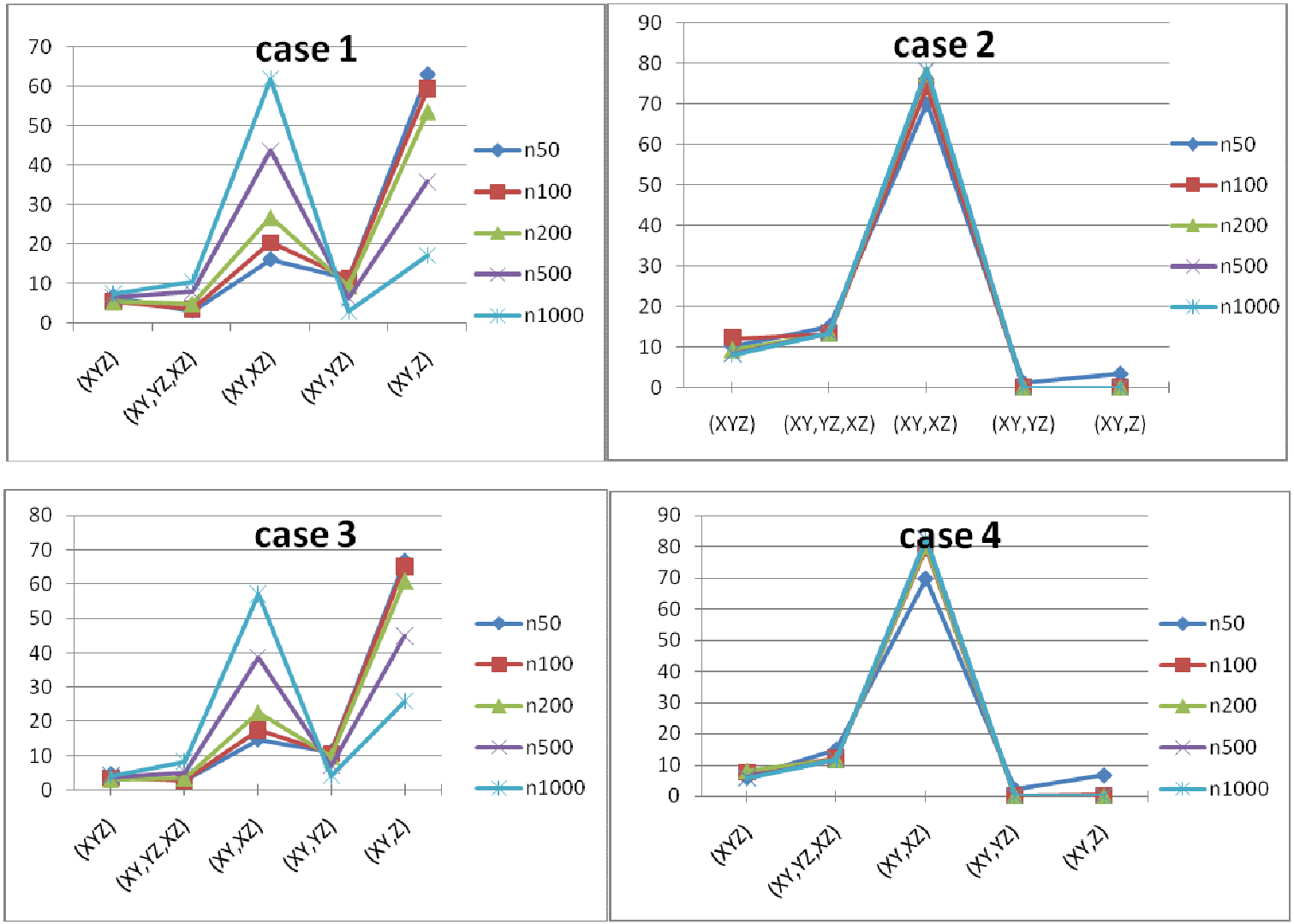

FIGURE 9. Percentage of each model being selected (case 1-case 4)

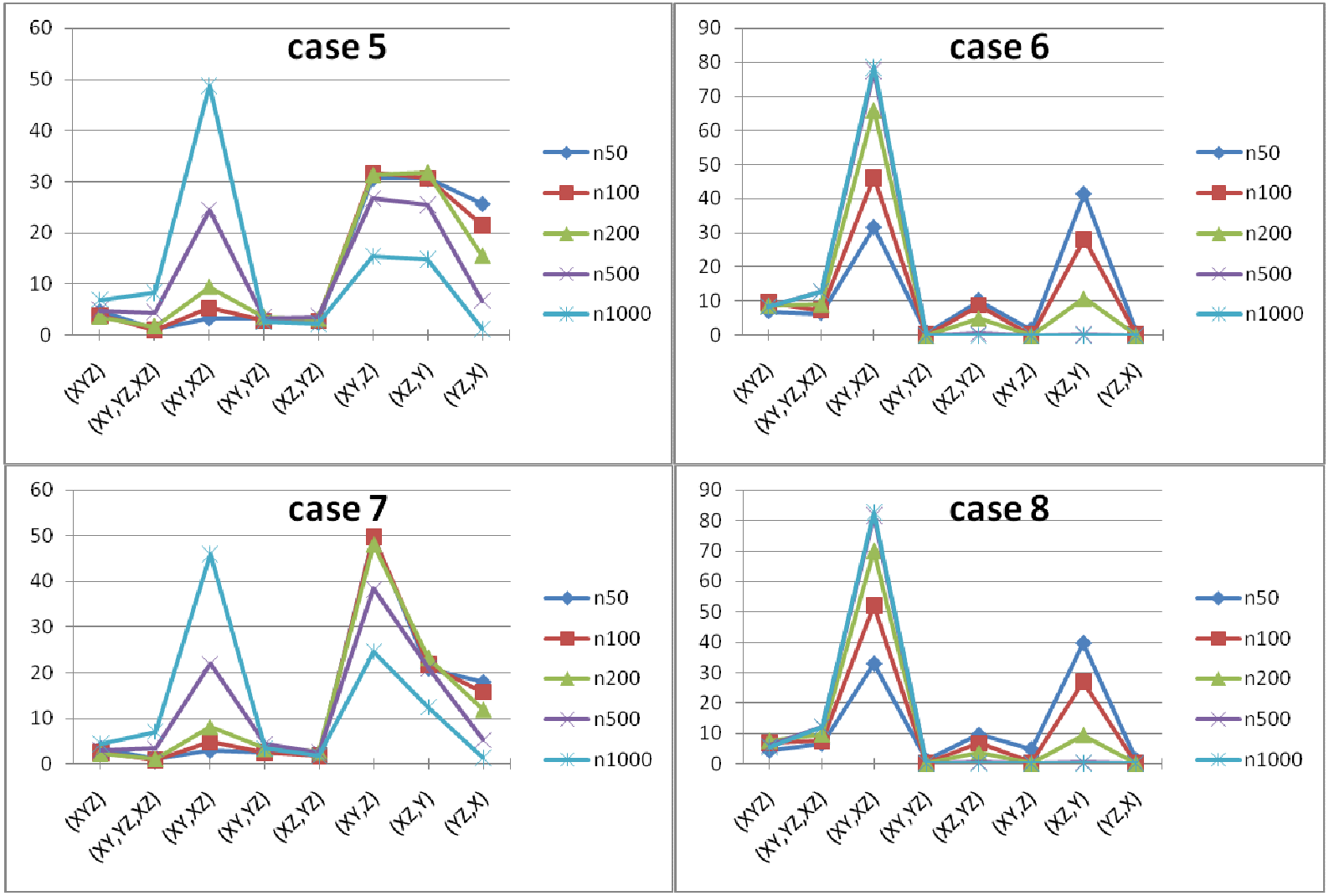

FIGURE 10. Percentage of each model being selected (case 5--case 8)

There are six major findings: (1) As for case 1, when the sample size increases, the bias, standard error and MSE becomes smaller for each of the five estimators; (2) The bias of the ML estimator under true model assumption essentially disappears when sample size is fairly large, whereas the bias of the post-model-selection estimator persists even at sample size 60 observations per cell; (3) The standard error of the post-model-selection estimator can become very large at moderate sample size of 16 observations per cell. This usually happens when the parameter is large valued. For example in case 4 (large parameter value, five models in the model set, $2 \times 2 \times 3$ table), at sample size 200, the values of $se(\hat{\lambda}_{11}^{XZ})$ are .429, 4.25, 2.011 for estimator under true model assumption, under model selection, and by the BMA[1] , respectively; (4) When the parameter is large valued, the contrast of the bias, standard error, and MSE of the post-model-selection estimator against estimator under true model assumption is more prominent than the scenario when the parameter is small value. For example, with case 2, the bias, standard error, and MSE of the post-model-selection estimator are -.478, 3.224, and 10.621, respectively, and its relative efficiency versus the estimator under the true model assumption is 84. When other factors remains the same, with the small value parameter scenario (for example, case 1), the estimates are.08 (bias), .358 (s.e.) , .134 (MSE), and 1.411(Rel EFF). On reflection, this is to be expected, as larger valued estimates tend to have larger variability. (5) BMA point estimates have MSE lower than standard variable selection methods, however, it might introduce larger bias in the estimator under some conditions. (6) Let $\boldsymbol{M_1}$ denote the model set under the product multinomial sampling scheme and $\boldsymbol{M_2}$ denote the model set under multinomial sampling scheme. Then $\boldsymbol{M_1}$ contains 8 models, $\boldsymbol{M_2}$ contains 5 models and $\boldsymbol{M_2} \subset \boldsymbol{M_1}$. Case 5 – case 8 are parallel to

case 1 – case 4, with the former selected over larger candidate model set. The percentage of each model being selected is more spread out in case 5 – case 8, as two misspecified models (XZ, Y), (YZ, X) and one underspecified model (XY, Z) are included. The comparison of these two categories shows intractable results. The interests here focus on the comparison of BMA results. The purpose is to see whether BMA, when performed over a larger set of models (including the smaller set), lead to bigger variability of the estimates, as suggested by the BMA variance formula (14). The results do not give a definitive answer to the question, since different scenarios point to different answers, and there seems to be no rules governing each result. However, the difference between them is of a small magnitude.

4.2.2 Distribution of the Z-statistic

Similar to the result in case 1, the Z-statistic of the parameters under model selection condition usually do not follow a normal distribution while the Z-statistic under the true model assumption always do, as is shown by the Q-Q plots in Figures B1-B7. To quantify the deviation from normality, Kolmogorov-Smirnov tests are performed on the empirical distributions of the Z-statistics of the parameters λ_{11}^{XY}, λ_{11}^{XZ} and λ_{12}^{XZ} under model selection and under true model assumption to detect the deviation from normality. Table B28 – B30 show the p-value of the Kolmogorov-Smirnov test statistic. Since the sample size in these tests are as large as 10000, some of the p-value under true model assumption conditions is also very small (below .05), but on the whole, they are bigger than the p-value under model selection conditions.

4.2.3 Conditional Odds Ratio

The Empirical means, standard error for the conditional odds ratio estimator under the true model assumption, model selection and BMA (phi=1.65) for different sample sizes are presented in Table B31-B37. The results of case 2-case 8 are similar to that of case 1. The standard error of the post-model-selection estimator always has larger standard error than that of the estimator under true model assumption. Their contrast is more prominent in the conditional odds ratio $\hat{\theta}_{YZ}$ rather than $\hat{\theta}_{XY}$ and $\hat{\theta}_{XZ}$. This is because the true model is (XY, XZ) in this simulation study, indicating that there is no association between factor Y and Z. Therefore, we expect the conditional odds ratio to be close to 1 with small variability under the true model assumption. Selecting a model other than the true model produces a quite different conditional odds ratio, since the odds ratio ($\hat{\theta}$) is based on the ratio of the expected cell mean, which is on the exponential scale of the of the parameter estimates. Hence, small bias in the λ's will produce bigger deviations in the μ's and $\hat{\theta}$'s.

What is informative of the result is that almost in all cases, in small sample size such as 50 and 100, conditional odds ratio based on BMA parameter estimators has bigger standard error than post-model-selection estimators. This suggests that using BMA in small samples cannot not guarantee smaller variability of the conditional odds ratio estimator than that under model selection. When sample size is greater than 200, the advantage of BMA is secured.

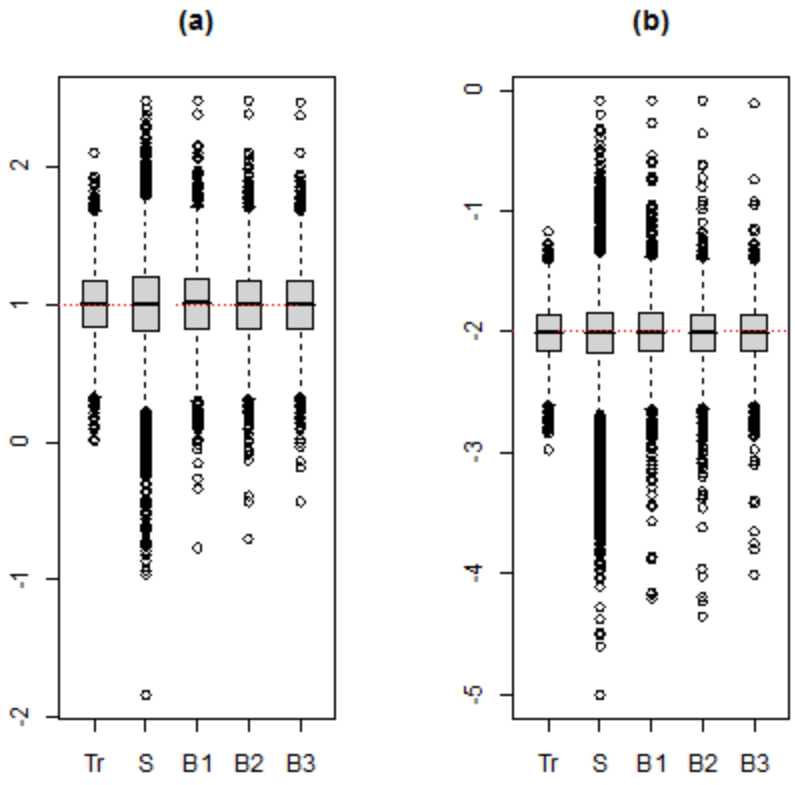

FIGURE 11. Boxplots of the empirical distributions of parameter estimators under true model assumptions (Tr), under model selection (S), BMA(phi=1) (B1), BMA (phi=1.65) (B2), BMA (phi=5) (B3) for λ_{11}^{XY} (plot a) and λ_{11}^{XZ} (plot b) for case 2, n=500.

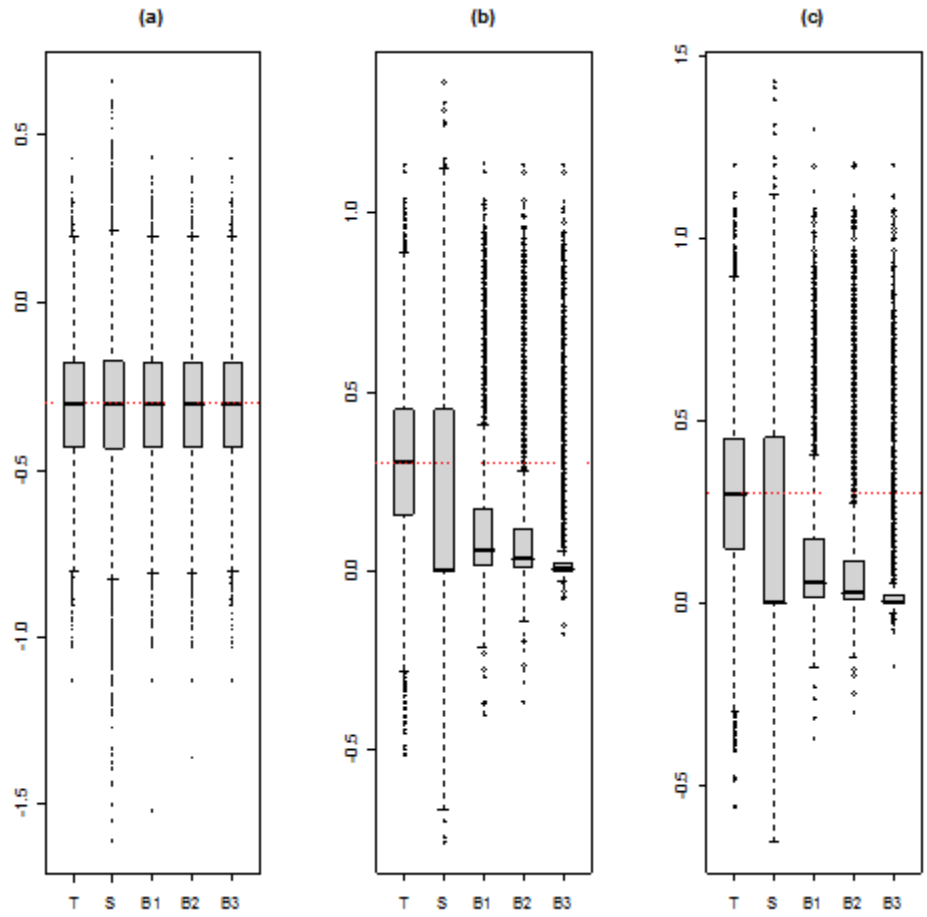

FIGURE 12. Boxplots of the empirical distributions of parameter estimators under true model assumptions (Tr), under model selection (S), BMA(phi=1) (B1), BMA (phi=1.65) (B2), BMA (phi=5) (B3) for λ_{11}^{XY} (plot a), λ_{11}^{XZ} (plot b) and λ_{12}^{XZ} (plot c) for case 3, n=500.

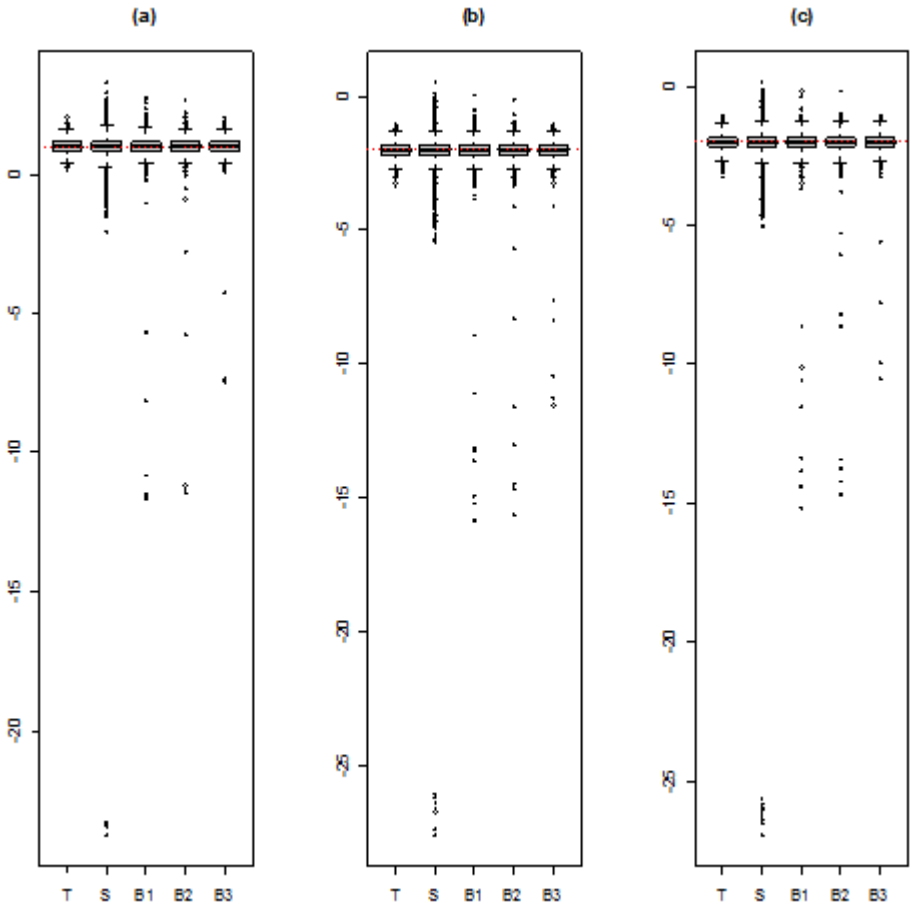

FIGURE 13. Boxplots of the empirical distributions of parameter estimators under true model assumptions (Tr), under model selection (S), BMA(phi=1) (B1), BMA (phi=1.65) (B2), BMA (phi=5) (B3) for λ_{11}^{XY} (plot a), λ_{11}^{XZ} (plot b) and λ_{12}^{XZ} (plot c) for case 4, n=500.

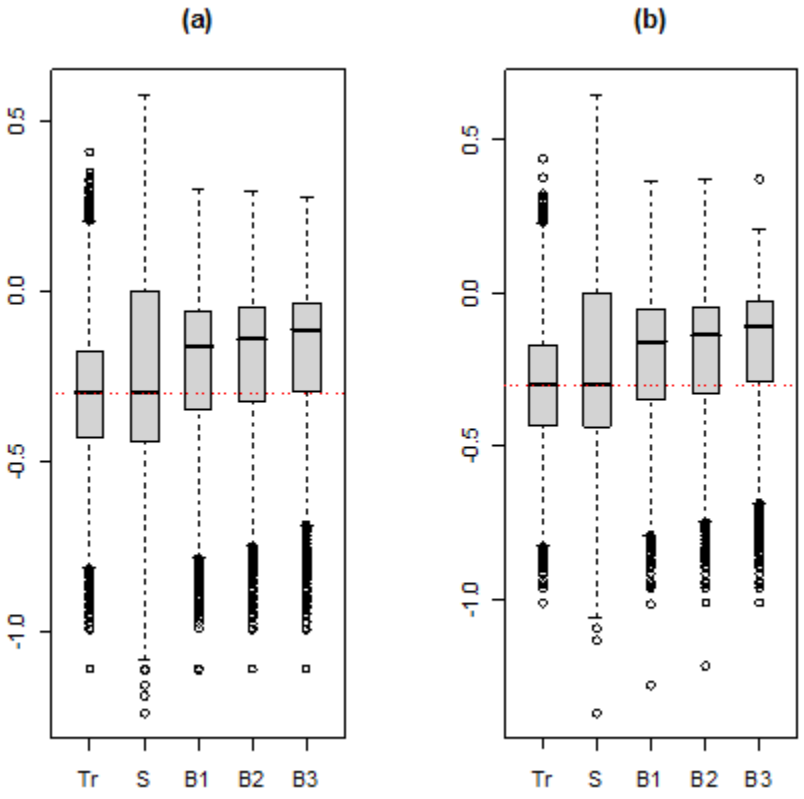

FIGURE 14. Boxplots of the empirical distributions of parameter estimators under true model assumptions (Tr), under model selection (S), BMA(phi=1) (B1), BMA (phi=1.65) (B2), BMA (phi=5) (B3) for λ_{11}^{XY} (plot a) and λ_{11}^{XZ} (plot b) for case 5, n=500.

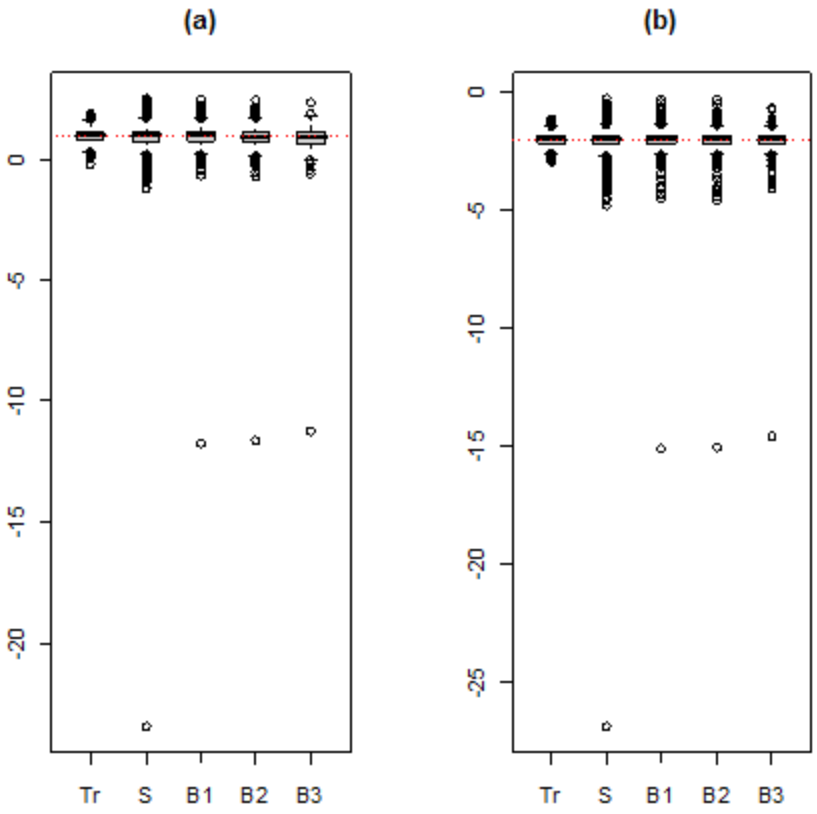

FIGURE 15. Boxplots of the empirical distributions of parameter estimators under true model assumptions (Tr), under model selection (S), BMA(phi=1) (B1), BMA (phi=1.65) (B2), BMA (phi=5) (B3) for λ_{11}^{XY} (plot a) and λ_{11}^{XZ} (plot b) for case 6, n=500.

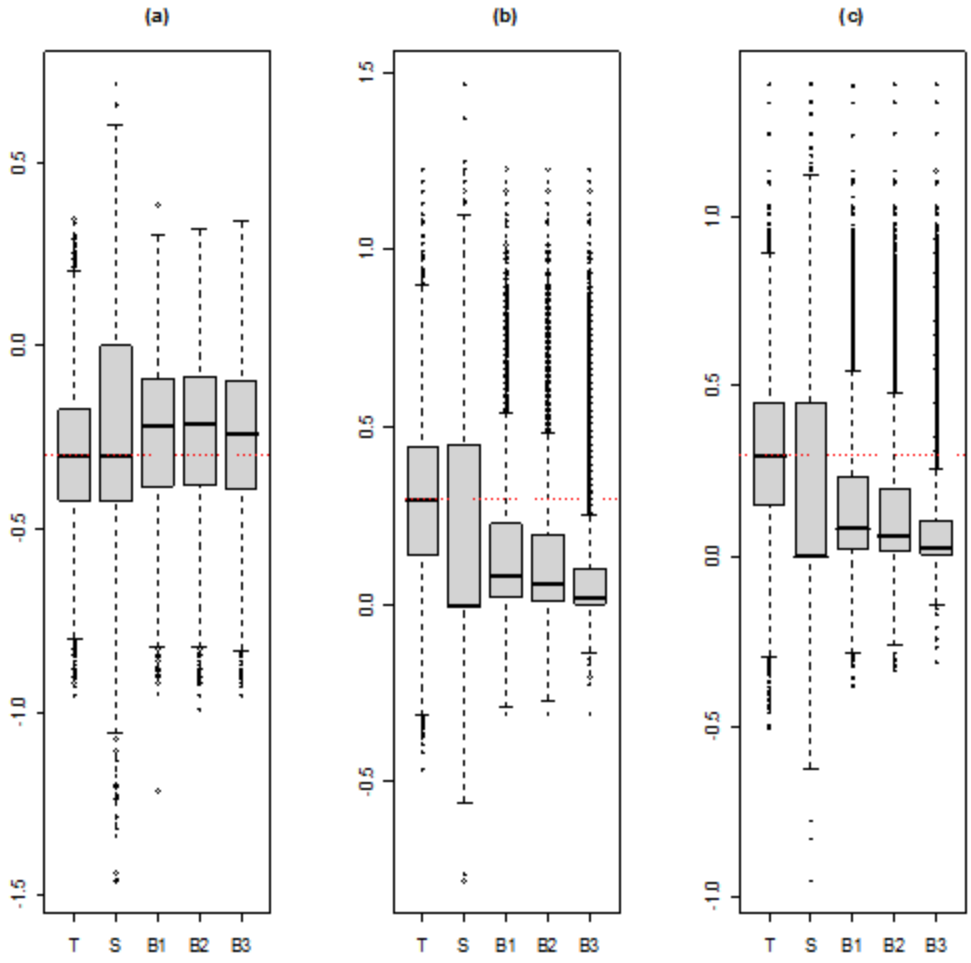

FIGURE 16. Boxplots of the empirical distributions of parameter estimators under true model assumptions (Tr), under model selection (S), BMA(phi=1) (B1), BMA (phi=1.65) (B2), BMA (phi=5) (B3) for λ_{11}^{XY} (plot a), λ_{11}^{XZ} (plot b) and λ_{12}^{XZ} (plot c) for case 7, n=500.

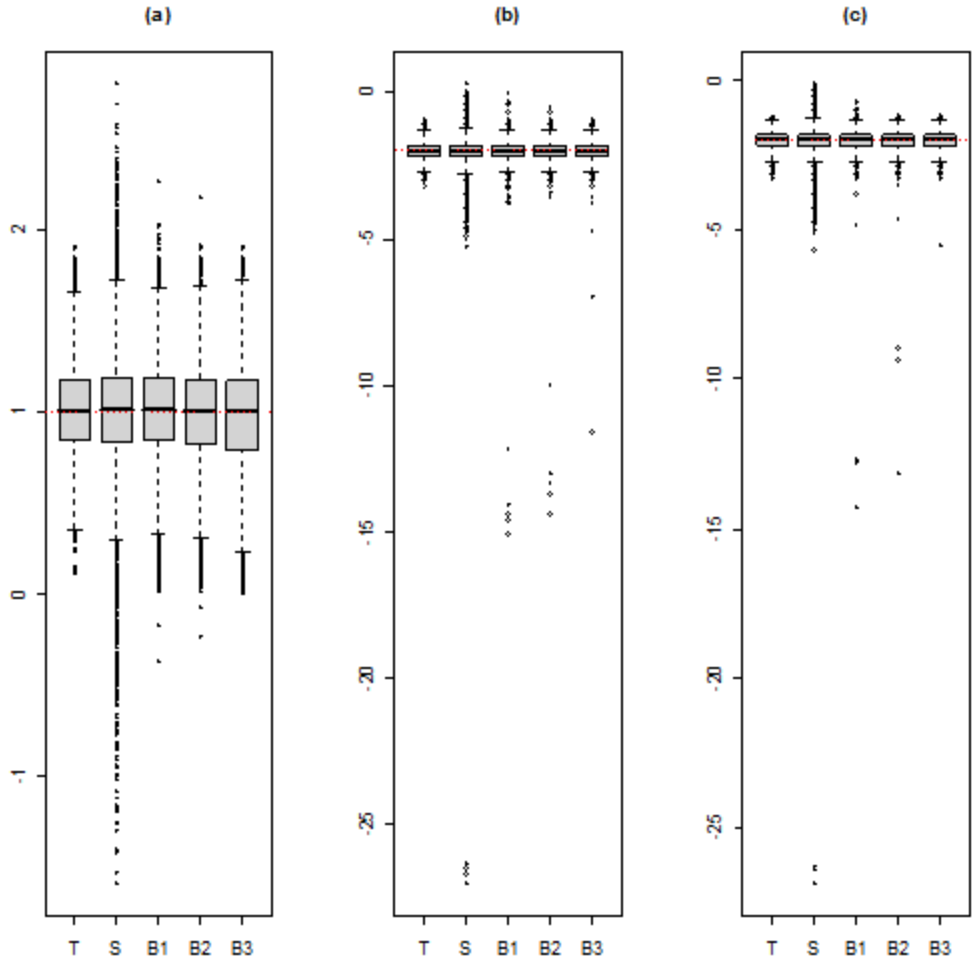

FIGURE 17. Boxplots of the empirical distributions of parameter estimators under true model assumptions (Tr), under model selection (S), BMA(phi=1) (B1), BMA (phi=1.65) (B2), BMA (phi=5) (B3) for λ_{11}^{XY} (plot a), λ_{11}^{XZ} (plot b) and λ_{12}^{XZ} (plot c) for case 8, n=500.

Chapter 5: Conclusion and Discussion

It is common practice for researchers in the social sciences and education to use a model selection technique to find a best fitting model and, then, to carry out inference as if this model were given *a priori*. This study examined the effect of model selection, variable selection in this context, on the inference of the log-linear model. The purpose was to (i) find out the consequences when the behavior of model selection is ignored; (ii) investigate the performance of the estimator provided by the Bayesian model averaging method, and evaluate usefulness of the multi-model inference as opposed to the single model inference.

The basic finding of this study were that inference based on a single "best fit" model chosen from a set of candidate models tends to underestimate the variability of the parameters and induce additional bias in estimation for a loglinear model. The results of the simulation study showed that the post-model-selection parameter estimator has larger bias, standard error, and mean square error than the estimator under the true model assumption due to model uncertainty. The same results applied to the conditional odds-ratio estimators. The fundamental reason is that the sampling distribution of the post-model-selection estimator is actually a mixture of distributions from a set of candidate models. The variability of the post-model-selection estimator has a large component from selection bias. While these problems are alleviated with the increase of sample size, the interpretation of the p-value of the Z-statistic of the parameters is erroneous even when sample size is quite large.

To avoid the problem of the inference based a single best model, Bayesian model averaging adopts a multi-model inference method, treating the weighted mean of the

estimates from each model in the set as the point estimator, where the weights are derived using Bayes theorem. In this thesis, real and simulated datasets were used to illustrate this method and results were compared with the single "best fit" model inference. The BMA method requires specification of prior probabilities for models and prior densities for the parameters. Under the uniform prior probability for models and normal distributions for the parameters, different specifications of the normal distribution lead to somewhat different results in terms of MSE of the parameter estimates in the simulation study. With large sample size, it was true that different and reasonable choices of prior distributions had minor effects on posterior inferences. Generally speaking, the simulation results confirm the efficacy of the BMA method as compared with data-driven single "best-fit" model inference.

The distribution of the post-model-selection estimator has proved annoyingly intractable because the type of models and the specific set of models vary from research to research. The known theoretical results fall short of what we would like to know for practical applications. Based on the current findings, several recommendations can be given: (i) Bayesian model averaging is a better alternative to the inference based on a single best model since it has smaller MSE, although additional bias is introduced by this method. Usually, multivariate normal priors are used for the parameters in the Bayesian model averaging method. Different specification of the multivariate normal distribution parameters leads to different results but when the sample size is relatively large (500 or larger, say), these differences are small. (ii) If model selection is vital in some research setting (for example comparing theories represented by different models), large sample size is needed. The question to ask is how large is considered "large?" Based on the

results of this simulation study, in the loglinear model setting, 60 observations per cell helps to keep the magnitude of the relative efficiency of the post-model-selection estimator versus the estimator under true model assumption at a relatively low level (below 10). (iii) The p-value of the post-model-selection Z-statistic is misleading, even under large sample size, and, therefore, its use is not recommended. Using the normal curve theory p-values will lead to either conservative or liberal conclusions depending on the shape of the sampling distribution of the Z-statistics. Usually, in the loglinear model cases studied here, some post-model-selection Z-statistics were leptokurtic, having a higher peak than normal distribution, i.e., more scores fall at 0 due to the fact some underestimated models are chosen.

Based on the current research, some implications for the future research are : (i) AIC has been criticized for not being dimension-consistent, in other words, as sample size grows, the probability of selecting the “true” model does not go to 1, while BIC is dimensional-consistent. It would be valuable to evaluate the performance of the post-model-selection estimator using the BIC as model selection criterion.; (ii) Due to the scope of this research, only the point estimators were investigated. In future studies, it would be beneficial to investigate the accuracy of the BMA variance estimators. Also, the coverage properties of the confidence intervals might be of interest in the future study.

As a final note, although the multi-model inference method (BMA) outperformed post-model-selection estimators using measures such as estimated mean square error, the interpretation of the estimates using the BMA method should proceed with caution. It is because meaning and interpretation of a coefficient might change over models. The interpretation of a parameter pertains to the particular model, even if the same symbol is

used to designate the parameter. If BMA results are used to explain a specific coefficient, one should make sure that the interpretation of the coefficient should be the same over the set of competing models. In summary, BMA gives more realistic estimates of model uncertainty and provides a structured way to deal with the model selection dilemma. However, model averaging method does not provide model parsimony, since it averages models across different dimensions. We can only interpret the parameters as from the saturated model with the average coefficients, which makes it difficult to answer such research questions as “among the three factors of marijuana, cigarette, and alcohol, what types of independence relationship can be inferred based on the dataset” in a drug data context.

Appendix A: Application of Bayesian Model Averaging to Real Data

In this section we applied the Bayesian model averaging method to some real data, which provided some insight into the performance of the method. The results reported in this section were analyzed using the software package, BMA, in R. First, we described the drug data. Second, we investigated the issue of sensitivity by changing the prior distributions on the posterior results. Third, a comparison of BMA results and classical results were presented. Fourth, we utilized the principle of "Occam's razor" to reduce the number of models in the model set when the number of model under consideration is huge. Finally, we compared the results of two model averaging methods, namely, Bayesian model averaging and frequentist model averaging. In the frequentist model averaging, two methods were used to generate the weights: the bootstrap method, and the AIC method.

Drug Data

We illustrate the use of BMA in loglinear analysis of contingency table with the student drug data in Agresti (2002). The data were from a 1992 survey by the Wright State University and the United Health Services. The Survey asked 2276 senior high school students in Dayton, Ohio whether they had ever used alcohol, cigarettes, or marijuana. The respondents were cross-classified by alcohol use (A), cigarette use (C) , and marijuana use (M).

Agresti gave an example of the model selection process in loglinear model building. He investigated nine models (all possible) for the three-way contingency table, namely, (A,C,M), (M, AC), (A,CM), (C, AM), (AM,CM),(AC, AM), (AC,CM), (AC,AM,CM),(ACM). He suggested AIC could be used in this type of model

comparison. Judging by minimum AIC, the all-two-factor-interaction model (AC, AM, CM) fits the data best. He also gave advice on the use of the likelihood ratio tests: "With large sample sizes, statistically significant effects can be weak and unimportant. A more relevant concern is whether the associations are strong enough to be important. Confidence intervals are more useful than tests for assessing this" (p.325). The drug data are summarized in Table A1.

TABLE A1
Drug data

Marijuana(M)	Cigarette (C)	Alcohol(A)	count
Yes	Yes	Yes	911
No	Yes	Yes	538
Yes	No	Yes	44
No	No	Yes	456
Yes	Yes	No	3
No	Yes	No	43
Yes	No	No	2
No	No	No	279

The log-linear model (AMC) can be written in matrix form as

$$\begin{bmatrix} log\ \mu_{111} \\ log\ \mu_{121} \\ log\ \mu_{211} \\ log\ \mu_{221} \\ log\ \mu_{112} \\ log\ \mu_{122} \\ log\ \mu_{212} \\ log\ \mu_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \lambda_1^C \\ \lambda_1^M \\ \lambda_1^A \\ \lambda_{11}^{CM} \\ \lambda_{11}^{AC} \\ \lambda_{11}^{AM} \\ \lambda_{111}^{AMC} \end{bmatrix}.$$

In simple form, a log-linear model is expressed as

$$\log \boldsymbol{\mu} = \mathbf{X}\,\boldsymbol{\beta}$$

where $\boldsymbol{\mu}$ is a $q \times 1$ vector of expected counts (q is the number of cells), $\mathbf{X}$ is a $q \times p$ design matrix with known values (p is the number of parameters including the

intercept and often $\mathbf{X}$ consists of 1s and 0s), and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters.

An Investigation of Sensitivity

In BMA, the conclusion is drawn based on the posterior probabilities of each model in the model set. In most applications of BMA, the standard practice of setting up priors is to use the flat prior model probability, which is the default in the software. These reference priors are chosen by "public agreement," much like units of length and weight. The analysts fall back to the default when there is insufficient information. However, the robustness of the BMA results under different prior specification has seldom been inspected. In this section, we use the change in posterior model probability and parameter probability as a sensitivity measure for BMA. Two types of priors will be investigated: the prior model probabilities and the prior parameter distributions.

In BMA, the choice of prior is to a certain degree arbitrary, which poses a difficulty for applied researchers. However, advocates of BMA tend to view specification of a prior as an important but controllable technical complexity. They are attracted to the overall logical consistency and its role as a formal way to solve the model uncertainty problem.

In the BMA sensitivity analysis of drug data, the prior probabilities were constructed based on substantive knowledge of substance uses. In this sense, the prior is viewed as a unique representation of our ignorance. Thus, we devote a section to a description of our knowledge of substance uses. We begin with a brief review of the literature on substance uses, then move on to how this body of knowledge assists us in

choosing subjective priors. Our aim is to find out whether the weights in model averaging are sensitive to the choice of specification of the prior probability. Or, in other words, we try to find out whether prior settings have a dramatic impact on the value of the Bayes factor, and subsequently, on the posterior inference resulting from BMA. The investigation is conducted using the glib function for generalized linear models in the R software package BMA. The R glib function obtains the posterior model probability via Bayes factors.

In the United States, alcohol and cigarette are licit drugs while marijuana is, an illegal drug. Drug use, on the whole, is a complex social phenomenon influenced by many interacting factors (Rob, Reynolds, & Finlayson,1990). Several competing theories of drug uses can be represented by different loglinear models. Three major theories were explored.

The problem-behavior theory, one of the most cited and influential theories on drug use, stated that the associations between drug use are due to a common etiology caused by similar psychological and environmental factors from a variety of domains including biological, genetic, social and behavioral (Donovan & Jessor, 1985). If the same factors are responsible for the associations between these drugs, measures used to asses these variables are psychometrically comparable, then one would expect the associations to be similar between the licit drugs and marijuana use. The loglinear representing this theory is (AC, AM, CM) or (ACM). These models posit that the use of these substances is a symptom of a larger set of destructive and deviant behaviors and thus they are all related.

In drug use literature, “gateway drugs” refer to alcohol and cigarette. Gateway theory suggested that individuals usually started drug use with alcohol and/cigarettes and then progressed to marijuana and other illicit drugs (Kandel & Faust, 1975; Yamaguchi & Kandel, 1984). The loglinear model representing this theory is (AM, CM).

A third theory, representing by the model (CM, A), is the lifestyle theory. Coffey, Carlin, Lynskey, Li, and Patton (2003) found that persistent, frequent alcohol use during the teen years negated the risk for developing marijuana dependence in regular marijuana users as young adults, whereas persistent cigarette use is a strong predictor of marijuana dependence, possibly due to similar modes of ingestion of cigarettes and marijuana. They hypothesized that their findings may illustrate a social process whereby individuals either become part of a predominantly alcohol-using or marijuana-using lifestyle. In model (CM, A), cigarette use and marijuana use are associated, but alcohol use are not associated with marijuana use.

Based on the above substantive knowledge, four sets of reference priors chosen in the sensitivity analysis were shown in table A2. The weights for set 1-3 were arranged such that each model in the sequence gets a prior weight 30% more than the previous one. This type of priors was proposed by Bartels (1997). The weight vector (1,1.3,1.3^2 1.3^3, 1.3^4, 1.3^5, 1.3^6, 1.3^7, 1.3^8), when normalized, became the prior probability vector (.031,.041,.053,.069,.089,.116,.151,.196,.255). The prior probabilities assigned reflected the Bayesian interpretation of probability as a measure of uncertainty. This construction of the priors provided a good approximation to the subjective knowledge on the drug use issue. Set four adopted uniform model priors. Its weight vector (1,1,1,1,1,1,1,1,1), when normalized, became the prior probability vector

(.111,.111,.111,.111,.111,.111,.111,.111,.111). The prior probabilities were graphically presented in Figure A1.

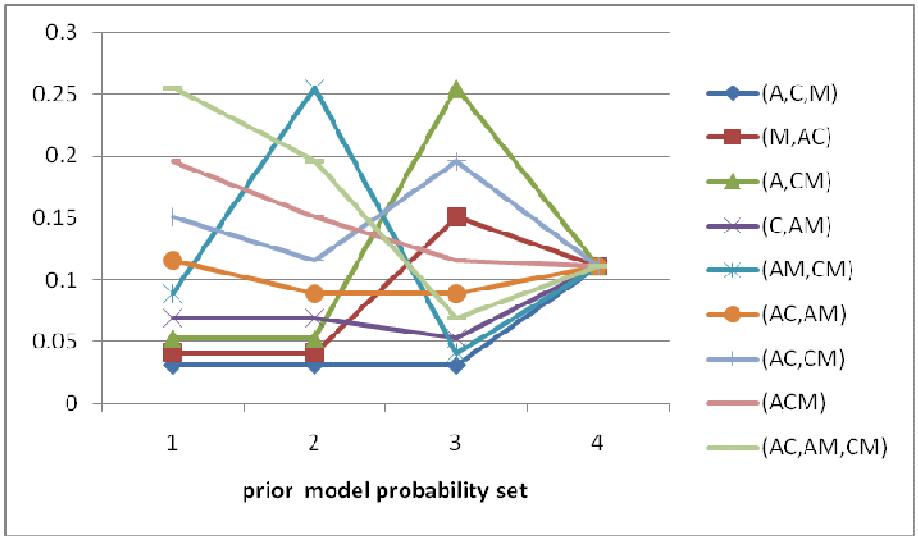

FIGURE A1. Prior model probabilities.

Table A2 presented the models arranged based on the theories. Since the problem-behavior theory suggested the most likely models are (ACM) or (AC,AM,CM), these two models were assigned the highest prior model weights, $(1.3)^7$ and $(1.3)^8$, respectively. The less association terms a model had, the smaller prior model probability was assigned. The gateway theory supported that the two interaction terms AM and CM be included in the model, therefore, three models containing these two terms, i.e., (ACM), (AM,CM, AC), and (AM,CM), were given relatively higher prior model weights, $(1.3)^6$, $(1.3)^7$ and $(1.3)^8$, respectively. The lifestyle theory suggested that only cigarette use was related to marijuana use and alcohol use was not associated with marijuana use. Thus, five models containing the terms AM, i.e., (AM,CM), (C,AM), (AC,AM,CM), (AC,AM) and (ACM) were given relatively less weights, $(1.3)^1$, $(1.3)^2$, $(1.3)^3$, $(1.3)^4$, and $(1.3)^5$. The model that posited that the three substance uses were mutually independent, i.e., (A, C, M) was given lowest prior weight in set 1-3, since all

three theories stated that at least some level of associations existed among the three types of substance uses. The setup represented prior opinions in an easily elicitable form.

TABLE A2
Four sets of reference priors in the sensitivity analysis

Prior model weight based on substantive knowledge				Uninformative priors
Prior model weight	Set 1: based on problem-behavior theory	Set 2: based on gateway theory	Set 3: based on lifestyle theory	Set 4: Uniform prior model weight
1	(A,C, M)	(A,C, M)	(A,C, M)	1
$(1.3)^1$	(M,AC)	(M,AC)	(AM,CM)	1
$(1.3)^2$	(A,CM)	(A,CM)	(C,AM)	1
$(1.3)^3$	(C,AM)	(C,AM)	(AC,AM,CM)	1
$(1.3)^4$	(AM,CM)	(AC,AM)	(AC,AM)	1
$(1.3)^5$	(AC,AM)	(AC,CM)	(ACM)	1
$(1.3)^6$	(AC,CM)	(ACM)	(M,AC)	1
$(1.3)^7$	(ACM)	(AC,AM,CM)	(AC,CM)	1
$(1.3)^8$	(AC,AM,CM)	(AM,CM)	(A,CM)	1

TABLE A3
Posterior probabilities with four sets of prior model weights (phi=1.65)

Models	Probabilities[1]	Probabilities[2]	Probabilities[3]	Probabilities[4]
(A,C, M)	0	0	0	0
(M,AC)	0	0	0	0
(A,CM)	0	0	0	0
(C,AM)	0	0	0	0
(AM,CM)	0	0	0	0
(AC,AM)	0	0	0	0
(AC,CM)	0	0	0	0
(ACM)	.494	.494	.682	.559
(AC,AM,CM)	.506	.506	.318	.441

[1]Results from prior set 1, [2]Results from prior set 2, [3]Results from prior set 3, and [4]Results from prior set 4.

TABLE A4
Maximum likelihood parameter estimates for the two most likely models

Parameters	Model (AC,AM,CM)	(ACM)
Intercept	5.633	5.631
A	0.488	0.491
C	-1.887	-1.870
M	-5.309	-4.938
A*M	2.986	2.600
A*C	2.055	2.035
C*M	2.848	2.275
A*C*M	------	0.590

Table A3 showed the results of posterior probabilities, i.e., the model weights in BMA. First, it was notable that the posterior probabilities put most mass on two models, namely, (ACM) and (AC,AM,CM), and the other seven models received zero probability in all four settings. Second, the differences in posterior probabilities under the four settings were relatively small except those of set three. Since the coefficients of model (ACM) and (AC, AM, CM) were very close, as shown in Table A4, weighted averages (BMA estimates) in set three does not differ a lot from those in other sets. In this case, the likelihood function (the data) yielded more information than the priors. Third, prior set one and two lead to exactly the same posterior probability (.506) for model (AC,AM,CM), although the prior probabilities for this model was .196 for set one and .255 for set two. The 30% increase of prior model probability from .196 to .255 did not cause any difference in the posterior probabilities.

We conclude that different principles of assigning model priors were not making large differences to the posterior inferences in the drug data. It showed that specifying a model prior using external information did not have a practical advantage. In other words, using the default uniform model priors provided robust results in this example.

The second part of the analyses focuses on comparing posterior inferences under different reasonable choices of prior parameter distribution. As was mentioned in literature review, prior distribution of the parameter was assumed to be normal, and distributed as $(\theta \mid M) \sim N(\mathbf{v}, \mathbf{U})$, where $\mathbf{v} = (v_1, 0, ..., 0)$ and $\mathbf{U} = \mathbf{diag}\{\psi^2, \phi^2, ...\phi^2\}$. Raftery (1996) suggested let $v_1 = 0$ and $\phi = 1$. We wanted to know if the posterior was impacted by different choices of ϕ. The reasonable range for ϕ suggested by Raftery (1996) is from 1 to 5, with 1.65 as a "central value." The ϕ values in this part of analysis were set to 1, 1.65 and 5. The posterior probabilities with different priors for parameters were displayed in Table A5. In this part of analyses, the prior model weights were assumed to be uniform.

TABLE A5
Posterior probabilities with different priors for parameters

Models	Probabilities[1]	Probabilities[2]	Probabilities[3]
(A,C, M)	0	0	0
(M,AC)	0	0	0
(A,CM)	0	0	0
(C,AM)	0	0	0
(AM,CM)	0	0	0
(AC,AM)	0	0	0
(AC,CM)	0	0	0
(ACM)	.945	.559	.147
(AC,AM,CM)	.055	.441	.853

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE A6
BMA results for Mean(θ|data) under different parameter prior assumptions

Parameters	Mean(θ|data)		
	phi=1	phi=1.65	phi=5
Intercept	5.631	5.632	5.633
A	0.491	0.490	0.488
C	-1.871	-1.878	-1.885
M	-4.958	-5.102	-5.254
A*M	2.621	2.770	2.929
A*C	2.036	2.044	2.052
C*M	2.307	2.528	2.764
A*C*M	0.558	0.330	0.087

Table A6 presented the BMA results for Mean(θ|data) under different parameter prior assumptions. The BMA Mean(θ|data) was quite similar for the three phi values except the parameter "A*C*M". This was because the term "A*C*M " was the only difference between these two models. In model (AC,AM,CM), the coefficient of "A*C*M " was considered to be zero, and it was given a weight of .853 when phi is set to 5, vs .055 when phi is set to 1. This made the coefficient of "A*C*M " vary more than other parameters. On the whole, the impact of constructing different parameter priors was minimal in this case.

Both parts of the sensitivity analyses results suggest that with large small sizes (N=2276, and minimum cell size is two), reasonable choices of prior distributions have minor effects on posterior probabilities in the drug data. Although posterior probabilities are functions of prior distributions, the likelihood function overweighs the priors when the sample size is large. Since this result might be just coincident, the conclusion should not be generalized to other dataset and/or other settings.

BMA vs. Classical Analysis

In classical loglinear analysis of drug data, the nine possible models were examined, the best fit model was chosen based on AIC or other criteria, and inference was based on the "best" model. This practice ignored model uncertainty. BMA had been advocated as a formal way to circumvent the problem of model uncertainty. Posterior means and standard deviations were used in BMA inferences instead of parameter estimates and standard errors. The BMA analogue of the p-value is the quantity $P(\theta \neq 0|data)$.

TABLE A7
Drug data example: comparison of BMA results (via Bayes factor) to post-model-selection estimates

	Bayesian model averaging (phi=1.65)			Model (AC,AM,CM)	
Parameter	Mean θ |Data	SD θ |Data	$P(\theta \neq 0|data)$	Estimate (se)	p-value
A	0.490	0.076	1	0.488 (0.076)	< 0.0001
C	-1.877	0.164	1	-1.887 (0.163)	< 0.0001
M	-5.102	0.645	1	-5.309 (0.475)	< 0.0001
A*M	2.770	0.661	1	2.986 (0.465)	< 0.0001
A*C	2.044	0.176	1	2.055 (0.174)	< 0.0001
C*M	2.528	0.844	1	2.848 (0.164)	< 0.0001
A*C*M	0.330	0.790	0.559	-------	------

Table A7 compared the BMA result to post-model-selection estimator. The model selection criteria used was the standard AIC, since the models were not hierarchical, G^2 was not appropriate. Model (AC,AM,CM) was chosen as the "best " model because it has the minimum AIC value. Nonetheless, (AC,AM,CM) represented only 44.1% of the total posterior probability, indicating the amount of model uncertainty is ineligible. The BMA result of Mean (θ

|Data) was very similar to the post-model-selection estimates. However, the standard errors of the estimates of model (AC, AM, CM) were much smaller than the SD(θ|Data) of BMA estimates, indicating that post-model-selection estimators substantially underestimated variability and produced too optimistic confidence interval. For example, the standard deviation of the posterior distribution of the coefficient of the term C*M is .844 under BMA, compared with .164 (s.e.) of the post-model-selection estimator. The 95% confidences interval formed for this parameter from BMA was about five times wider than the one from model selection. The results in table A7 suggested that the confidence intervals formed for BMA estimates were much wider than those of a single chosen model. Inference based on a set of candidate models rather than a single model incorporated model selection into inference. The posterior probabilities and the p-values agreed that there were very strong evidence for all the one way effect and two way associations. "Strong" can be interpreted based on p-value smaller than 0.001, and P(θ≠0|data) greater than 0.99.

Figure A2 graphically presented the marginal posterior distribution of the coefficients. The spike in the plot of "X7" (the three way association term ACM) corresponded to P($\theta \neq 0$ |Data)= 0.559.

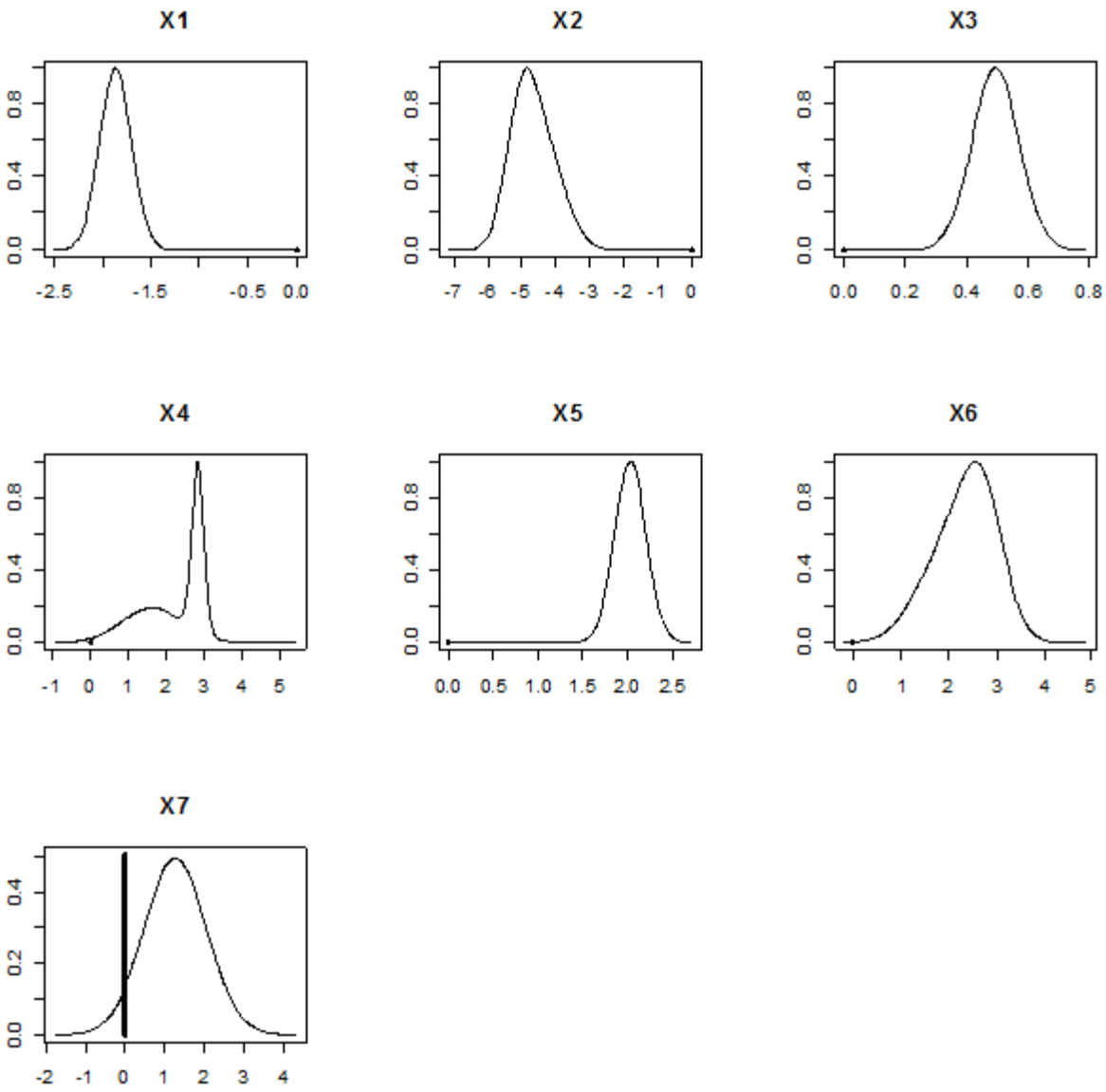

FIGURE A2. Posterior densities of coefficients in BMA results, X1-X7 representing variables C, M, A, CM, AC, AM, AMC, in that order (intercept is not plotted).

BMA: with vs. without Following "Principle of Hierarchy"

For the two-by-two-by-two drug data, if the "principle of hierarchy" (Reynolds, 1977) as followed, the permissible models were nine, which is a small number. If the "principle of hierarchy" was not followed, the permissible models were 2^7=128, since each term could be included or excluded. Although it was difficult to interpret the model having terms AMC but not A, it is mathematically feasible. The R glib function requires the user to specify all the models to be considered. It was practically unfeasible to manually input 128 models in one function (consider the possible mistakes one will make). A better alternative, in this case, is the R bic.glm function in the BMA package, which utilizes BIC to approximate the Bayes factor in large samples (Kass & Raftery, 1995).

The posterior probability is given by

$$\Pr(M^{(i)} \mid \text{data}) \approx \frac{\text{pr}(M^{(i)})\ \exp(-.5\text{BIC}(M^{(i)}))}{\sum_i \text{pr}(M^{(i)})\ \exp(-.5\text{BIC}(M^{(i)}))}, \qquad (27)$$

The function bic.glm could carry out Bayesian model averaging analysis for generalized linear models. Another feature of this function was the utilization of "Occam's razor" in reducing number of models to make the summation manageable. The principle of Occam's razor states that one should not increase, beyond what is necessary, to number of entities required to explain things. This principle is useful in the statistical model building process, because the subject domain could become unlimited complex in some cases (Jefferys & Berger, 1991). Under this principle, the models with posterior probability far less than the best model in the model space are excluded. The rule was to get rid of those models belonging to the set (Madigan & Raftery, 1994):

$$S = \left\{ \mathrm{M}^{(k)} : \frac{\max(\mathrm{pr}(\mathrm{M} \mid \mathrm{data}))}{\mathrm{pr}(\mathrm{M}^{(k)} \mid \mathrm{data})} > c \right\},$$

where the value of *c* are determined by the context. Madigan and Raftery (1994) suggested in a general context, the default *c* value is 20, compared with a *c* value of 1,000 for forensic evidence in criminal cases suggested by Evett (1991).

Reducing the number of models greatly reduced the amount of computation in the analysis of three dimensional table, and it was especially important in accounting for model uncertainty in higher dimensional tables. For example, the number of all possible loglinear models for four dimensional contingency table is 2^{15}=32768, if the law of hierarchy is not followed. Table A8 containd the results of the BMA analysis, averaged over a set of parsimonious, data-supported models. Table A9 listed the models with highest posterior model probability (PMP).

TABLE A8
BMA results utilizing Occam's razor when the "hierarchy principle" is not followed

parameter	Mean θ|data	SD θ|data	P(θ≠0|data)
A	0.489	0.076	1
C	-1.879	0.164	1
M	-5.152	0.614	1
A*M	2.824	0.618	1
A*C	2.046	0.175	1
C*M	2.557	0.797	.95
A*C*M	0.297	0.796	.33

TABLE A9
Models with highest posterior probability

	C	M	A	C*M	A*C	A*M	A*M*C	PMP[1]
Model 1	•	•	•	•	•	•		0.670
Model 2	•	•	•	•	•	•	•	0.285
Model 3	•	•	•		•	•	•	0.045

[1]PMP denotes posterior probability
[2]dot denotes the inclusion of the term in the model

When the law of hierarchy was not followed, and the Ockham's razor constant c was set to 20, three models were left in the parsimonious model set: the saturated model, the all two-way-association model and a model consisting of the terms C, M, A, A*C, A*M, and A*M*C, which could not be represented by the model symbols we used. Model 3 in table A9 represented 4.5% of the total posterior probability. These three models accounted for virtually 100 percent of the posterior probability, which means the other models excluded by the Ockham's razor represented negligible amount of posterior probability. Nonetheless, it is still possible that in some datasets, the many models with small posterior probabilities contribute collectively a fair amount of the posterior probabilities. If that happens, the researcher need to reset the Ockham's razor constant to allow more models to enter, thus increasing the amount of posterior probability accounted for by the parsimonious model set. Moulton (1991) gave an extreme example of such occasion: in about 4000 models, over 800 models were required to account for the 90% of the posterior probability.

Model Weights Obtained Via Bootstrap Model Averaging, AIC, and BMA

Although most of the statistical approaches to handling model uncertainty are Bayesian, frequentist alternatives do exist. Bootstrapping has been proposed to determine

the relative frequencies of each model being designated as "best," thus constructing a set of weights. Little justifications are given in these works and the advocators acknowledge the need for more research in this area. Efron and Gong (1983) considered a data-based process of explanatory variable selection for a logistic model. They applied the selection process to bootstrap replications of the data, obtaining a distribution of logistic models, which represent uncertainty about the models.

Buckland, Burnham, and Augustin (1997) investigated the performance of different bootstrap resampling methods in producing the model averaging weights: nonparametric bootstrap, parametric bootstrap, and bootstrap from the residuals. Martin and Roberts (2006) did a bootstrap model averaging in time series studies and compared its results with BMA in their simulation studies. They found that bootstrap model averaging and BMA offered very similar results and they did not favor one method over the other.

In this section, we generated bootstrap resamples, and applied the model selection procedure to each resample. The relative frequencies for each model was treated as weight for each model in the bootstrap model averaging. As Candolo, Davison, and Demetrio (2003) pointed out the bootstrap variance thus obtained was too large to make sense, we did not use the bootstrap method to compute variances of the parameters in the averaged model. Our goal was to compare the weights given by bootstrap method, the AIC approximation method, and the BMA method.

The method of obtaining bootstrapping model averaging weights consisted of three steps:

Step 1. Sample with replacement from the cases until the resample contains the same number of individuals as the original sample. We sampled the case numbers, indexed from 1 to N (2276). We calculated the cumulative cell frequencies of the original data, and then calculate how many resampled cases indices fell between the cumulative cell frequencies, thus generating the resampled contingency table.

Step 2. Follow through the estimation procedure, including model selection by minimum AIC, on this resample exactly as if it had been the observed sample.

Step 3. Repeat this process 1000 times to generate the bootstrap estimate of the relative frequencies of each model.

Clyde (2000) noted that BIC could be used in approximating the model weights in large samples. Suppose a_i is the prior probability placed on model *k*. The model weights could be approximated as (Clyde, 2000):

$$W_k^{BIC} = \frac{\exp(-\frac{1}{2}\Delta BIC(M^{(i)}))a_i}{\sum_{k=1}^{K}\exp(-\frac{1}{2}\Delta BIC(M^{(k)}))a_k}, \quad (28)$$

where $\Delta BIC(M^{(i)}) = BIC(M^{(i)}) - \min(BIC(M))$. Buckland, Burnham, and Augustin (1997) proposed an approximation of the model weights based on AIC under the frequentist framework. The weight is given by (Buckland, Burnham, & Augustin, 1997):

$$W_k^{AIC} = \frac{\exp(\text{-AIC}(\text{M}^{(\text{k})})/2)}{\sum_{\text{i=1}}^{\text{K}}\exp(\text{-AIC}(\text{M}^{(\text{i})})/2)}, \quad (29)$$

This definition of weight leads to the fact that two models with same AICs would be given the same weight, even if they have different number of parameters. They obtained

the averaged estimator and its variance for the weighted estimator, which were given by (Buckland, Burnham, & Augustin, 1997):

$$\hat{\bar{\theta}} = \sum_{k=0}^{K} \mathrm{W_k}\hat{\theta}_k ,(30) \qquad (30)$$

$$\mathrm{var}(\hat{\theta}) = \left[\sum_{\mathrm{k}=0}^{\mathrm{K}} \mathrm{W_k}\sqrt{\mathrm{var_k}\left(\hat{\theta}_k\right) + B_k^2}\right]^2, \qquad (31)$$

where B_k can be replaced by $\hat{B}_k = \hat{\theta}_k - \hat{\bar{\theta}}$, and $\mathrm{var_k}\left(\hat{\theta}_k\right)$ can be obtained from the fitted models by their standard method. Burnham and Anderson (2004) noted that there was a connection between W_k^{AIC} and W_k^{BIC}. They pointed out W_k^{AIC} was a form of W_k^{BIC}, if the prior weight a_i was specified in a "savvy" way. See Burnham and Anderson (2002) for further discussion of the topic.

TABLE A10
Model weights given by the bootstrap method, AIC approximation, and BMA

	Bootstrap method	AIC approximation	BMA (phi=1.65)
(AMC)	.300	.307	.559
(AM,AC,CM)	.700	.697	.441
(AM,CM)	0	0	0
(AC,AM)	0	0	0
(AC,CM)	0	0	0
(M,AC)	0	0	0
(A,CM)	0	0	0
(C,AM)	0	0	0
(A,C, M)	0	0	0

The results in Table A10 showed that all three methods put weights are on two models, (ACM) and (AC,AM,CM), exclusively. The two frequentist methods, Bootstrap method and AIC approximation, gave essentially the same weight to each model, while the BMA method allocated different weight on the two models. Although the model weights were somewhat different for the three methods, the averaged means of the parameters are still similar across the methods. It was because the parameter estimates for

the two models (AMC) and (AM,AC,CM) are very close in magnitude. Buckland, Burnham, and Augustin (1997) pointed out that for generalized linear models, observations were assumed to be independently distributed, but their variance is a function of their expectation, so the cases were not independently and identically distributed, thus nonparametric bootstrap should be replaced by the parametric bootstrap. Candolo, Davison, and Demetrio (2003) showed that for Poisson regression setting, the parameteric bootstrap and nonparametric bootstrap gave very similar results in terms of model weight. If we consider loglinear model as a special case of Poisson regression, the nonparametric bootstrap method used in this section was a reasonable approximation to the ideal result. The bootstrap method has a conceptual advantage in that the analysis of the same data does not lead to incompatible results when different priors are specified, thereby minimizing the need for incorporating relevant prior information. However, the difficulty of bootstrap methods lies in the fact that for some type of models, nonparametric bootstrap is still a very difficult topic.

Appendix B: Simulation Results

TABLE B1
Percentage of time each model is selected (%) under model selection for case 1-4

Models		Case 1	Case 2	Case 3	Case 4
n=50					
$M^{(1)}$: (XYZ)	Overspecified	6.61	10.22	4.51	6.14
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	3.06	15.09	2.94	15.03
$M^{(3)}$: (XY, XZ)	**True model**	16.06	70.08	14.75	69.74
$M^{(4)}$: (XY, YZ)	Misspecified	11.30	1.19	11.12	2.28
$M^{(5)}$: (XY, Z)	Underspecified	62.97	3.42	66.68	6.81
n=100					
$M^{(1)}$: (XYZ)	Overspecified	5.47	12.21	3.74	7.77
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	3.47	13.36	2.81	12.24
$M^{(3)}$: (XY, XZ)	**True model**	20.30	74.24	17.52	79.45
$M^{(4)}$: (XY, YZ)	Misspecified	11.38	.04	10.69	.18
$M^{(5)}$: (XY, Z)	Underspecified	59.38	.15	65.24	.36
n=200					
$M^{(1)}$: (XYZ)	Overspecified	5.38	9.32	3.16	8.09
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	4.90	13.29	3.61	11.67
$M^{(3)}$: (XY, XZ)	**True model**	26.76	77.39	22.58	80.24
$M^{(4)}$: (XY, YZ)	Misspecified	9.41	.00	9.83	.00
$M^{(5)}$: (XY, Z)	Underspecified	53.55	.00	60.82	.00
n=500					
$M^{(1)}$: (XYZ)	Overspecified	6.40	8.46	3.91	6.10
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	7.87	13.59	5.28	11.65
$M^{(3)}$: (XY, XZ)	**True model**	43.67	77.95	38.70	82.25
$M^{(4)}$: (XY, YZ)	Misspecified	6.23	.00	7.14	.00
$M^{(5)}$: (XY, Z)	Underspecified	35.83	.00	44.97	.00
n=1000					
$M^{(1)}$: (XYZ)	Overspecified	7.46	8.16	4.31	5.70
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	10.49	13.38	8.33	11.63
$M^{(3)}$: (XY, XZ)	**True model**	61.89	78.46	57.21	82.67
$M^{(4)}$: (XY, YZ)	Misspecified	2.97	.00	4.21	.00
$M^{(5)}$: (XY, Z)	Underspecified	17.19	.00	25.94	.00

TABLE B2
Percentage of time each model is selected (%) for case 5-8

Models		Case 5	Case 6	Case 7	Case 8
n=50					
$M^{(1)}$: (XYZ)	Overspecified	4.62	7.02	3.41	4.28
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	1.11	6.45	1.00	6.52
$M^{(3)}$: (XY, XZ)	**True model**	3.26	31.67	2.82	32.89
$M^{(4)}$: (XY, YZ)	Misspecified	3.29	.80	2.47	1.11
$M^{(5)}$: (XZ, YZ)	Misspecified	2.36	10.26	1.84	9.46
$M^{(6)}$: (XY, Z)	Underspecified	30.63	1.63	49.70	4.77
$M^{(7)}$: (XZ, Y)	Misspecified	30.64	41.47	20.87	39.66
$M^{(8)}$: (YZ, X)	Misspecified	25.69	.70	17.87	1.31
n=100					
$M^{(1)}$: (XYZ)	Overspecified	3.92	9.59	2.55	6.71
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	1.04	7.30	.86	7.40
$M^{(3)}$: (XY, XZ)	**True model**	5.37	46.26	4.75	52.06
$M^{(4)}$: (XY, YZ)	Misspecified	2.84	.02	2.60	.12
$M^{(5)}$: (XZ, YZ)	Misspecified	2.87	8.76	1.87	6.64
$M^{(6)}$: (XY, Z)	Underspecified	31.69	.06	49.77	.21
$M^{(7)}$: (XZ, Y)	Misspecified	30.71	27.99	21.82	26.85
$M^{(8)}$: (YZ, X)	Misspecified	21.56	.02	15.78	.01
n=200					
$M^{(1)}$: (XYZ)	Overspecified	3.64	8.87	2.12	7.55
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	1.72	9.21	.94	9.60
$M^{(3)}$: (XY, XZ)	**True model**	9.43	66.05	7.99	70.07
$M^{(4)}$: (XY, YZ)	Misspecified	3.56	.00	3.21	.00
$M^{(5)}$: (XZ, YZ)	Misspecified	3.17	5.06	2.50	3.42
$M^{(6)}$: (XY, Z)	Underspecified	31.25	.00	48.15	.00
$M^{(7)}$: (XZ, Y)	Misspecified	31.71	10.81	23.24	9.36
$M^{(8)}$: (YZ, X)	Misspecified	15.52	.00	11.85	.00
n=500					
$M^{(1)}$: (XYZ)	Overspecified	4.90	8.50	3.04	5.87
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	4.42	12.88	3.44	12.12
$M^{(3)}$: (XY, XZ)	**True model**	24.61	77.58	21.98	81.56
$M^{(4)}$: (XY, YZ)	Misspecified	3.37	.00	4.37	.00
$M^{(5)}$: (XZ, YZ)	Misspecified	3.62	.59	2.53	.24
$M^{(6)}$: (XY, Z)	Underspecified	26.82	.00	38.31	.00
$M^{(7)}$: (XZ, Y)	Misspecified	25.53	.45	21.01	.21
$M^{(8)}$: (YZ, X)	Misspecified	6.73	.00	5.32	.00
n=1000					
$M^{(1)}$: (XYZ)	Overspecified	6.90	8.48	4.30	5.37
$M^{(2)}$: (XY, YZ, XZ)	Overspecified	8.26	12.90	6.96	11.97
$M^{(3)}$: (XY, XZ)	**True model**	48.74	78.61	45.88	82.65
$M^{(4)}$: (XY, YZ)	Misspecified	2.57	.00	3.20	.00
$M^{(5)}$: (XZ, YZ)	Misspecified	2.22	.01	1.68	.01
$M^{(6)}$: (XY, Z)	Underspecified	15.41	.00	24.42	.00
$M^{(7)}$: (XZ, Y)	Misspecified	14.76	.00	12.36	.00
$M^{(8)}$: (YZ, X)	Misspecified	1.14	.00	1.20	.00

TABLE B3

Simulation results of the estimates of λ_{11}^{XY} in case 2

Sample size	Estimate	Method				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.031	-0.343	-1.065	-0.796	-0.543
	SE	3.757	7.795	4.497	4.047	3.573
	MSE	14.118	60.878	21.352	17.013	13.063
	Rel EFF		4.312	1.512	1.205	0.925
100	Bias	0.025	-0.859	-0.624	-0.476	-0.294
	SE	0.666	4.998	2.580	2.174	1.709
	MSE	0.445	25.715	7.043	4.953	3.007
	Rel EFF		57.851	15.844	11.142	6.765
200	Bias	0.011	-0.177	-0.067	-0.061	-0.042
	SE	0.412	2.169	1.063	0.935	0.755
	MSE	0.170	4.737	1.133	0.879	0.572
	Rel EFF		27.908	6.676	5.176	3.367
500	Bias	0.003	-0.002	0.009	0.004	0.003
	SE	0.253	0.340	0.272	0.268	0.259
	MSE	0.064	0.116	0.074	0.072	0.067
	Rel EFF		1.800	1.154	1.117	1.044
1000	Bias	-0.001	-0.004	0.002	0.000	0.000
	SE	0.175	0.232	0.186	0.182	0.177
	MSE	0.031	0.054	0.035	0.033	0.031
	Rel EFF		1.766	1.132	1.088	1.025

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B4

Simulation results of the estimates of λ_{11}^{XZ} in case 2

Sample size	Estimate	Method				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.382	-2.148	-3.032	-2.503	-1.752
	SE	2.573	10.581	5.904	5.429	5.036
	MSE	6.765	116.561	44.04	35.739	28.426
	Rel EFF		17.231	6.510	5.283	4.202
100	Bias	-0.059	-2.008	-1.482	-1.235	-0.837
	SE	0.525	7.664	3.781	3.398	2.914
	MSE	0.279	62.755	16.487	13.070	9.190
	Rel EFF		224.900	59.087	46.839	32.934
200	Bias	-0.024	-0.478	-0.235	-0.220	-0.173
	SE	0.354	3.224	1.508	1.378	1.154
	MSE	0.126	10.621	2.329	1.946	1.361
	Rel EFF		84.490	18.527	15.479	10.827
500	Bias	-0.009	-0.029	-0.013	-0.017	-0.013
	SE	0.223	0.379	0.255	0.247	0.231
	MSE	0.050	0.144	0.065	0.061	0.054
	Rel EFF		2.898	1.312	1.231	1.078
1000	Bias	-0.005	-0.016	-0.006	-0.007	-0.006
	SE	0.156	0.260	0.174	0.167	0.158
	MSE	0.024	0.068	0.030	0.028	0.025
	Rel EFF		2.777	1.245	1.150	1.030

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B5

Simulation results of the estimates of λ_{11}^{XY} in case 3

Sample size	Estimate	Method				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.040	-0.177	-0.268	-0.162	-0.057
	SE	0.722	2.960	1.872	1.350	0.759
	MSE	0.524	8.789	3.577	1.849	0.579
	Rel EFF		16.788	6.833	3.531	1.106
100	Bias	-0.006	-0.056	-0.044	-0.029	-0.009
	SE	0.431	1.259	0.742	0.621	0.461
	MSE	0.186	1.589	0.552	0.386	0.212
	Rel EFF		8.552	2.972	2.080	1.144
200	Bias	-0.006	-0.007	-0.009	-0.007	-0.006
	SE	0.298	0.340	0.300	0.299	0.298
	MSE	0.089	0.115	0.090	0.089	0.089
	Rel EFF		1.299	1.017	1.005	1.000
500	Bias	-0.005	-0.007	-0.006	-0.005	-0.005
	SE	0.188	0.212	0.189	0.189	0.188
	MSE	0.035	0.045	0.036	0.036	0.035
	Rel EFF		1.272	1.013	1.006	1.000
1000	Bias	-0.002	-0.002	-0.002	-0.002	-0.002
	SE	0.130	0.149	0.131	0.131	0.130
	MSE	0.017	0.022	0.017	0.017	0.017
	Rel EFF		1.309	1.014	1.003	1.000

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B6

Simulation results of the estimates of λ_{11}^{XZ} in case 3

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	0.034	-0.071	-0.089	-0.138	-0.237
	SE	0.910	2.097	1.242	0.899	0.432
	MSE	0.829	4.403	1.549	0.828	0.243
	Rel EFF		5.310	1.869	0.999	0.293
100	Bias	0.016	-0.131	-0.166	-0.193	-0.26
	SE	0.508	0.586	0.344	0.313	0.172
	MSE	0.258	0.360	0.146	0.135	0.097
	Rel EFF		1.393	0.564	0.522	0.375
200	Bias	0.006	-0.133	-0.178	-0.205	-0.264
	SE	0.351	0.355	0.219	0.199	0.128
	MSE	0.123	0.144	0.080	0.082	0.086
	Rel EFF		1.166	0.648	0.664	0.698
500	Bias	0.004	-0.091	-0.167	-0.195	-0.254
	SE	0.221	0.270	0.184	0.174	0.126
	MSE	0.049	0.081	0.062	0.068	0.080
	Rel EFF		1.659	1.264	1.393	1.641
1000	Bias	0.002	-0.047	-0.136	-0.164	-0.229
	SE	0.155	0.207	0.176	0.173	0.140
	MSE	0.024	0.045	0.050	0.057	0.072
	Rel EFF		1.888	2.071	2.362	3.021

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B7

Simulation results of the estimates of λ_{12}^{XZ} in case 3

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	0.024	-0.112	-0.111	-0.154	-0.238
	SE	1.072	1.988	1.254	0.921	0.471
	MSE	1.150	3.965	1.584	0.872	0.279
	Rel EFF		3.448	1.377	0.758	0.242
100	Bias	0.022	-0.124	-0.162	-0.190	-0.258
	SE	0.517	0.583	0.347	0.313	0.168
	MSE	0.268	0.356	0.146	0.134	0.095
	Rel EFF		1.327	0.546	0.500	0.355
200	Bias	0.004	-0.128	-0.178	-0.204	-0.263
	SE	0.353	0.360	0.220	0.201	0.127
	MSE	0.125	0.146	0.080	0.082	0.086
	Rel EFF		1.166	0.641	0.656	0.686
500	Bias	0.002	-0.089	-0.167	-0.194	-0.253
	SE	0.222	0.272	0.187	0.177	0.130
	MSE	0.049	0.082	0.063	0.069	0.081
	Rel EFF		1.661	1.276	1.403	1.644
1000	Bias	0.001	-0.047	-0.136	-0.163	-0.229
	SE	0.156	0.207	0.177	0.174	0.142
	MSE	0.024	0.045	0.050	0.057	0.072
	Rel EFF		1.845	2.043	2.326	2.966

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B8

Simulation results of the estimates of λ_{11}^{XY} in case 4

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	0.356	0.915	-0.493	-0.115	0.198
	SE	3.440	7.255	5.352	4.407	3.069
	MSE	11.957	53.466	28.887	19.429	9.457
	Rel EFF		4.472	2.416	1.625	0.791
100	Bias	0.046	-0.401	-0.738	-0.382	-0.071
	SE	0.869	4.739	3.045	2.245	1.269
	MSE	0.758	22.613	9.818	5.187	1.616
	Rel EFF		29.828	12.951	6.843	2.132
200	Bias	0.013	-0.337	-0.203	-0.088	-0.007
	SE	0.397	2.947	1.363	0.904	0.528
	MSE	0.158	8.797	1.899	0.826	0.279
	Rel EFF		55.777	12.039	5.234	1.771
500	Bias	0.008	-0.013	0.006	0.004	0.006
	SE	0.240	0.685	0.369	0.334	0.274
	MSE	0.058	0.469	0.136	0.111	0.075
	Rel EFF		8.119	2.353	1.930	1.297
1000	Bias	0.005	0.004	0.007	0.005	0.005
	SE	0.171	0.228	0.173	0.172	0.171
	MSE	0.029	0.052	0.030	0.030	0.029
	Rel EFF		1.777	1.027	1.012	1.002

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B9

Simulation results of the estimates of λ_{11}^{XZ} in case 4

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-1.083	-1.622	-3.166	-2.089	-0.488
	SE	4.481	8.890	6.541	5.489	4.075
	MSE	21.246	81.656	52.809	34.492	16.840
	Rel EFF		3.843	2.486	1.623	0.793
100	Bias	-0.161	-1.262	-1.884	-1.129	-0.198
	SE	1.315	7.174	4.471	3.457	2.076
	MSE	1.755	53.051	23.534	13.226	4.348
	Rel EFF		30.233	13.411	7.537	2.478
200	Bias	-0.044	-0.825	-0.510	-0.297	-0.095
	SE	0.429	4.505	2.011	1.449	0.803
	MSE	0.186	20.972	4.305	2.186	0.654
	Rel EFF		112.986	23.194	11.779	3.522
500	Bias	-0.012	-0.053	-0.024	-0.021	-0.016
	SE	0.265	0.838	0.445	0.409	0.32
	MSE	0.070	0.705	0.198	0.168	0.103
	Rel EFF		10.031	2.823	2.385	1.462
1000	Bias	-0.009	-0.012	-0.008	-0.009	-0.009
	SE	0.185	0.261	0.187	0.186	0.185
	MSE	0.034	0.068	0.035	0.035	0.034
	Rel EFF		1.991	1.028	1.009	1.001

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B10

Simulation results of the estimates of λ_{12}^{XZ} in case 4

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-1.068	-1.671	-3.261	-2.152	-0.512
	SE	4.465	8.774	6.533	5.482	4.097
	MSE	21.079	79.761	53.314	34.68	17.045
	Rel EFF		3.784	2.529	1.645	0.809
100	Bias	-0.166	-1.258	-1.863	-1.111	-0.187
	SE	1.323	7.000	4.377	3.390	2.061
	MSE	1.777	50.581	22.622	12.721	4.281
	Rel EFF		28.458	12.728	7.157	2.408
200	Bias	-0.044	-0.796	-0.500	-0.292	-0.094
	SE	0.429	4.404	1.989	1.428	0.789
	MSE	0.186	20.027	4.204	2.125	0.631
	Rel EFF		107.698	22.610	11.430	3.396
500	Bias	-0.011	-0.053	-0.023	-0.020	-0.014
	SE	0.262	0.822	0.419	0.374	0.295
	MSE	0.069	0.678	0.176	0.140	0.087
	Rel EFF		9.819	2.551	2.028	1.266
1000	Bias	-0.010	-0.018	-0.01	-0.011	-0.01
	SE	0.182	0.261	0.185	0.183	0.182
	MSE	0.033	0.069	0.034	0.033	0.033
	Rel EFF		2.066	1.032	1.008	1.001

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B11

Simulation results of the estimates of λ_{11}^{XY} in case 5

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.064	-0.116	-0.047	-0.007	0.064
	SE	1.061	2.010	1.109	0.979	0.734
	MSE	1.129	4.052	1.231	0.958	0.542
	Rel EFF		3.590	1.091	0.848	0.480
100	Bias	-0.017	0.038	0.068	0.087	0.114
	SE	0.446	0.589	0.404	0.392	0.359
	MSE	0.199	0.349	0.168	0.161	0.142
	Rel EFF		1.750	0.841	0.808	0.712
200	Bias	-0.001	0.053	0.086	0.103	0.125
	SE	0.309	0.341	0.265	0.258	0.247
	MSE	0.096	0.119	0.077	0.077	0.077
	Rel EFF		1.246	0.810	0.804	0.800
500	Bias	-0.004	0.039	0.077	0.092	0.112
	SE	0.194	0.242	0.202	0.202	0.200
	MSE	0.038	0.060	0.047	0.049	0.052
	Rel EFF		1.596	1.245	1.302	1.392
1000	Bias	-0.001	0.018	0.063	0.076	0.098
	SE	0.136	0.177	0.168	0.170	0.174
	MSE	0.018	0.032	0.032	0.035	0.040
	Rel EFF		1.713	1.731	1.885	2.154

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B12

Simulation results of the estimates of λ_{11}^{XZ} in case 5

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.048	-0.044	0.002	0.035	0.088
	SE	1.063	1.780	0.973	0.871	0.703
	MSE	1.132	3.170	0.946	0.760	0.502
	Rel EFF		2.800	0.836	0.672	0.444
100	Bias	-0.016	0.043	0.071	0.091	0.118
	SE	0.442	0.528	0.364	0.348	0.322
	MSE	0.196	0.281	0.138	0.129	0.118
	Rel EFF		1.433	0.703	0.660	0.602
200	Bias	-0.007	0.050	0.082	0.099	0.121
	SE	0.307	0.342	0.266	0.259	0.249
	MSE	0.094	0.120	0.077	0.077	0.076
	Rel EFF		1.270	0.823	0.816	0.812
500	Bias	0.000	0.042	0.080	0.095	0.115
	SE	0.195	0.244	0.202	0.201	0.199
	MSE	0.038	0.061	0.047	0.049	0.053
	Rel EFF		1.611	1.247	1.302	1.388
1000	Bias	-0.001	0.019	0.062	0.076	0.097
	SE	0.138	0.179	0.169	0.172	0.175
	MSE	0.019	0.032	0.032	0.035	0.040
	Rel EFF		1.704	1.709	1.859	2.120

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B13

Simulation results of the estimates of λ_{11}^{XY} in case 6

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	0.066	-0.074	-0.830	-0.626	-0.577
	SE	3.794	6.956	4.010	3.474	2.840
	MSE	14.397	48.382	16.766	12.458	8.398
	Rel EFF		3.361	1.165	0.865	0.583
100	Bias	0.002	-0.701	-0.667	-0.565	-0.589
	SE	0.727	4.118	2.152	1.690	1.265
	MSE	0.529	17.451	5.073	3.176	1.947
	Rel EFF		33.007	9.595	6.007	3.683
200	Bias	0.008	-0.243	-0.198	-0.250	-0.366
	SE	0.403	2.176	1.033	0.865	0.677
	MSE	0.162	4.792	1.105	0.811	0.593
	Rel EFF		29.559	6.818	5.001	3.655
500	Bias	0.001	-0.010	-0.023	-0.045	-0.089
	SE	0.249	0.425	0.335	0.351	0.388
	MSE	0.062	0.181	0.112	0.125	0.158
	Rel EFF		2.911	1.808	2.009	2.546
1000	Bias	0.000	-0.004	0.001	-0.002	-0.005
	SE	0.175	0.231	0.189	0.189	0.191
	MSE	0.031	0.054	0.036	0.036	0.037
	Rel EFF		1.751	1.172	1.168	1.199

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B14

Simulation results of the estimates of λ_{11}^{XZ} in case 6

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.384	-1.665	-2.564	-1.93	-1.095
	SE	2.551	9.153	5.367	4.639	3.654
	MSE	6.654	86.547	35.377	25.240	14.547
	Rel EFF		13.006	5.316	3.793	2.186
100	Bias	-0.066	-1.625	-1.292	-0.963	-0.521
	SE	0.584	7.027	3.442	2.944	2.258
	MSE	0.346	52.021	13.512	9.595	5.370
	Rel EFF		150.537	39.100	27.765	15.539
200	Bias	-0.027	-0.490	-0.243	-0.212	-0.145
	SE	0.358	3.264	1.500	1.320	1.012
	MSE	0.129	10.895	2.308	1.788	1.044
	Rel EFF		84.530	17.905	13.872	8.102
500	Bias	-0.016	-0.037	-0.021	-0.025	-0.021
	SE	0.222	0.452	0.286	0.278	0.262
	MSE	0.050	0.206	0.082	0.078	0.069
	Rel EFF		4.141	1.659	1.565	1.395
1000	Bias	-0.004	-0.016	-0.006	-0.008	-0.006
	SE	0.154	0.257	0.169	0.164	0.156
	MSE	0.024	0.066	0.029	0.027	0.024
	Rel EFF		2.804	1.214	1.135	1.031

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B15

Simulation results of the estimates of λ_{11}^{XY} in case 7

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-0.029	-0.071	-0.118	-0.030	0.024
	SE	0.735	2.541	1.524	1.083	0.673
	MSE	0.541	6.461	2.337	1.174	0.454
	Rel EFF		11.947	4.321	2.171	0.839
100	Bias	-0.012	-0.020	0.016	0.037	0.040
	SE	0.429	1.294	0.735	0.566	0.401
	MSE	0.184	1.674	0.541	0.322	0.162
	Rel EFF		9.107	2.941	1.751	0.883
200	Bias	-0.005	0.039	0.058	0.060	0.043
	SE	0.297	0.340	0.269	0.271	0.281
	MSE	0.088	0.117	0.076	0.077	0.081
	Rel EFF		1.321	0.857	0.872	0.913
500	Bias	0.000	0.034	0.051	0.054	0.042
	SE	0.186	0.235	0.194	0.195	0.195
	MSE	0.222	0.267	0.197	0.195	0.174
	Rel EFF		1.204	0.887	0.880	0.785
1000	Bias	-0.002	0.015	0.035	0.040	0.038
	SE	0.132	0.174	0.155	0.157	0.156
	MSE	0.017	0.030	0.025	0.026	0.026
	Rel EFF		1.750	1.446	1.510	1.486

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B16

Simulation results of the estimates of λ_{11}^{XZ} in case 7

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	0.023	-0.064	-0.077	-0.109	-0.165
	SE	0.989	2.016	1.162	0.872	0.555
	MSE	0.979	4.068	1.355	0.772	0.335
	Rel EFF		4.155	1.384	0.788	0.342
100	Bias	0.010	-0.106	-0.137	-0.147	-0.191
	SE	0.513	0.542	0.376	0.344	0.276
	MSE	0.264	0.305	0.160	0.140	0.112
	Rel EFF		1.157	0.606	0.531	0.426
200	Bias	0.002	-0.107	-0.145	-0.157	-0.201
	SE	0.352	0.355	0.248	0.243	0.209
	MSE	0.124	0.137	0.082	0.084	0.084
	Rel EFF		1.108	0.665	0.677	0.678
500	Bias	-0.003	-0.080	-0.141	-0.156	-0.202
	SE	0.223	0.266	0.196	0.195	0.174
	MSE	0.050	0.077	0.058	0.062	0.071
	Rel EFF		1.557	1.178	1.258	1.432
1000	Bias	-0.002	-0.046	-0.121	-0.140	-0.189
	SE	0.155	0.205	0.177	0.178	0.165
	MSE	0.024	0.044	0.046	0.051	0.063
	Rel EFF		1.830	1.915	2.128	2.630

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B17

Simulation results of the estimates of λ_{12}^{XZ} in case 7

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	0.020	-0.057	-0.075	-0.110	-0.168
	SE	1.008	1.987	1.117	0.840	0.553
	MSE	1.016	3.951	1.252	0.718	0.334
	Rel EFF		3.890	1.233	0.707	0.329
100	Bias	0.009	-0.104	-0.137	-0.147	-0.190
	SE	0.518	0.483	0.345	0.331	0.279
	MSE	0.269	0.244	0.138	0.131	0.114
	Rel EFF		0.909	0.513	0.487	0.423
200	Bias	0.001	-0.106	-0.145	-0.158	-0.201
	SE	0.352	0.357	0.247	0.242	0.208
	MSE	0.124	0.139	0.082	0.084	0.084
	Rel EFF		1.118	0.662	0.674	0.675
500	Bias	-0.001	-0.080	-0.140	-0.156	-0.201
	SE	0.222	0.267	0.197	0.195	0.174
	MSE	0.049	0.077	0.058	0.062	0.071
	Rel EFF		1.578	1.186	1.267	1.443
1000	Bias	0.000	-0.043	-0.119	-0.138	-0.188
	SE	0.158	0.207	0.18	0.181	0.168
	MSE	0.025	0.045	0.047	0.052	0.064
	Rel EFF		1.787	1.863	2.066	2.541

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B18

Simulation results of the estimates of λ_{11}^{XY} in case 8

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	0.410	0.842	-0.273	-0.041	-0.011
	SE	3.616	6.819	4.912	3.953	2.755
	MSE	13.245	47.200	24.198	15.624	7.587
	Rel EFF		3.564	1.827	1.180	0.573
100	Bias	0.036	-0.501	-0.782	-0.542	-0.448
	SE	0.844	4.424	2.706	1.901	1.042
	MSE	0.714	19.825	7.935	3.908	1.287
	Rel EFF		27.776	11.117	5.476	1.803
200	Bias	0.013	-0.371	-0.279	-0.243	-0.307
	SE	0.398	2.964	1.315	0.949	0.671
	MSE	0.158	8.925	1.808	0.960	0.545
	Rel EFF		56.376	11.422	6.066	3.440
500	Bias	0.009	0.003	-0.002	-0.016	-0.047
	SE	0.244	0.341	0.277	0.290	0.327
	MSE	0.059	0.116	0.077	0.084	0.109
	Rel EFF		1.958	1.294	1.419	1.837
1000	Bias	0.001	-0.002	0.002	0.000	-0.001
	SE	0.169	0.227	0.173	0.172	0.174
	MSE	0.028	0.052	0.030	0.030	0.030
	Rel EFF		1.809	1.050	1.043	1.062

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B19

Simulation results of the estimates of λ_{11}^{XZ} in case 8

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-1.067	-1.538	-2.763	-1.723	-0.448
	SE	4.414	7.879	5.862	4.718	3.347
	MSE	20.617	64.44	41.993	25.227	11.400
	Rel EFF		3.126	2.037	1.224	0.553
100	Bias	-0.143	-1.205	-1.636	-0.919	-0.167
	SE	1.209	6.840	4.149	3.045	1.652
	MSE	1.481	48.231	19.890	10.116	2.756
	Rel EFF		32.568	13.431	6.831	1.861
200	Bias	-0.041	-0.785	-0.461	-0.260	-0.081
	SE	0.425	4.513	1.962	1.422	0.778
	MSE	0.182	20.981	4.062	2.089	0.612
	Rel EFF		115.306	22.323	11.480	3.361
500	Bias	-0.016	-0.044	-0.024	-0.023	-0.018
	SE	0.266	0.672	0.386	0.365	0.289
	MSE	0.071	0.454	0.150	0.134	0.084
	Rel EFF		6.379	2.104	1.880	1.182
1000	Bias	-0.009	-0.016	-0.010	-0.010	-0.009
	SE	0.185	0.261	0.188	0.186	0.185
	MSE	0.034	0.068	0.035	0.035	0.034
	Rel EFF		2.001	1.030	1.012	1.003

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B20

Simulation results of the estimates of λ_{12}^{XZ} in case 8

Sample size	Estimates	Methods				
		True Model Assumption	Model Selection	BMA[1]	BMA[2]	BMA[3]
50	Bias	-1.060	-1.472	-2.709	-1.682	-0.437
	SE	4.408	7.794	5.844	4.709	3.345
	MSE	20.554	62.900	41.487	25.002	11.376
	Rel EFF		3.060	2.018	1.216	0.553
100	Bias	-0.148	-1.140	-1.624	-0.907	-0.169
	SE	1.210	6.755	4.131	3.047	1.692
	MSE	1.486	46.924	19.702	10.106	2.893
	Rel EFF		31.581	13.260	6.802	1.947
200	Bias	-0.038	-0.764	-0.442	-0.245	-0.079
	SE	0.424	4.424	1.916	1.392	0.811
	MSE	0.181	20.152	3.865	1.997	0.664
	Rel EFF		111.351	21.355	11.035	3.667
500	Bias	-0.017	-0.045	-0.024	-0.022	-0.018
	SE	0.264	0.581	0.336	0.307	0.267
	MSE	0.070	0.340	0.114	0.095	0.072
	Rel EFF		4.847	1.618	1.352	1.021
1000	Bias	-0.008	-0.017	-0.009	-0.009	-0.008
	SE	0.185	0.262	0.188	0.186	0.185
	MSE	0.034	0.069	0.036	0.035	0.034
	Rel EFF		2.006	1.036	1.015	1.003

[1]Results from phi=1, [2]results from phi=1.65, [3]results from phi=5

TABLE B21

Simulation results: percentiles of the Z-statistic under model-selection for parameter λ_{11}^{XY} and λ_{11}^{XZ} for case 2

n	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
λ_{11}^{XY}									
50	-2.064	-1.784	-1.521	-1.200	0.001	1.294	1.590	1.866	2.086
100	-2.226	-1.918	-1.631	-1.257	0.001	1.428	1.789	2.040	2.319
200	-2.309	-2.004	-1.716	-1.366	-0.001	1.436	1.790	2.128	2.475
500	-2.475	-2.154	-1.849	-1.447	-0.017	1.430	1.836	2.175	2.582
1000	-2.508	-2.130	-1.820	-1.437	0.015	1.446	1.866	2.187	2.587
λ_{11}^{XZ}									
50	-1.947	-1.729	-1.495	-1.231	-0.001	1.036	1.368	1.604	1.830
100	-2.156	-1.912	-1.636	-1.304	-0.001	1.392	1.812	2.162	2.463
200	-2.274	-2.017	-1.771	-1.408	-0.042	1.379	1.819	2.120	2.462
500	-2.467	-2.162	-1.868	-1.484	-0.036	1.405	1.833	2.221	2.611
1000	-2.507	-2.208	-1.857	-1.479	-0.041	1.429	1.880	2.200	2.610
Z-statistic	-2.326	-1.960	-1.645	-1.282	0.000	1.282	1.645	1.960	2.326

TABLE B22
Simulation results: percentiles of the Z-statistic under model-selection for parameter λ_{11}^{XY} and λ_{11}^{XZ} for case 3

	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
λ_{11}^{XY}									
n=50	-2.223	-1.943	-1.665	-1.302	-0.037	1.277	1.660	1.949	2.336
n=100	-2.349	-2.044	-1.731	-1.327	-0.001	1.332	1.723	2.077	2.449
n=200	-2.400	-2.031	-1.713	-1.323	-0.013	1.308	1.683	2.031	2.478
n=500	-2.512	-2.071	-1.725	-1.331	-0.019	1.294	1.700	2.054	2.419
n=1000	-2.457	-2.056	-1.686	-1.318	-0.017	1.322	1.677	1.999	2.403
λ_{11}^{XZ}									
n=50	-2.296	-1.547	-0.546	0.000	0.000	1.142	1.679	1.946	2.245
n=100	-2.104	-1.259	-0.284	0.000	0.000	1.304	1.701	2.015	2.295
n=200	-1.992	-1.073	-0.256	0.000	0.000	1.297	1.696	2.004	2.390
n=500	-1.835	-1.209	-0.634	-0.023	0.000	1.294	1.701	2.050	2.429
n=1000	-1.840	-1.354	-0.937	-0.464	0.000	1.319	1.680	2.001	2.386
λ_{12}^{XZ}									
n=50	-2.495	-1.759	-0.629	0.000	0.000	1.384	2.123	2.441	2.749
n=100	-2.922	-1.406	0.000	0.000	0.000	2.628	3.625	4.174	4.692
n=200	-2.734	-1.185	0.000	0.000	0.000	4.834	5.848	6.780	7.730
n=500	-2.441	-0.910	0.000	0.000	0.000	10.081	11.783	13.266	15.152
n=1000	-1.722	-0.808	-0.054	0.000	7.194	18.028	20.389	22.602	24.954
Z-statistic	-2.326	-1.960	-1.645	-1.282	0.000	1.282	1.645	1.960	2.326

TABLE B23
Simulation results: percentiles of the Z-statistic under model-selection for parameter λ_{11}^{XY} and λ_{11}^{XZ} for case 4

	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
λ_{11}^{XY}									
n=50	-2.167	-1.849	-1.573	-1.200	0.001	1.212	1.489	1.705	1.930
n=100	-2.327	-1.96	-1.603	-1.257	0.021	1.364	1.698	1.963	2.260
n=200	-2.324	-1.982	-1.681	-1.305	0.005	1.373	1.777	2.062	2.416
n=500	-2.348	-2.039	-1.742	-1.338	0.028	1.356	1.728	2.045	2.441
n=1000	-2.500	-2.123	-1.784	-1.366	0.038	1.384	1.771	2.132	2.518
λ_{11}^{XZ}									
n=50	-1.775	-1.558	-1.357	-1.111	-0.001	0.829	1.221	1.527	1.884
n=100	-2.051	-1.825	-1.605	-1.302	-0.024	1.209	1.593	1.906	2.237
n=200	-2.235	-1.890	-1.647	-1.307	-0.023	1.301	1.710	2.061	2.425
n=500	-2.332	-2.050	-1.724	-1.360	-0.040	1.340	1.755	2.088	2.512
n=1000	-2.474	-2.089	-1.750	-1.394	-0.031	1.342	1.749	2.092	2.527
λ_{12}^{XZ}									
n=50	-1.741	-1.549	-1.395	-1.185	-0.215	0.516	0.952	1.329	1.775
n=100	-4.396	-4.113	-3.849	-3.500	-1.908	0.049	0.740	1.289	1.888
n=200	-9.817	-9.356	-8.952	-8.455	-6.266	-0.001	0.625	1.354	2.042
n=500	-26.727	-25.951	-25.277	-24.433	-20.922	-0.401	0.648	1.275	2.110
n=1000	-55.643	-54.381	-53.367	-52.118	-47.268	-0.312	0.640	1.389	2.088
Z-statistic	-2.326	-1.960	-1.645	-1.282	0.000	1.282	1.645	1.960	2.326

TABLE B24
Simulation results: percentiles of the Z-statistic under model-selection for parameter λ_{11}^{XY} and λ_{11}^{XZ} for case 5

n	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
λ_{11}^{XY}									
50	-2.179	-1.958	-1.682	-1.307	0.000	0.145	1.576	1.990	2.353
100	-2.386	-2.028	-1.739	-1.341	0.000	0.000	1.329	1.983	2.390
200	-2.290	-1.910	-1.582	-1.258	-0.003	1.315	1.658	1.931	2.345
500	-2.347	-2.008	-1.648	-1.299	0.001	1.260	1.647	1.980	2.398
1000	-2.477	-2.086	-1.736	-1.340	-0.009	0.616	0.773	1.097	1.666
λ_{11}^{XZ}									
50	-2.516	-2.162	-1.835	-1.426	0.000	0.234	1.613	2.036	2.563
100	-4.281	-3.482	-2.627	-1.610	0.000	0.000	1.222	1.927	2.496
200	-2.249	-1.897	-1.621	-1.282	-0.012	1.265	1.641	1.933	2.279
500	-2.282	-1.946	-1.660	-1.304	0.012	1.297	1.669	1.997	2.365
1000	-29.213	-26.526	-24.22	-21.346	-2.047	0.301	0.697	1.051	1.605
Z-statistic	-2.326	-1.960	-1.645	-1.282	0.000	1.282	1.645	1.960	2.326

TABLE B25
Simulation results: percentiles of the Z-statistic under model-selection for parameter λ_{11}^{XY} and λ_{11}^{XZ} for case 6

n	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
λ_{11}^{XY}									
50	-0.103	-0.001	0.000	0.000	0.000	1.284	1.598	1.855	2.078
100	-1.107	-0.531	-0.306	-0.100	0.000	1.423	1.788	2.069	2.376
200	-1.615	-1.195	-0.970	-0.730	0.000	1.439	1.805	2.106	2.494
500	-2.223	-1.998	-1.705	-1.385	0.006	1.427	1.842	2.177	2.540
1000	-2.552	-2.145	-1.801	-1.424	0.000	1.420	1.818	2.140	2.559
λ_{11}^{XZ}									
50	-2.510	-2.250	-2.012	-1.690	-0.368	0.834	1.265	1.579	1.926
100	-6.228	-5.862	-5.485	-5.019	-1.324	0.860	1.419	1.873	2.321
200	-14.116	-13.556	-13.068	-12.445	-9.416	0.498	1.264	1.837	2.363
500	-38.171	-37.287	-36.472	-35.489	-31.411	0.004	1.167	1.806	2.280
1000	-78.442	-77.100	-76.119	-74.748	-69.017	-0.078	1.092	1.822	2.338
Z-statistic	-2.326	-1.960	-1.645	-1.282	0.000	1.282	1.645	1.960	2.326

TABLE B26
Simulation results: percentiles of the Z-statistic under model-selection for parameter λ_{11}^{XY} and λ_{11}^{XZ} for case 7

	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
λ_{11}^{XY}									
n=50	-2.220	-1.911	-1.627	-1.258	0.000	1.006	1.552	1.937	2.268
n=100	-2.318	-1.997	-1.681	-1.327	0.000	0.839	1.519	1.992	2.390
n=200	-2.311	-2.002	-1.678	-1.324	0.000	0.608	1.265	1.813	2.373
n=500	-2.392	-2.009	-1.699	-1.313	0.000	0.449	0.970	1.412	1.975
n=1000	-2.439	-2.066	-1.735	-1.354	-0.017	0.683	0.872	1.240	1.875
λ_{11}^{XZ}									
n=50	-2.297	-1.841	-1.104	0.000	0.000	1.274	1.661	1.959	2.229
n=100	-2.307	-1.724	-0.822	0.000	0.000	1.309	1.695	1.994	2.348
n=200	-2.130	-1.284	-0.601	0.000	0.000	1.293	1.651	1.995	2.341
n=500	-1.886	-1.314	-0.756	-0.163	0.000	1.306	1.680	1.982	2.383
n=1000	-1.957	-1.400	-0.955	-0.527	0.000	1.332	1.686	2.022	2.333
λ_{12}^{XZ}									
n=50	-2.237	-1.826	-1.046	0.000	0.000	1.271	1.691	2.024	2.412
n=100	-2.364	-1.665	-0.818	0.000	0.000	1.418	1.994	2.895	3.864
n=200	-2.366	-1.388	-0.540	0.000	0.000	1.661	3.633	5.402	6.439
n=500	-2.199	-1.148	-0.559	0.000	0.000	8.513	10.437	11.897	13.556
n=1000	-1.955	-1.147	-0.557	0.000	1.146	17.111	19.871	22.03	24.608
Z-statistic	-2.326	-1.960	-1.645	-1.282	0.000	1.282	1.645	1.960	2.326

TABLE B27
Simulation results: percentiles of the Z-statistic under model-selection for parameter λ_{11}^{XY} and λ_{11}^{XZ} for case 8

	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99
λ_{11}^{XY}									
n=50	-0.504	-0.003	0.000	0.000	0.000	1.197	1.486	1.675	1.869
n=100	-0.608	-0.441	-0.327	-0.141	0.000	1.355	1.695	1.948	2.191
n=200	-1.313	-1.147	-0.987	-0.779	0.002	1.387	1.755	2.060	2.448
n=500	-2.276	-1.983	-1.685	-1.317	0.029	1.419	1.825	2.139	2.530
n=1000	-2.460	-2.085	-1.754	-1.381	0.003	1.366	1.740	2.087	2.473
λ_{11}^{XZ}									
n=50	-1.837	-1.577	-1.376	-1.113	-0.002	0.889	1.238	1.597	1.978
n=100	-2.045	-1.795	-1.567	-1.282	-0.025	1.223	1.621	1.981	2.326
n=200	-2.193	-1.887	-1.638	-1.295	-0.008	1.295	1.740	2.063	2.512
n=500	-2.305	-2.052	-1.768	-1.424	-0.035	1.339	1.750	2.106	2.587
n=1000	-2.419	-2.080	-1.755	-1.369	-0.056	1.317	1.762	2.122	2.481
λ_{12}^{XZ}									
n=50	-1.776	-1.576	-1.390	-1.163	-0.101	0.799	1.179	1.540	1.944
n=100	-4.245	-3.920	-3.605	-3.220	-1.117	0.802	1.337	1.697	2.114
n=200	-9.864	-9.316	-8.850	-8.300	-5.843	0.320	1.009	1.542	2.129
n=500	-26.933	-26.055	-25.277	-24.408	-20.965	-0.318	0.642	1.397	2.060
n=1000	-55.644	-54.546	-53.541	-52.203	-47.169	-0.402	0.565	1.297	2.074
Z-statistic	-2.326	-1.960	-1.645	-1.282	0.000	1.282	1.645	1.960	2.326

TABLE B28
Simulation results: p-values of Kolmogorov-Smirnov test of the empirical distribution of the Z statistics of λ_{11}^{XY}

	case 1		case 2		case 3		case 4	
	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT
n=50	.0496	.0001	.0000	.0000	.0484	.0030	.0000	.0000
n=100	.1270	.0280	.0006	.0000	.1795	.0121	.0001	.0000
n=200	.5178	.0020	.4569	.0000	.3969	.0427	.0012	.0000
n=500	.3643	.0009	.2821	.0000	.7838	.1038	.4963	.0021
n=1000	.1884	.0104	.6222	.0000	.4456	.2251	.5646	.0001

	case 5		case 6		case 7		case 8	
	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT
n=50	.0000	.0000	.0000	.0000	.0001	.0000	.0000	.0000
n=100	.1838	.0000	.0016	.0000	.3710	.0000	.0001	.0000
n=200	.7089	.0000	.0101	.0000	.3159	.0000	.7388	.0000
n=500	.1509	.0000	.0493	.0000	.8741	.0000	.1169	.0001
n=1000	.8958	.0000	.3493	.0000	.5736	.0000	.4852	.0000

TABLE B29
Simulation results: p-values of Kolmogorov-Smirnov test of the empirical distribution of the Z statistics of λ_{11}^{XZ}

	case 1		case 2		case 3		case 4	
	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT
n=50	.0014	.0000	.0000	.0000	.0024	.0000	.0000	.0000
n=100	.0550	.0000	.0000	.0000	.9881	.0000	.0000	.0000
n=200	.5164	.0000	.0000	.0000	.4166	.0000	.0000	.0000
n=500	.8330	.0000	.0478	.0000	.0117	.0000	.0005	.0000
n=1000	.4193	.0000	.1198	.0000	.0855	.0000	.0883	.0000

	case 5		case 6		case 7		case 8	
	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT
n=50	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
n=100	.0000	.0000	.0000	.0000	.3272	.0000	.0000	.0000
n=200	.0000	.0000	.0001	.0000	.3589	.0000	.0000	.0000
n=500	.9783	.0000	.0062	.0000	.6876	.0000	.0035	.0000
n=1000	.6070	.0000	.0007	.0000	.7990	.0000	.0817	.0000

TABLE B30
Simulation results: p-values of Kolmogorov-Smirnov test of the empirical distribution of the Z statistics of λ_{12}^{XZ}

	case 3		case 4		case 7		case 8	
	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT	TRUE	SELECT
n=50	.0060	.0000	.0000	.0000	.0082	.0000	.0000	.0000
n=100	.8260	.0000	.0000	.0000	.0221	.0000	.0000	.0000
n=200	.3967	.0000	.0039	.0000	.9528	.0000	.0000	.0000
n=500	.4497	.0000	.0005	.0000	.1870	.0000	.0008	.0000
n=1000	.9575	.0000	.5246	.0000	.6765	.0000	.0835	.0000

TABLE B31
Empirical means, standard error for the conditional odds ratio estimator under true model assumption, model selection, and BMA (phi=1.65) for n=50, 100, 200, 500 and 1000 for case 2

	TRUE MODEL		MODEL SELECTION		BMA (phi=1.65)	
$\hat{\theta}_{XY}$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$
n=50	2.655	2.032	3.077	3.443	4.022	4.731
n=100	2.745	1.518	3.444	3.889	3.880	4.416
n=200	2.754	1.097	3.164	3.253	3.062	2.888
n=500	2.731	0.685	2.826	1.067	2.761	0.770
n=1000	2.719	0.473	2.763	0.654	2.737	0.498
$\hat{\theta}_{XZ}$						
n=50	0.176	0.123	0.196	0.199	0.228	0.199
n=100	0.155	0.076	0.157	0.089	0.164	0.104
n=200	0.145	0.050	0.145	0.053	0.145	0.051
n=500	0.139	0.031	0.139	0.033	0.139	0.031
n=1000	0.137	0.021	0.137	0.022	0.137	0.021
$\hat{\theta}_{YZ}$						
n=50	0.801	0.125	1.063	1.174	1.313	1.277
n=100	0.887	0.057	1.216	1.384	1.324	1.437
n=200	0.941	0.022	1.132	1.130	1.085	1.014
n=500	0.976	0.006	1.029	0.356	0.993	0.178
n=1000	0.988	0.002	1.012	0.217	0.994	0.089

TABLE B32
Empirical means, standard error for the conditional odds ratio estimator under true model assumption, model selection, and BMA (phi=1.65) for n=50, 100, 200, 500 and 1000 for case 3

	TRUE MODEL		MODEL SELECTION		BMA (phi=1.65)	
$\hat{\theta}_{XY}$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$
n=50	0.870	0.486	0.985	1.737	1.012	1.776
n=100	0.818	0.336	0.847	0.541	0.826	0.483
n=200	0.777	0.227	0.790	0.305	0.779	0.230
n=500	0.754	0.141	0.759	0.172	0.758	0.141
n=1000	0.748	0.097	0.750	0.117	0.749	0.098

TABLE B33
Empirical means, standard error for the conditional odds ratio estimator under true model assumption, model selection, and BMA (phi=1.65) for n=50, 100, 200, 500 and 1000 for case 4

	TRUE MODEL		MODEL SELECTION		BMA (phi=1.65)	
$\hat{\theta}_{XY}$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$
n=50	2.429	1.653	2.746	3.694	3.521	4.560
n=100	2.589	1.350	2.967	3.081	3.437	3.241
n=200	2.659	1.001	3.085	3.162	3.101	2.856
n=500	2.704	0.636	2.805	1.286	2.721	0.839
n=1000	2.714	0.457	2.742	0.635	2.707	0.461

TABLE B34
Empirical means, standard error for the conditional odds ratio estimator under true model assumption, model selection, and BMA (phi=1.65) for n=50, 100, 200, 500 and 1000 for case 5

	TRUE MODEL		MODEL SELECTION		BMA (phi=1.65)	
$\hat{\theta}_{XY}$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$
n=50	0.91	0.541	1.043	1.435	0.986	1.25
n=100	0.825	0.349	0.901	0.587	0.874	0.278
n=200	0.789	0.239	0.846	0.306	0.855	0.186
n=500	0.757	0.145	0.801	0.206	0.831	0.151
n=1000	0.749	0.101	0.771	0.151	0.816	0.13
$\hat{\theta}_{XZ}$						
n=50	0.921	0.536	1.098	1.902	1.007	1.788
n=100	0.825	0.346	0.903	0.577	0.878	0.284
n=200	0.784	0.234	0.844	0.304	0.853	0.189
n=500	0.76	0.146	0.802	0.202	0.835	0.148
n=1000	0.75	0.103	0.773	0.154	0.812	0.131
$\hat{\theta}_{YZ}$						
n=50	1.056	0.139	1.385	2.42	1.262	1.954
n=100	1.024	0.027	1.15	1.081	1.061	0.392
n=200	1.011	0.008	1.07	0.404	1.03	0.195
n=500	1.004	0.002	1.026	0.222	1.013	0.103
n=1000	1.002	0.001	1.014	0.159	1.008	0.062

TABLE B35

Empirical means, standard error for the conditional odds ratio estimator under true model assumption, model selection, and BMA (phi=1.65) for n=50, 100, 200, 500 and 1000 for case 6

	TRUE MODEL		MODEL SELECTION		BMA (phi=1.65)	
$\hat{\theta}_{XY}$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$
n=50	2.728	2.146	2.881	4.160	3.343	4.361
n=100	2.727	1.530	3.245	4.035	3.219	4.244
n=200	2.736	1.064	3.070	3.166	2.678	2.833
n=500	2.725	0.670	2.806	1.042	2.675	0.820
n=1000	2.721	0.472	2.767	0.662	2.733	0.508
$\hat{\theta}_{XZ}$						
n=50	0.174	0.118	0.188	0.183	0.214	0.184
n=100	0.154	0.075	0.156	0.086	0.161	0.096
n=200	0.145	0.050	0.145	0.053	0.145	0.051
n=500	0.138	0.030	0.138	0.032	0.139	0.031
n=1000	0.137	0.021	0.137	0.022	0.137	0.022
$\hat{\theta}_{YZ}$						
n=50	0.797	0.125	0.989	1.034	1.282	1.250
n=100	0.887	0.057	1.184	1.357	1.260	1.387
n=200	0.940	0.023	1.119	1.127	1.046	0.989
n=500	0.976	0.006	1.022	0.360	0.984	0.162
n=1000	0.988	0.002	1.012	0.216	0.992	0.085

TABLE B36
Empirical means, standard error for the conditional odds ratio estimator under true model assumption, model selection, and BMA (phi=1.65) for n=50, 100, 200, 500 and 1000 for case 7

	TRUE MODEL		MODEL SELECTION		BMA (phi=1.65)	
$\hat{\theta}_{XY}$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$
n=50	0.879	0.486	1.015	1.696	0.955	1.038
n=100	0.813	0.334	0.875	0.651	0.850	0.313
n=200	0.778	0.227	0.826	0.281	0.820	0.197
n=500	0.757	0.140	0.790	0.186	0.799	0.145
n=1000	0.747	0.098	0.764	0.138	0.782	0.120

TABLE B37
Empirical means, standard error for the conditional odds ratio estimator under true model assumption, model selection, and BMA (phi=1.65) for n=50, 100, 200, 500 and 1000 for case 8

	TRUE MODEL		MODEL SELECTION		BMA (phi=1.65)	
$\hat{\theta}_{XY}$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$	$E(\hat{\theta})$	$se(\hat{\theta})$
n=50	2.429	1.693	2.493	3.575	2.982	5.103
n=100	2.551	1.304	2.743	3.017	2.831	2.931
n=200	2.657	0.994	3.010	3.303	2.734	2.622
n=500	2.709	0.65	2.799	1.315	2.658	0.805
n=1000	2.704	0.451	2.737	0.640	2.710	0.461

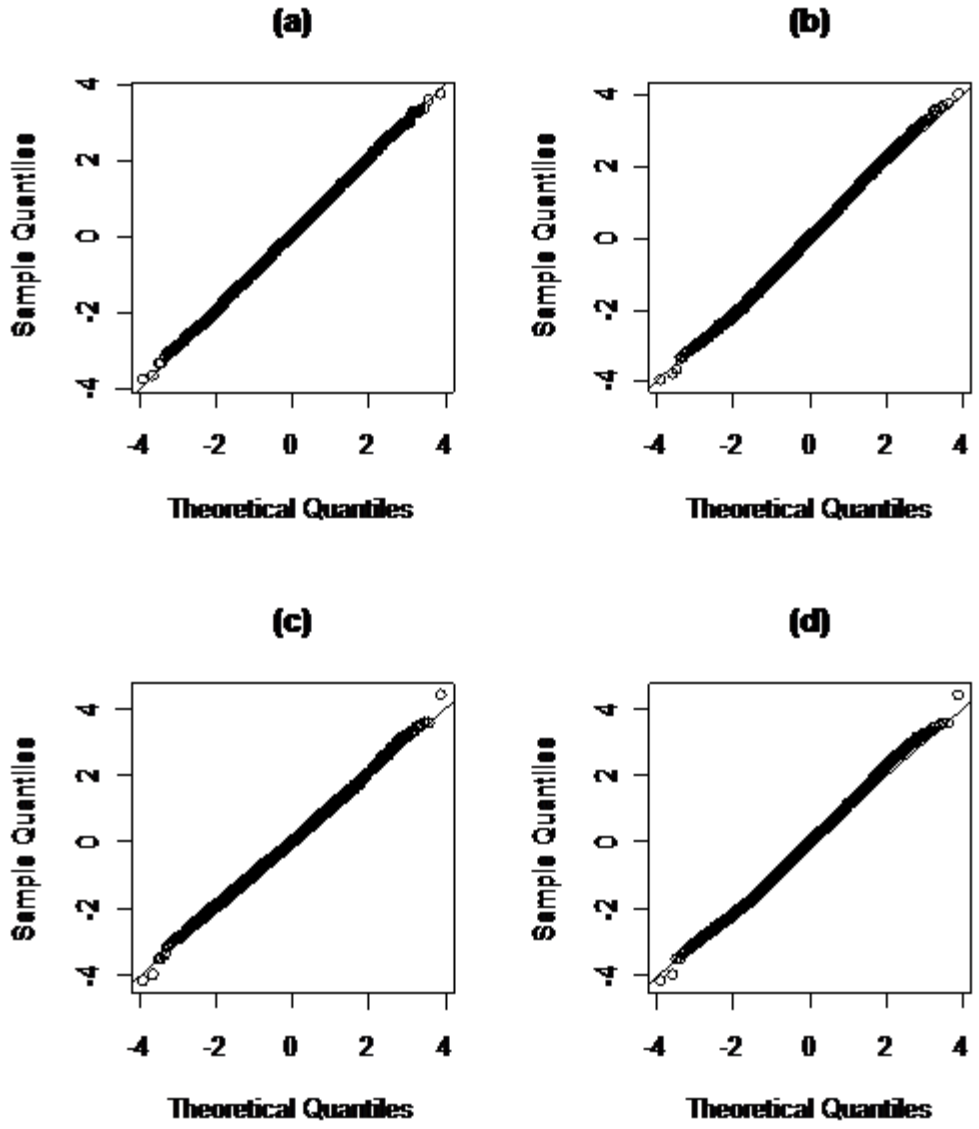

FIGURE B1. Q-Q plot of the Z-statistics for testing parameters (a) λ_{11}^{XY} under true model assumption; (b) λ_{11}^{XY} under model selection condition; (c) λ_{11}^{XZ} under true model assumption; (d) λ_{11}^{XZ} under true model assumption for case 2, n=500.

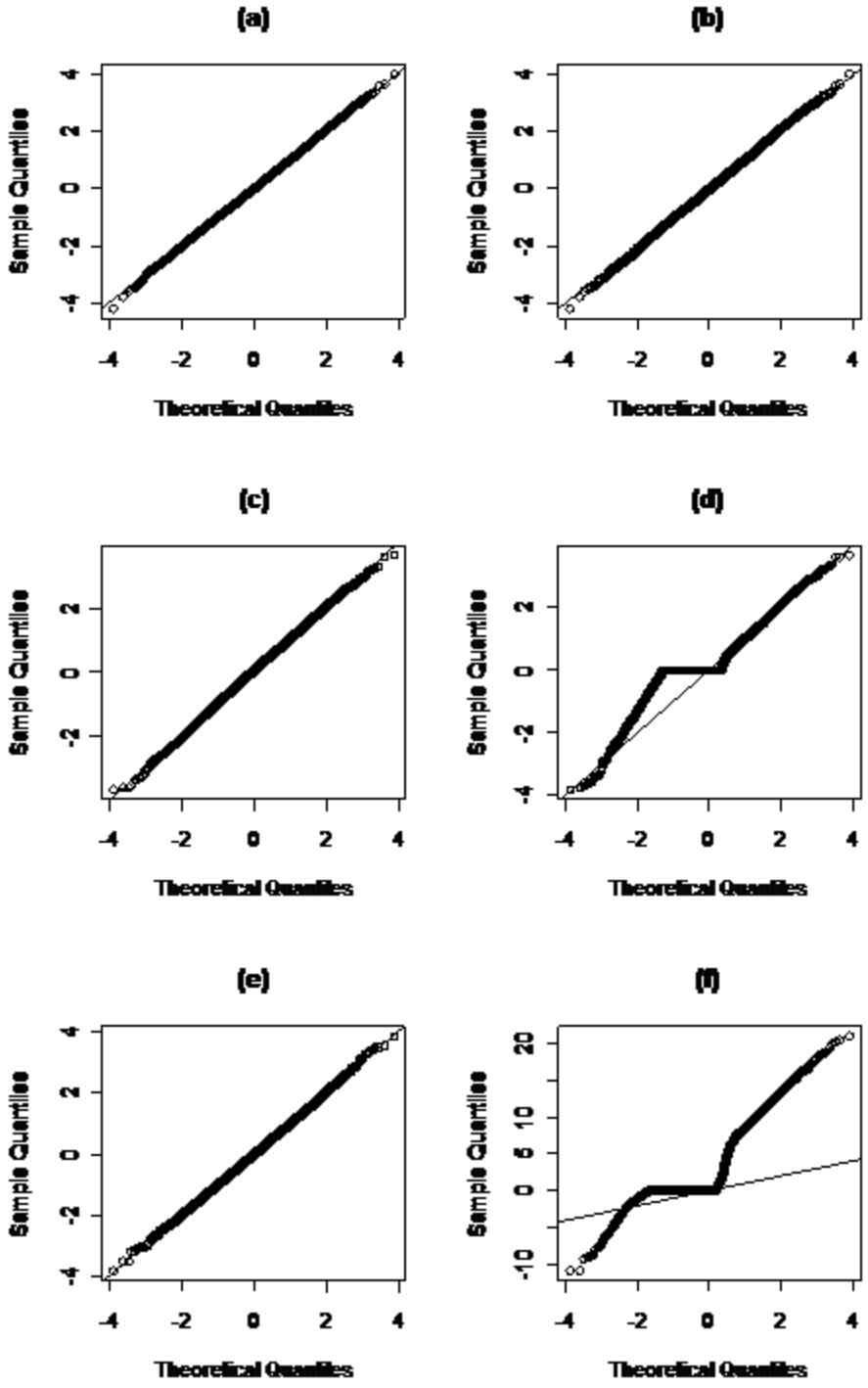

FIGURE B2. Q-Q plot of the Z-statistics for testing parameters (a) λ_{11}^{XY} under true model assumption; (b) λ_{11}^{XY} under model selection condition; (c) λ_{11}^{XZ} under true model assumption; (d) λ_{11}^{XZ} under true model assumption; (e) λ_{12}^{XZ} under true model assumption; (f) λ_{12}^{XZ} under model selection for case 3, n=500.

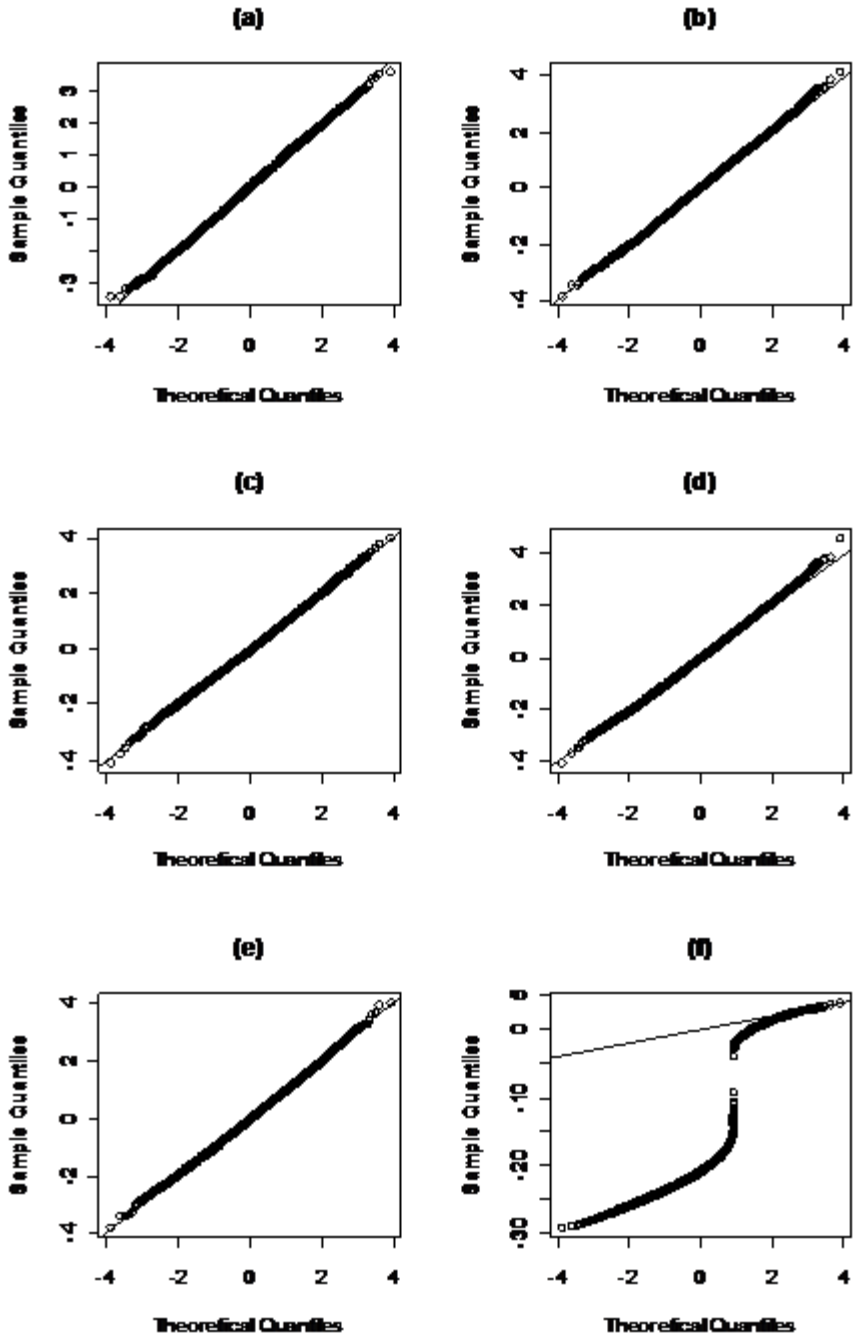

FIGURE B3. Q-Q plot of the Z-statistics for testing parameters (a) λ_{11}^{XY} under true model assumption; (b) λ_{11}^{XY} under model selection condition; (c) λ_{11}^{XZ} under true model assumption; (d) λ_{11}^{XZ} under true model assumption; (e) λ_{12}^{XZ} under true model assumption; (f) λ_{12}^{XZ} under model selection for case 4, n=500.

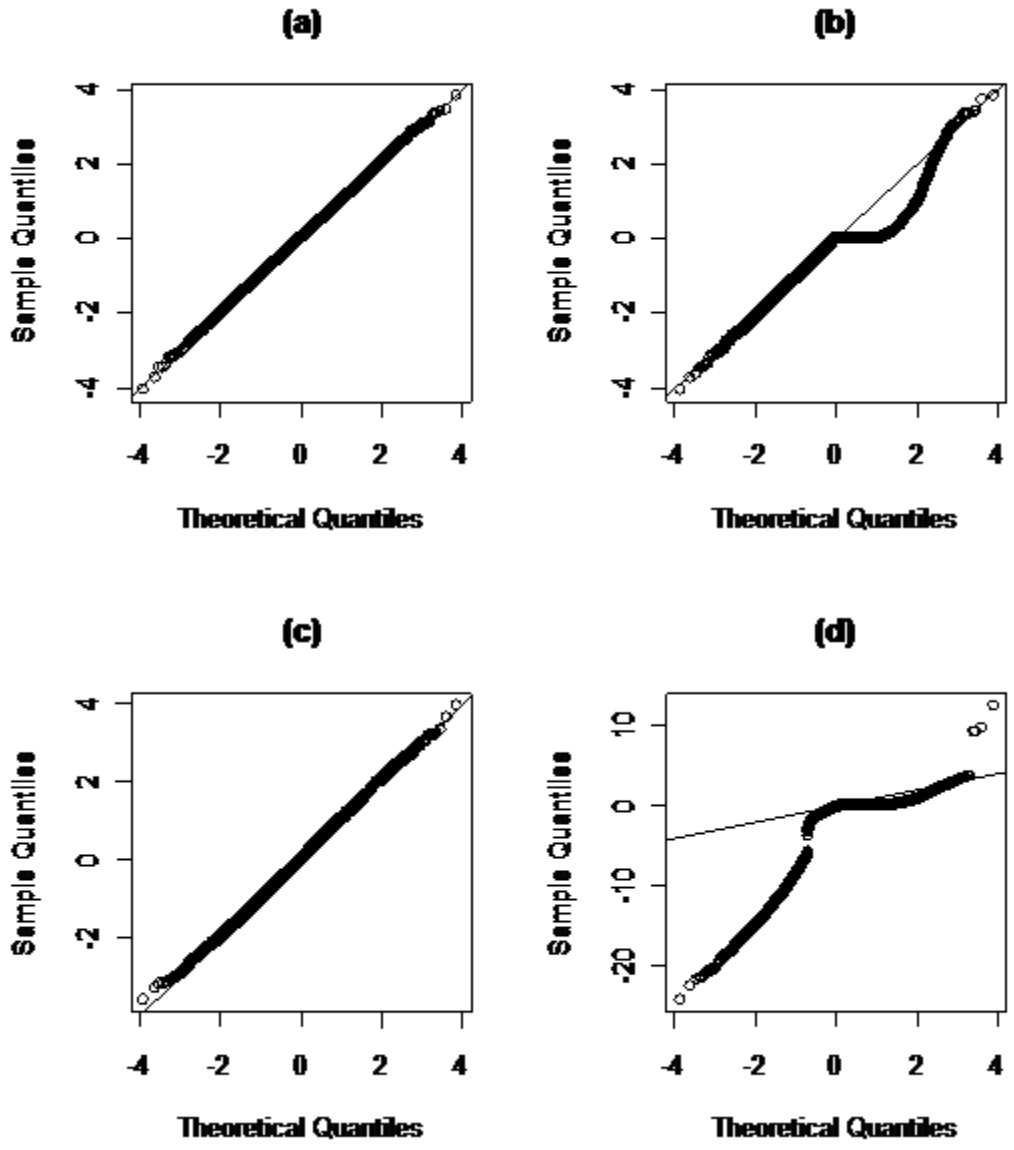

FIGURE B4. Q-Q plot of the Z-statistics for testing parameters (a) λ_{11}^{XY} under true model assumption; (b) λ_{11}^{XY} under model selection condition; (c) λ_{11}^{XZ} under true model assumption; (d) λ_{11}^{XZ} under true model assumption for case 5, n=500.

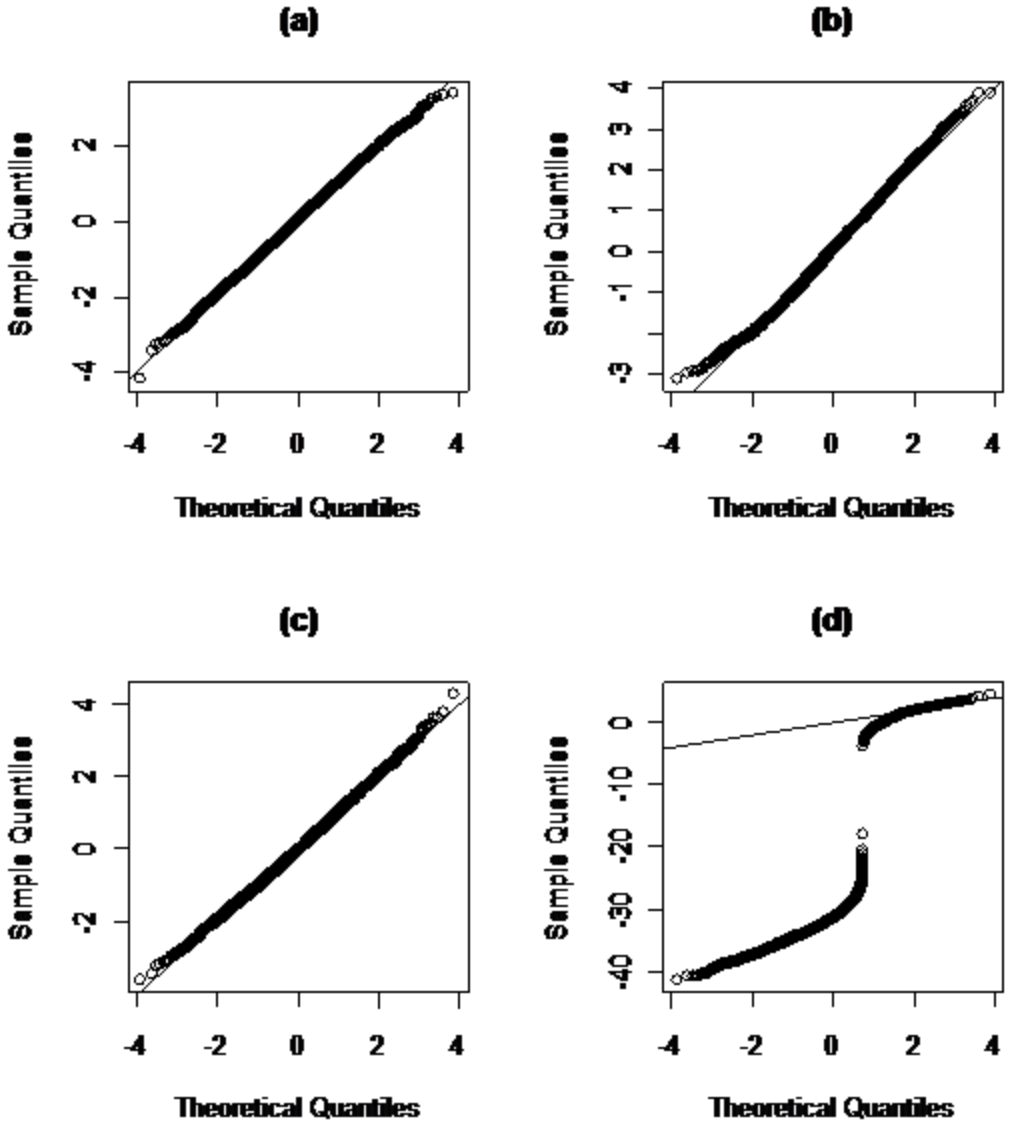

FIGURE B5. Q-Q plot of the Z-statistics for testing parameters (a) λ_{11}^{XY} under true model assumption; (b) λ_{11}^{XY} under model selection condition; (c) λ_{11}^{XZ} under true model assumption; (d) λ_{11}^{XZ} under true model assumption for case 6, n=500.

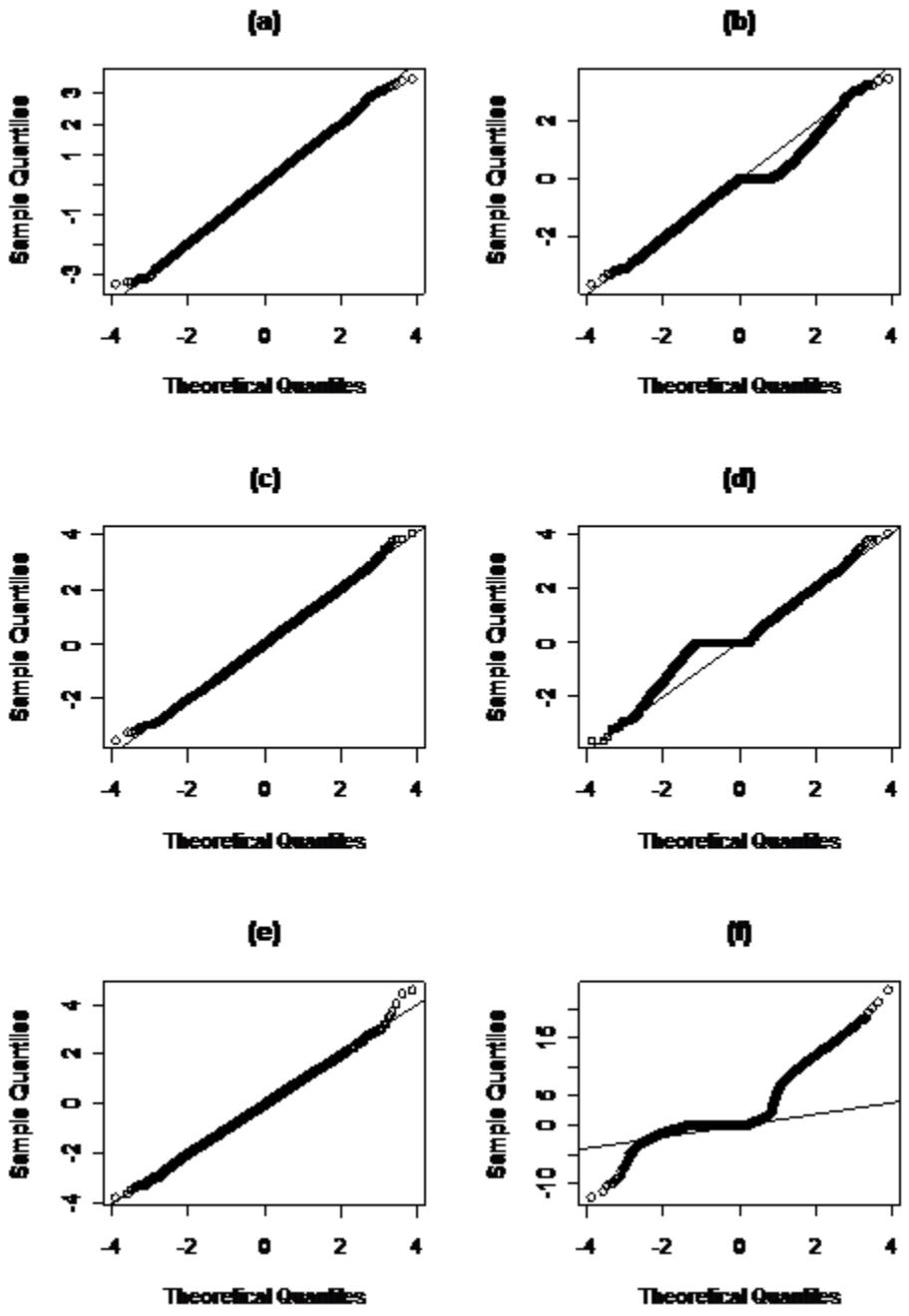

FIGURE B6. Q-Q plot of the Z-statistics for testing parameters (a) λ_{11}^{XY} under true model assumption; (b) λ_{11}^{XY} under model selection condition; (c) λ_{11}^{XZ} under true model assumption; (d) λ_{11}^{XZ} under true model assumption; (e) λ_{12}^{XZ} under true model assumption; (f) λ_{12}^{XZ} under model selection for case 7, n=500.

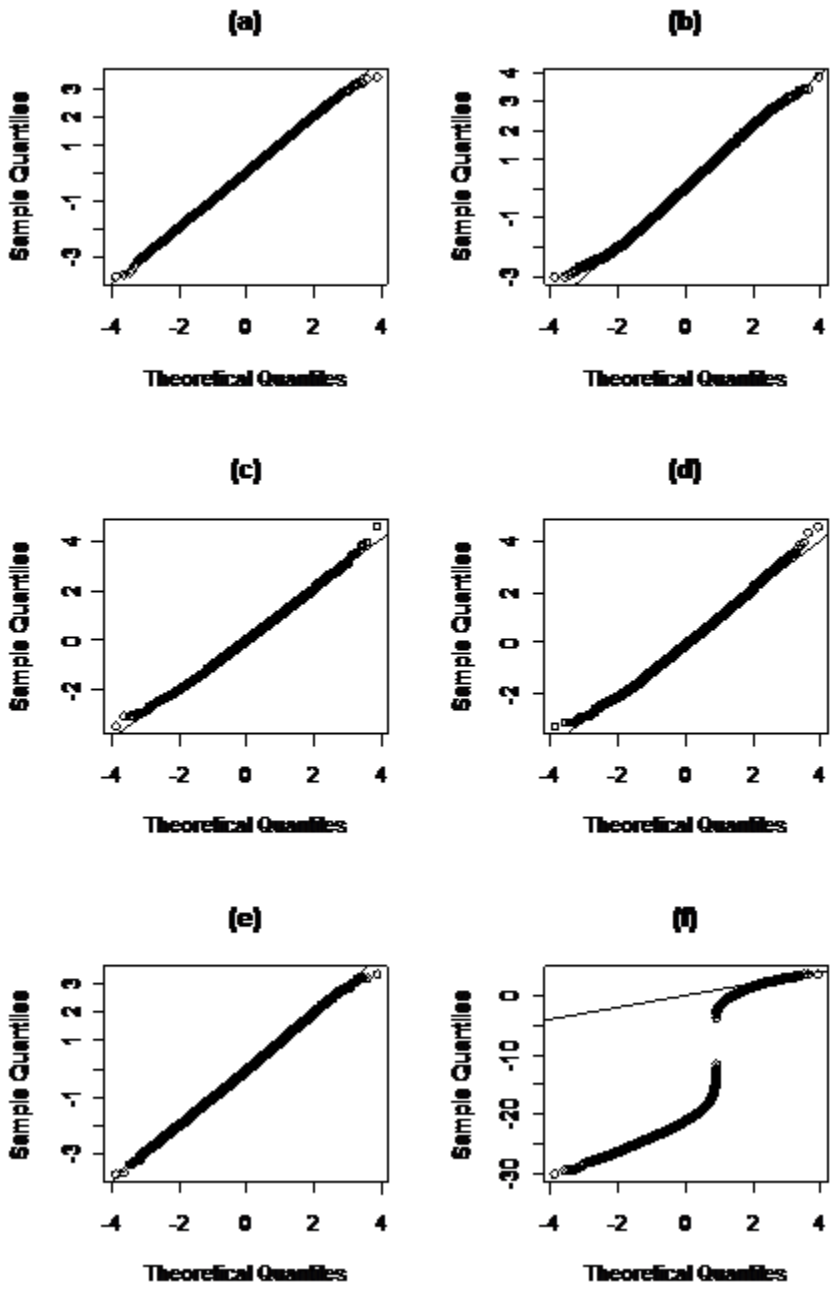

FIGURE B7. Q-Q plot of the Z-statistics for testing parameters (a) λ_{11}^{XY} under true model assumption; (b) λ_{11}^{XY} under model selection condition; (c) λ_{11}^{XZ} under true model assumption; (d) λ_{11}^{XZ} under true model assumption; (e) λ_{12}^{XZ} under true model assumption; (f) λ_{12}^{XZ} under model selection for case 8, n=500.

Appendix C: Pilot Study

Data Generation

All simulation was carried out by the programs written and executed using R and 1,000 iterations were conducted for each condition. The IPF algorithm was used to estimate expected cell frequencies.

This simulation study was built around the Drug data in Agresti (2002)'s work. Agresti (2002, p323) gave an example of the model selection process. The data were from a 1992 survey by the Wright State University and the United Health Services. The Survey asked 2276 final year of high school students in Dayton, Ohio whether they had ever used alcohol, cigarettes, or marirjuana. The respondents were cross-classified by alcohol use, cigarette use, and marijuana use. With three factors, it was easy to look at all possible models. Agresti investigated 9 possible models for the three-way contingency table. He suggested AIC could be used in this type of model comparison. Judging by minimum AIC, the all-two-factor-interaction model fit the best. His data were reproduced in Table C1.

The true model that generated the data in this pilot study were (AC, AM, CM), since this was the model chosen based on the AIC criteria. The sampling scheme was assumed to be multinomial: MN(n,p_{111}, p_{112},p_{121},p_{122},p_{211},p_{212},p_{221},p_{222}). Fienberg (2007) argued that "Few, if any, large-scale sample surveys, …, use simple random samples leading to the multinomial sampling model." In this study, complex survey design was not taken into consideration and a simple random sample was assumed.

The effects used to generate data are from the SAS analysis of drug data based on the model (AC, AM, CM). First, the cell probabilities and cumulative cell probabilities

are calculated (see Table C1). Second, for each observation, a uniform distribution is generated, $U_i \sim Uniform(0,1)$ and is compared to the cumulative cell probabilities. If U_i is smaller than 0.399992619, it is assigned to the cell of "marijuana (Yes) cigarette (Yes) alcohol (Yes)." If U_i is between 0.399992619 and 0.636643234, it is assigned to the cell of "marijuana (No) cigarette (Yes) alcohol (Yes)," etc. Third, for each sample, five possible models are compared: (A,C,M), (AC,M), (AM,CM),(AC, AM,CM), (ACM). The best model was selected by minimizing AIC

AIC= -2(maximized log likelihood – number of parameters in the model)

or, equivalently, minimizing (G^2-2*df). The distributions of the estimated odds ratio were the interest in the pilot study.

TABLE C1
Fitted values, proportions and cumulative proportions of the model (AC,AM, CM)

marijuana	cigarette	alcohol	Model (AC,AM,CM)	proportions	cumulative proportions
Yes	Yes	Yes	910.3832	0.3999926	0.399992619
No	Yes	Yes	538.6168	0.2366506	0.636643234
Yes	No	Yes	44.61683	0.0196031	0.656246410
No	No	Yes	455.3832	0.2000804	0.856326902
Yes	Yes	No	3.61683	0.0015891	0.857916019
No	Yes	No	42.38317	0.0186217	0.876537799
Yes	No	No	1.38317	0.0006077	0.877145518
No	No	No	279.6168	0.1228544	1

Preliminary Results

Preliminary results of this simulation are shown in Table C2. The sample size varies from 50, 100, 200, 500, 1000, to 2000.

The relative bias is defined as the empirical mean under model selection minus the empirical mean under the true model. The standard error ratio is defined as the empirical standard error under model selection divided by the empirical standard error

under the true model. Figure C1 shows the distributions of the two estimators of $\hat{\theta}_{CM}$ when sample size is 500. At this relative large sample size, the disparity of the distribution is very obvious. The distribution of $\hat{\theta}_{CM}$ under model selection had two peaks while the one under true model had only one. In this case, the relative bias was .499, and the standard error ratio was 2.761.

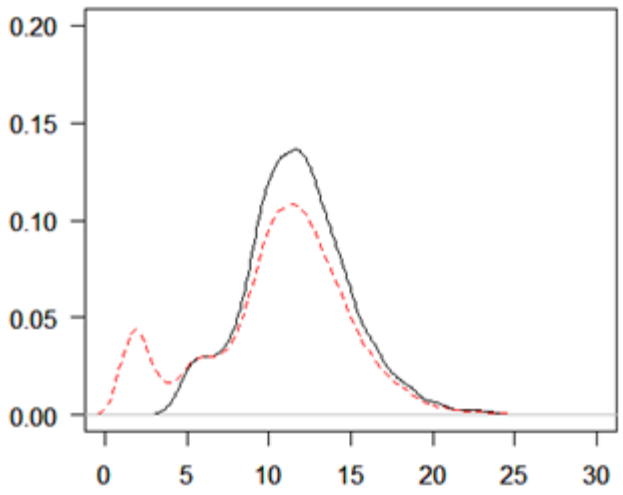

FIGURE C1. The distributions of two odds ratio estimator $\hat{\theta}_{CM}$, post-model-selection estimator (dashed) vs. true model assumption (solid) when sample size equals 500.

Take the odds ratio estimates of association between Alcohol and Cigarettes ($\hat{\theta}_{AC}$) for example. The percentage of selecting the true model ranged from .76 to .93, it did not seem to have a directional relationship with the sample size, i.e., increasing the sample size does not indicate the rate of choosing the correct model rises. The mean of odds ratio estimators (both obtained by model selection and by the true model) went down from around 9.7 to around 7.8 as sample size increases. The relative bias changed

sign unpredictably when sample increases. The s.e. ratio was relatively stable irrespective of sample size.

Odds ratio $\hat{\theta}_{CM}$ showed much more dramatic results. The relative bias increased from .081 to 7.273 when sample size increased from 50 to 2000. The standard error ratio increased from 1.242 to 8.995 as sample size increased from 50 to 2000. This was against the anticipation that similar with linear regression, the difference in the empirical mean would diminish and the difference in the standard error would stay when sample size increases. When sample size was 2000, under model selection, the empirical mean is 21.108 and the standard error 22.281; one would draw a conclusion that this odds ratio was not significantly different from 1. However, at the same sample size, under the true model, the empirical mean was 13.835 and standard error is 2.477; one would draw a conclusion that this odds ratio was significantly different from 1. Consequently, the independence relationship between variable Marijuana and Cigarettes was different under these two conditions.

TABLE C2
Empirical Means, standard error for the conditional odds ratio estimators (model selection vs. no model selection)

sample size	percentage of time a true model is selected	True model		Model selection			
$\hat{\theta}_{AC}$		$E(\hat{\theta}_{AC})$	$se(\hat{\theta}_{AC})$	$E(\hat{\theta}_{AC})$	$se(\hat{\theta}_{AC})$	Difference in empirical means	se ratio
n=50	0.752	9.714	9.629	9.562	9.996	-0.152	1.038
n=100	0.900	10.494	10.730	10.645	11.217	0.151	1.045
n=200	0.936	9.706	9.092	9.916	10.215	0.210	1.124
n=500	0.870	8.343	3.560	8.304	3.680	-0.039	1.034
n=1000	0.816	8.082	2.309	8.071	2.350	-0.011	1.018
n=2000	0.769	7.931	1.536	7.950	1.563	0.019	1.018
$\hat{\theta}_{AM}$		$E(\hat{\theta}_{AM})$	$se(\hat{\theta}_{AM})$	$E(\hat{\theta}_{AM})$	$se(\hat{\theta}_{AM})$		
n=50	0.752	1.983	1.942	1.930	1.970	-0.053	1.014
n=100	0.900	2.965	2.185	2.963	2.269	-0.002	1.038
n=200	0.936	4.787	2.830	4.765	2.940	-0.022	1.040
n=500	0.870	8.858	4.276	8.783	4.761	-0.075	1.113
n=1000	0.816	12.847	5.937	13.489	7.504	0.642	1.264
n=2000	0.769	16.015	7.002	20.425	14.470	4.410	2.067
$\hat{\theta}_{CM}$		$E(\hat{\theta}_{CM})$	$se(\hat{\theta}_{CM})$	$E(\hat{\theta}_{CM})$	$se(\hat{\theta}_{CM})$		
n=50	0.752	7.749	6.435	7.830	7.993	0.081	1.242
n=100	0.900	9.127	8.496	9.907	12.841	0.780	1.511
n=200	0.936	9.242	7.196	10.184	15.974	0.942	2.220
n=500	0.870	10.045	3.900	10.544	10.769	0.499	2.761
n=1000	0.816	11.800	3.306	14.305	15.093	2.505	4.565
n=2000	0.769	13.835	2.477	21.108	22.281	7.273	8.995

Appendix D: R Code

Pilot Study

```
library(MASS)
repnumber=10000
theta.select=c()
theta=c()
theta2=c()
theta2.select=c()
theta3=c()
theta3.select=c()
counter=0
samplesize=100
for (j in 1:repnumber){
U=c()
c1=0;c2=0;c3=0;c4=0;c5=0;c6=0;c7=0;c8=0

for (i in 1:samplesize){
U[i]=runif(1,0,1)
if (U[i]>=0 & U[i]<=0.399992619) c1=c1+1
if(U[i]>0.399992619 & U[i]<=0.636643234) c2=c2+1
if (U[i]>0.636643234 & U[i]<=0.65624641) c3=c3+1
if (U[i]>0.65624641 & U[i]<=0.856326902)  c4=c4+1
if (U[i]>=0.856326902 & U[i]<=0.857916019) c5=c5+1
if(U[i]>0.857916019 & U[i]<=0.876537799) c6=c6+1
if (U[i]>0.876537799 & U[i]<=0.877145518) c7=c7+1
if (U[i]>0.877145518 & U[i]<=1)  c8=c8+1
}
#c1;c2;c3;c4;c5;c6;c7;c8

table.1=data.frame(expand.grid(marijuana=factor(c("Yes","No"),levels=c("No","Yes")),
                               cigarette=factor(c("Yes","No"),levels=c("No","Yes")),
                              alcohol=factor(c("Yes","No"),levels=c("No","Yes"))),
                             count=c(c1,c2,c3,c4,c5,c6,c7,c8))
```

```
fitACM=loglm(count~alcohol*cigarette*marijuana,data=table.1,param=T,fit=T) #ACM--------model  1
fitAC.AM.CM=update(fitACM,.~.-alcohol:cigarette:marijuana)                 #AC,AM,CM---model  2
fitAM.CM=update(fitAC.AM.CM,.~.-alcohol:cigarette)                         #AM,CM------model  3
fitAC.M=update(fitAC.AM.CM,.~.-alcohol:marijuana-cigarette:marijuana)      #AC,M-------model  4
fitA.C.M=update(fitAC.M,.~.-alcohol:cigarette)                             #A,C,M------model  5

m1=fitACM$deviance-2*fitACM$df                          #calculate AIC's
m2=fitAC.AM.CM$deviance-2*fitAC.AM.CM$df
m3=fitAM.CM$deviance-2*fitAM.CM$df
m4=fitAC.M$deviance-2*fitAC.M$df
m5=fitA.C.M$deviance-2*fitA.C.M$df

aic=c(m1,m2,m3,m4,m5)
ind=which.min(aic)     #find the index of the minimum of the vector aic

                                                    #compute theta of AC of this best fit model
if (ind==1){f.1=c(aperm(fitted(fitACM))); theta.select[j]=(f.1[1]+.5)*(f.1[7]+.5)/((f.1[3]+.5)*(f.1[5]+.5));
                                         theta2.select[j]=(f.1[1]+.5)*(f.1[6]+.5)/((f.1[2]+.5)*(f.1[5]+.5))
                                         theta3.select[j]=(f.1[1]+.5)*(f.1[4]+.5)/((f.1[2]+.5)*(f.1[3]+.5))}

if
(ind==2){f.2=c(aperm(fitted(fitAC.AM.CM)));theta.select[j]=(f.2[1]+.5)*(f.2[7]+.5)/((f.2[3]+.5)*(f.2[5]+.5));
                                           counter=counter+1;

      theta2.select[j]=(f.2[1]+.5)*(f.2[6]+.5)/((f.2[2]+.5)*(f.2[5]+.5))

      theta3.select[j]=(f.2[1]+.5)*(f.2[4]+.5)/((f.2[2]+.5)*(f.2[3]+.5))}

if (ind==3){f.3=c(aperm(fitted(fitAM.CM)));theta.select[j]=(f.3[1]+.5)*(f.3[7]+.5)/((f.3[3]+.5)*(f.3[5]+.5))
                                         theta2.select[j]=(f.3[1]+.5)*(f.3[6]+.5)/((f.3[2]+.5)*(f.3[5]+.5))
                                         theta3.select[j]=(f.3[1]+.5)*(f.3[4]+.5)/((f.3[2]+.5)*(f.3[3]+.5))}

if (ind==4){f.4=c(aperm(fitted(fitAC.M)));theta.select[j]=(f.4[1]+.5)*(f.4[7]+.5)/((f.4[3]+.5)*(f.4[5]+.5))
                                         theta2.select[j]=(f.4[1]+.5)*(f.4[6]+.5)/((f.4[2]+.5)*(f.4[5]+.5))
                                         theta3.select[j]=(f.4[1]+.5)*(f.4[4]+.5)/((f.4[2]+.5)*(f.4[3]+.5))}

if (ind==5){f.5=c(aperm(fitted(fitA.C.M)));theta.select[j]=(f.5[1]+.5)*(f.5[7]+.5)/((f.5[3]+.5)*(f.5[5]+.5))
```

```
                                        theta2.select[j]=(f.5[1]+.5)*(f.5[6]+.5)/((f.5[2]+.5)*(f.5[5]+.5))
                                        theta3.select[j]=(f.5[1]+.5)*(f.5[4]+.5)/((f.5[2]+.5)*(f.5[3]+.5))}

d.2=c(aperm(fitted(fitAC.AM.CM)))
theta[j]=(d.2[1]+.5)*(d.2[7]+.5)/((d.2[3]+.5)*(d.2[5]+.5))
theta2[j]=(d.2[1]+.5)*(d.2[6]+.5)/((d.2[2]+.5)*(d.2[5]+.5))
theta3[j]=(d.2[1]+.5)*(d.2[4]+.5)/((d.2[2]+.5)*(d.2[3]+.5))
}
counter
mean(theta);var(theta)
mean(theta.select);var(theta.select)
mean(theta2);var(theta2)
mean(theta2.select);var(theta2.select)
mean(theta3);var(theta3)
mean(theta3.select);var(theta3.select)
plot(density(theta),lwd=1,lty=1,col="black",ann=FALSE, las=1,ylim=c(0,2),xlim=c(-2,2))
plot(density(theta.select),lwd=1,lty=2,col="red",ann=FALSE, las=1, ylim=c(0,30),xlim=c(0,5))
```

Classical Analysis of Model (AM,CM,AC) of the Drug Data

```
table.8.3<-data.frame(expand.grid(
 marijuana=factor(c("Yes","No"),levels=c("No","Yes")),
 cigarette=factor(c("Yes","No"),levels=c("No","Yes")),
 alcohol=factor(c("Yes","No"),levels=c("No","Yes"))),
count=c(911,538,44,456,3,43,2,279))

options(contrasts=c("contr.treatment","contr.poly"))
fit.glm<-glm(count~.^2, data=table.8.3, family=poisson)
```

BMA Sensitivity Analysis of the Drug Data

```
library(MASS)
library(BMA)
#####################SET FOUR: uniform priors ############

# columns correspond to C, M, A, CM, AC, AM, AMC
x=rbind(
      c(1,1,1,1,1,1,1),
      c(1,0,1,0,1,0,0),
      c(0,1,1,0,0,1,0),
      c(0,0,1,0,0,0,0),
      c(1,1,0,1,0,0,0),
      c(1,0,0,0,0,0,0),
      c(0,1,0,0,0,0,0),
      c(0,0,0,0,0,0,0))
n=c(1,1,1,1,1,1,1,1)
y=c(911,538,44,456,3,43,2,279)
model9=rbind(
c(1,1,1,0,0,0,0),  #model (A,C,M)
c(1,1,1,0,1,0,0),  #model (M,AC)
c(1,1,1,1,0,0,0),  #model (A,CM)
c(1,1,1,0,0,1,0),  #model (C,AM)
c(1,1,1,1,0,1,0),  #model (AM,CM)
c(1,1,1,0,1,1,0),  #model (AC,AM)
c(1,1,1,1,1,0,0),  #model (AC,CM)
c(1,1,1,1,1,1,1),  #model (ACM)
c(1,1,1,1,1,1,0)   #model (AC,AM,CM)
)
glib.drug <- glib(x,y,n,error="poisson",link="log",models=model9)

glib.drug$glim.est$coef
glib.drug$inputs$phi
glib.drug$bf$postprob

#####################SET ONE: based on problem-behavior theory
############

# columns correspond to C, M, A, CM, AC, AM, AMC
x=rbind(
      c(1,1,1,1,1,1,1),
      c(1,0,1,0,1,0,0),
      c(0,1,1,0,0,1,0),
      c(0,0,1,0,0,0,0),
      c(1,1,0,1,0,0,0),
```

```
      c(1,0,0,0,0,0,0),
      c(0,1,0,0,0,0,0),
      c(0,0,0,0,0,0,0))
n=c(1,1,1,1,1,1,1,1)
y=c(911,538,44,456,3,43,2,279)
model9=rbind(
c(1,1,1,0,0,0,0),  #model (A,C,M)
c(1,1,1,0,1,0,0),  #model (M,AC)
c(1,1,1,1,0,0,0),  #model (A,CM)
c(1,1,1,0,0,1,0),  #model (C,AM)
c(1,1,1,1,0,1,0),  #model (AM,CM)
c(1,1,1,0,1,1,0),  #model (AC,AM)
c(1,1,1,1,1,0,0),  #model (AC,CM)
c(1,1,1,1,1,1,1),  #model (ACM)
c(1,1,1,1,1,1,0)   #model (AC,AM,CM)
)
glib.drug <- glib(x,y,n,error="poisson",link="log",
pmw=c(1,1.3,1.3^2,1.3^3,1.3^4,1.3^5,1.3^6,1.3^7,1.3^8),models=model9)

glib.drug$bf$postprob

####################SET TWO: based on lifestyle theory #############
model9=rbind(
c(1,1,1,0,0,0,0),  #model (A,C,M)
c(1,1,1,0,1,0,0),  #model (M,AC)
c(1,1,1,1,0,0,0),  #model (A,CM)
c(1,1,1,0,0,1,0),  #model (C,AM)
c(1,1,1,0,1,1,0),  #model (AC,AM)
c(1,1,1,1,1,0,0),  #model (AC,CM)
c(1,1,1,1,1,1,1),  #model (ACM)
c(1,1,1,1,1,1,0),  #model (AC,AM,CM)
c(1,1,1,1,0,1,0)   #model (AM,CM)
)

glib.drug <- glib(x,y,n,error="poisson",link="log",
pmw=c(1,1.3,1.3^2,1.3^3,1.3^4,1.3^5,1.3^6,1.3^7,1.3^8),models=model9)

glib.drug$bf$postprob


################  SET THREE: based on lifestyle theory###############
model9=rbind(
c(1,1,1,0,0,0,0),  #model (A,C,M)
c(1,1,1,1,0,1,0),  #model (AM,CM)
c(1,1,1,0,0,1,0),  #model (C,AM)
c(1,1,1,1,1,1,0),  #model (AC,AM,CM)
c(1,1,1,0,1,1,0),  #model (AC,AM)
c(1,1,1,1,1,1,1),  #model (ACM)
c(1,1,1,0,1,0,0),  #model (M,AC)
c(1,1,1,1,1,0,0),  #model (AC,CM)
c(1,1,1,1,0,0,0)   #model (A,CM)
)

glib.drug <- glib(x,y,n,error="poisson",link="log",
pmw=c(1,1.3,1.3^2,1.3^3,1.3^4,1.3^5,1.3^6,1.3^7,1.3^8),models=model9)

glib.drug$bf$postprob
```

Model Weights via Bootstrapping Model Averaging

```
library(MASS)
countdata=c(911,538,44,456,3,43,2,279)
cumsum(countdata)
f1=0;f2=0;f3=0;f4=0;f5=0;f6=0;f7=0;f8=0;f9=0    #frequencies
for ( i in 1:1000){
this.ind<-sample(2276,2276,replace=TRUE)
#calculate the cell frequencies of this bootstrape sample
c1=length(this.ind[this.ind<=911])
c2=length(this.ind[911<this.ind & this.ind<=1449])
c3=length(this.ind[1449<this.ind & this.ind<=1493])
c4=length(this.ind[1493<this.ind & this.ind<=1949])
c5=length(this.ind[1949<this.ind & this.ind<=1952])
c6=length(this.ind[1952<this.ind & this.ind<=1995])
c7=length(this.ind[1995<this.ind & this.ind<=1997])
c8=length(this.ind[1997<this.ind & this.ind<=2276])

table.1=data.frame(expand.grid(
marijuana=factor(c("Yes","No"),levels=c("No","Yes")),
         cigarette=factor(c("Yes","No"),levels=c("No","Yes")),
alcohol=factor(c("Yes","No"),levels=c("No","Yes"))),
count=c(c1,c2,c3,c4,c5,c6,c7,c8))

fitACM=glm(count~alcohol*cigarette*marijuana,data=table.1,family=poisso
n(link=log))                                                    #ACM
fitAC.AM.CM=update(fitACM,.~.-alcohol:cigarette:marijuana)#AC,AM,CM
fitAM.CM=update(fitAC.AM.CM,.~.-alcohol:cigarette)        #AM,CM
fitAM.AC=update(fitAC.AM.CM,.~.-cigarette:marijuana)      #AM,AM
fitAC.CM=update(fitAC.AM.CM,.~.-alcohol:marijuana)        #AC,CM
fitAC.M=update(fitAC.AM.CM,.~.-alcohol:marijuana-cigarette:marijuana)
#AC,M
fitAM.C=update(fitAC.AM.CM,.~.-alcohol:cigarette-cigarette:marijuana)
#AM,C
fitCM.A=update(fitAC.AM.CM,.~.-alcohol:cigarette-alcohol:marijuana)
#CM,A
fitA.C.M=update(fitAC.M,.~.-alcohol:cigarette)
#A,C,M
m1=fitACM$aic
m2=fitAC.AM.CM$aic
m3=fitAM.CM$aic
m4=fitAM.AC$aic
m5=fitAC.CM$aic
m6=fitAC.M$aic
m7=fitAM.C$aic
m8=fitCM.A$aic
m9=fitA.C.M$aic
aic=c(m1,m2,m3,m4,m5,m7,m8,m9)
ind=which.min(aic)
if (ind==1){f1=f1+1}
if (ind==2){f2=f2+1}
if (ind==3){f3=f3+1}
if (ind==4){f4=f4+1}
if (ind==5){f5=f5+1}
```

```
if (ind==6){f6=f6+1}
if (ind==7){f7=f7+1}
if (ind==8){f8=f8+1}
if (ind==9){f9=f9+1}}
fre=c(f1,f2,f3,f4,f5,f6,f7,f8,f9)
weight=fre/sum(fre)
```

Model Weights via AIC Approximation

```
countdata=c(911,538,44,456,3,43,2,279)
table.1=data.frame(expand.grid(
marijuana=factor(c("Yes","No"),levels=c("No","Yes")),
cigarette=factor(c("Yes","No"),levels=c("No","Yes")),
alcohol=factor(c("Yes","No"),levels=c("No","Yes"))),
count=countdata)

fitACM=glm(count~alcohol*cigarette*marijuana,data=table.1,family=poisso
n(link=log))                                                        #ACM
fitAC.AM.CM=update(fitACM,.~.-alcohol:cigarette:marijuana)  #AC,AM,CM
fitAM.CM=update(fitAC.AM.CM,.~.-alcohol:cigarette)          #AM,CM
fitAM.AC=update(fitAC.AM.CM,.~.-cigarette:marijuana)        #AM,AM
fitAC.CM=update(fitAC.AM.CM,.~.-alcohol:marijuana)          #AC,CM
fitAC.M=update(fitAC.AM.CM,.~.-alcohol:marijuana-cigarette:marijuana)
#AC,M
fitAM.C=update(fitAC.AM.CM,.~.-alcohol:cigarette-cigarette:marijuana)
#AM,C
fitCM.A=update(fitAC.AM.CM,.~.-alcohol:cigarette-alcohol:marijuana)
#CM,A
fitA.C.M=update(fitAC.M,.~.-alcohol:cigarette)
#A,C,M
m1=fitACM$aic
m2=fitAC.AM.CM$aic
m3=fitAM.CM$aic
m4=fitAM.AC$aic
m5=fitAC.CM$aic
m6=fitAC.M$aic
m7=fitAM.C$aic
m8=fitCM.A$aic
m9=fitA.C.M$aic
aic=c(m1,m2,m3,m4,m5,m7,m8,m9)
weight=exp(-aic/2)
normedwt=weight/sum(weight)
```

References

Agresti, A. (2002). Categorical Data Analysis. New York: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Albert, J. H. (1989). A Bayesian test for a two-way contingency table using independence priors. *Canadian Journal of Statistics*, 18, 347-363.

Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics*, 15, 190-204.

Bancroft, T. A., & Han, C. P. (1977). Inference based on conditional specification: a note and a bibliography. *International Statistical Review*, 45, 117-127.

Bartels, L. M. (1997). Specification uncertainty and model averaging. *American Journal of Political Science*, 41, 641-674.

Bishop, Y. M. M., Fienberg, S. E. , & Holland, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Massachusetts: MIT Press.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tales. *Journal of the Royal Statistical Society, Series B*, 25, 220-233.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, 53, 603-618.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer, New York.

Burnham, K. P., & Anderson, D. R. (2004). Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33, 261-304.

Candolo, C., Davison, A. C., & Demetrio, C. G. B. (2003). A note on model uncertainty in linear regression. *The Statistician*, 52, 165-177.

Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference. *Journal of the Royal Statistical Society, Series A*, 158, 419-466.

Christensen, R. (1997). *Log-linear Models and Logistic Regression.* New York: Springer.

Christensen, R. (1990). *Log-linear Models.* New York: Springer.

Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics*, 11, 745-763.

Coffey, C., Carlin, J. B., Lynskey, M., Li, N., & Patton, G. C. (2003). Adolescent precursors of cannabis dependence: Findings from the Victorian adolescent health cohort study. *British Journal of Psychiatry*, 182, 330-336.

Danilov, D., & Magnus, J.R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122, 27-46.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

Donovan, J. E., & Jessor, R. (1985). Structure of problem behavior in adolescence and young adulthood. *Journal of Consulting and Clinical Psychology*, 53, 890-904.

Efron, B., & Gong, G. (1983). A leisurely look at bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37, 36-49.

Everitt, B. S. (1977). *The Analysis of Contingency Tables*. New York: John Wiley and Sons.
Evett, I. W. (1991). Implementing Bayesian methods in forensic science. Paper presented to the Fourth International Meeting on Bayesian Statistics.
Feinstein, A. R. (1973). Clinical biostatistics. XX. The epidemiologic trohoc, the ablative risk ratio, and "retrospective" research. *Clinical Pharmacology and Therapeutics*, 14, 291-307.
Fernandez, C., Ley, E., & Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16, 563-576.
Fienberg, S. E. (2007). *The Analysis of Cross-Classified Categorical Data*. New York: Springer.
Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27-38.
Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, 37, 152-155.
Gao, J., & Dayton, C. (2008). Finite sample properties of post-model-selection estimators in linear regression: An empirical investigation. Poster session presented at the annual meeting of the AERA, New York, NY.
Gart, J. J. (1966). Alternative analyses of contingency tables. *Journal of Royal Statistical Society, Series B*, 28, 164-179.
Gart, J. J., & Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biomentrika*, *54*, 181-187.
Geer, J., & Lau, R. R. (2006). Filling the blanks: A new method for estimating campaign effects. *British Journal of Political Science*, 36, 269-290.
Gilbert, G. N. (1981). Modelling Society: An introduction to loglinear analysis for social researchers. London, U.K.: George Allen & Unwin.
Good, I. J., & Crook, J. F. (1987). The robustness and sensitivity of the mixed-Dirichlet Bayesian test for "independence" in contingency tables. *Annals of Statistics*, 15, 670-693.
Haldane, J. B. (1956). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics, 20*, 309-311.
Hoeting, J., Madigan, D., Raftery, A. E. , & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, 14, 382-417.
Huntsberger, D.V. (1955). A generalization of a preliminary testing procedure for pooling data. *Annals of Mathematical Statistics*, 26, 734-743.
Hurvich, C. M., & Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214-217.
Jefferys, W. H., & Berger, J. O. (1991). Ockham's razor and Bayesian statistics. *American Scientist*, 80, 64-72.
Judge, G. G., & Bock, M. E. (1978). *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*. North-Holland, Amsterdam.
Kabaila, P. (2005). On the coverage probability of confidence interval in regression after variable selection. *Australian & New Zealand Journal of Statistics*, 47, 549-562.
Kandel, D. B., & Faust, R. (1975). Sequence and stages in patterns of adolescent drug use. *Archives of General Psychiatry*, 32, 923-932.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Knoke, D. P., & Burke, P. J. (1980). Log-linear Models. Beverly Hills, CA: Sage Publications.

Lahiri, P. (Ed.). (2001). *Model Selection IMS Lecture Notes Monograph Series 38*. Beachwood, OH: Lecture Notes-Monograph Series, Institute of Mathematical Statistics.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, 21, 21-59.

Linhart, H., & Zucchini, W. (1986). *Model Selection*. New York: John Wiley.

Madigan, D., & Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1535-1546.

Martin, M., & Roberts, S. (2006). Bootstrap model averaging in time series studies of particular matter air pollution and mortality. *Journal of Exposure Science and Environmental Epidemiology*, 16, 242-250.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models (2nd ed.). Boca Raton: Chapman & Hall/CRC.

McQuarrie, A., & Tsai, C. (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific Publishing Company.

Miller, A. J. (2002). *Subset Selection in Regression*. (2nd edition). London, Chapman and Hall.

Morales, K. H., Ibrahim, J.G., Chen, C. J., & Ryan, L.M. (2006). Bayesian Model Averaging with applications to benchmark dose estimation for Arsenic in Drinking Water. *Journal of the American Statistical Association*, 101, 9-17.

Mosteller, F. (1948). On pooling data. *Journal of the American Statistical Association*, 43, 231-243.

Moulton, B. R. (1991). A Bayesian approach to regression selection and estimation with application to a price index for radio services. *Journal of Econometrics*, 49, 169-193.

Pardo, L., & Pardo, M. D. C. (2003). Minimum power-divergence estimator in three-way contingency tables. *Journal of Statistical Computation and Simulation*, 73, 819-831.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7,163-185.

Pötscher, B. M., & Novák, A. J. (1998). The distribution of estimators after model selection: large and small sample results. *Journal of Statistical Computation and Simulation*, 60, 19-56.

Raftery, A. E. (1986). Choosing models for cross-classifications, *American Sociological Review*, 51, 145-146.

Raftery, A. E. (1995). Rejoinder: Model selection is unavoidable in social research. *Sociological Methodology*, 25, 185-195.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83, 251-266.

Reynolds, H. T. (1977). The Analysis of Cross-classifications. New York: The Free Press.

Rob, M., Reynolds, I., & Finlayson, P. F. (1990). Adolescent marijuana use: Risk factors and implications. *Australian and New Zealand Journal of Psychiatry*, 24, 47-56.

Rosay, A., Najaka, S., & Herz, D. (2007). Differences in the validity of self-reported drug use across five factors: gender, race, age, type of drug, and offense seriousness. *Journal of Quantitative Criminology*, 23, 41-58.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Sclove, S. L., Morris, C., & Radhakrishnan, R. (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, 43, 1481-1490.

Sen, P. K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics*, 7, 1019-1033.

Shaddick, G., Wakefield, J., & Elliott, P. (1998). Estimating the association between air pollution and daily mortality using Bayesian model averaging. *Epidemiology*, 9, S-60.

Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine*, 2, 71-78.

Thompson, L. (2009). *R (and S-PLUS) Manual to Accompany Agresti's Categorical Data Analysis (2002)*. Retrieved August 21, 2009 from the World Wide Web: https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf

Tukey, J. W. (1980). Methodological comments focused on opportunities. In Jones, L.V. (Ed.). *The Collected Works of John W. Tukey*. Murray Hill, NJ: Bell Telephone Laboratories.

Viallefont, V., Raftery, A. E., & Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics of Medicine*, 20, 3215-3230.

Waldorp, L., & Wagenmakers, E. (ed.). (2006). *Journal of Mathematical Psychology Special Issue on Model Selection: Theoretical Developments and Applications*, 50, 99-214.

Yang, Z., & Xie, M. (2003). Efficient estimation of the Weibull shape parameter based on a modified profile likelihood. *Journal of Statistical Computation and Simulation*, 73, 115-123.

Yamaguchi, K., & Kandel, D. B. (1984). Patterns of drug use from adolescence to young adulthood: II. Sequences of progression. *American Journal of Public Health*, 74, 668-672.

Zhang, P.(1992), Influence after variable selection in linear regression models. *Biometrika*, 79, 741-746.