11-30-2007

Evaluating Variance of the Model Credibility Index

Yan Xiao

EVALUATING VARIANCE OF THE MODEL CREDIBILITY INDEX

By

Yan Xiao

Under the Direction of Dr. Jiawei Liu and Dr.Yu-Sheng Hsu

ABSTRACT

Model credibility index is defined to be a sample size under which the power of rejection equals 0.5. It applies goodness-of-fit testing thinking and uses a one-number summary statistic as an assessment tool in a false model world. The estimation of the model credibility index involves a bootstrap resampling technique. To assess the consistency of the estimator of model credibility index, we instead study the variance of the power achieved at a fixed sample size. An improved subsampling method is proposed to obtain an unbiased estimator of the variance of power. We present two examples to interpret the mechanics of building model credibility index and estimate its error in model selection. One example is two-way independent model by Pearson Chi-square test, and another example is multi-dimensional logistic regression model using likelihood ratio test.

INDEX WORDS: Model credibility index, Goodness-of-fit test, Bootstrap resampling, Consistency of the estimation

EVALUATING VARIANCE OF THE MODEL CREDIBILITY INDEX

by

Yan Xiao

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2007

EVALUATING VARIANCE OF THE MODEL CREDIBILITY INDEX

by

Yan Xiao

Committee Chair: Yu-Sheng Hsu

Jiawei Liu

Committee: Yichuan Zhao

Electronic Version Approved:

Office of Graduate Studies
College of Art and Sciences
Georgia State University
December 2007

ACKNOWLEDGEMENTS

First and foremost, I am so pleased to have such an opportunity to acknowledge my thesis advisor, Dr. Jiawei Liu for her directing my thesis work with great patience and valuable support. I thank my thesis advisor Dr. Yu-Sheng Hsu for his reviewing my thesis and making important revised suggestions throughout the whole thesis. I thank my committee member Dr. Yichuan Zhao for his spending time to read my thesis and providing useful comments.

I also like to thank all my professors and friends whose constant support and encouragement made the past two years a joyful experience during my graduate education.

Finally, I thank my husband and my son for their understanding, support and love throughout my graduate study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Basic ideas

Models derived from data provide insights for understanding certain phenomena and make predictions (Giudici, 2003, Hand, 2000) [1] [2]. In order to understand the data, the goodness-of-fit of the model plays important roles. However, in a real data analysis, one will often obtain a wrong model. The distance between the fitted and the true models can be measured with statistical distance. The fitted model is usually a convenient conceptual representation of the observed phenomenon (Linhart and Zucchini 1986) [3].

A probability model is an abstract mechanism from which one can imagine to generate the data. To fit a model, one needs to replace the empirical distribution with a theoretical probability distribution. A probability model M is an adequate approximation for the data set $(x_{1,\ldots\ldots,} x_n)$ if "typical" samples $(X_1(\theta),_{\ldots\ldots,} X_n(\theta))$ of sample size n generated using M resamples the data set " $(x_{1,\ldots\ldots,} x_n)$ (Davies 1995)[4]. To assess model, it is necessary to check if the model is in agreement with the data of size n by using goodness-of-fit statistic. Other standard techniques for the model assessment include the AIC etc, which involve the sample size issue. Model credibility index N^* proposed in Lindsay and Liu (2005) [5] is a new measure that applies goodness-of-fit testing thinking and is independent on the sample size. It uses a one-number summary statistic as an assessment tool in a false model world. In this study, we subsequently continue the work of Lindsay and Liu (2005) [5]. Accuracy of model credibility index will be evaluated and its confidence interval will be constructed.

1.2 Definition and Assumption

Lindsay and Liu (2005) [5] defined the model credibility index N^*, as a sample size (less than total sample size) under which the power of the test is 0.5. Lindsay and Liu(2005) [5] also set the assumption for the model credibility index, that is, total sample size n should be sufficiently large that many of the models under investigation are clearly false. Intuitively, N^* is relative to the model selection. N^* is inversely proportional to the squared distance measure that was used to construct the test statistics, such as Kullback-Leibler likelihood deviation (1959) [6]. The value of N^* is independent of sample size n but depends on which goodness-of-fit test is chosen. Traditional hypothesis test has played a prominent role in the assessment of models since the development of Pearson's chi-square statistic (Lindsay and Liu 2005) [5]. Under the assumption of requiring large enough sample size, models may be rejected due to the law of large numbers. The purpose to building N^* is to measure the quality of the approximation of models to true data generating mechanism under the model false assumption. Then one can find out one model which can capture main features of data most economically (Lindsay and Liu 2005) [5]. The larger the N^* value, the better the model fits the data.

1.3 The goals

After the model credibility index N^* is defined, the error of N^* estimation will be the main focus of this study. At first we apply the index N^* in the model selection process. Secondly, we can make better interpretation of the data generating mechanism and the characteristics of the model. Unlike standard methods for model selection, N^* is independent on sample size n, and has a straightforward interpretation based on the hypothesis testing methodology.

Finding the right variance of the model credibility index N^* is an important issue in our

study, because it can be used to judge the quality of the new statistic, as a follow up study of Lindsay and Liu (2005) [5]. However, directly obtaining the variance of N^* will be difficult because of the discontinuity of the integers N^*. To explore the variability of N^*, we consider the power obtained by N^*, and evaluate the variance of the power by some resampling techniques.

In chapter 2, we will review detailed information about the model credibility index N^*, including asymptotic approximation of N^*, determination of N^* using bootstrap. Beside the consistency of N^*, we will propose an improved bootstrap method to estimate the variance of the power obtained by N^*. In chapter 3, we will focus on applying the method on a two-way independent model by using Chi-Square goodness-of-fit test. Furthermore, we will carry a simulation study to verify the consistency of N^*. In chapter 4, a logistic model example will be used to illustrate the process of building model credibility index N^* and determining its error using likelihood ratio test. Finally, in chapter 5, based on the description within several chapters, we will make the conclusions about our study and discuss the possible further work.

CHAPTER 2: METHODOLOGY

2.1 Review

In previous study**,** Lindsay and Liu (2005) [5] introduced the definition of model credibility index N^* and used bootstrap method to determinate N^*. Before discussing consistency of N^*, we briefly introduce N^* in detail as in this section.

2.1.1 Basic ideas

As described in Lindsay and Liu (2005) [5], model credibility index is a statistic to measure goodness-of-fit models. It depends on the hypothesis testing methodology, but does not depend on the sample size used to estimate it. Data $X_1,\ldots,X_n$ are iid from distribution F; one is interested to test the goodness-of-fit of model M. Let $N^*(F,M)$ be the value of sample size that gives 0.5 power test between true distribution F and model M, we call the index N^* to be model credibility index for model M. One can generate this statistic with a known goodness-of-fit procedure, such as likelihood ratio test. Typically goodness-of-fit test statistics is based on distance measure. In general, the model credibility index $N^*(F,M)$ increases when the distance decreases. This is because a larger size is required to discriminate F and M when their distribution distance becomes smaller.

To construct model credibility index, we will use significant level $\alpha=0.05$ for the goodness-of-fit test, and 0.5 for the test power. These critical values seem to be arbitrary. The value of α only plays a minor role. The power 0.5 is motivated by the 50/50 model decision,

which greatly facilitates the asymptotic analysis. Those will be clearly shown in the following section.

2.1.2 Approximation of *N**

Lindsay and Liu (2005) [5] introduced methodology of how to construct asymptotic approximation of the model credibility index based on the likelihood ratio test in multinomial model. However, the approximation could be generalized. We show the generalized method in this section.

Lindsay and Liu (2005) [5] derived a simple asymptotic version of the testing index and showed that it is proportional to the reciprocal of the squared distance. This in turn leads to an elementary consistent estimator of the asymptotic index. This estimator has two important usages: first, it can also itself be bootstrapped; second, it provides a simple way to assess the variability of the estimated index.

The distance between a population distribution F and a model element M is defined as the likelihood deviation $2nL^2(F, M)$. This is a version of the Kullback-Leibler distance (Kullback, S., 1959) [6]. It technically operates as a squared distance, which is why we use the superscript 2.

Let $\hat{F}$ represents the empirical distribution of data $X_{1,\ldots,}X_{n,}$ likelihood statistic is then $2nL^2(\hat{F}, M)$. The likelihood ratio test statistic will have the asymptotic chi-squared distributions under the null hypothesis.

In the likelihood ratio test, one rejected the null hypothesis H_0: $F \in M$ at significant size α, if the likelihood ratio test statistic is large enough, that is

$$2n\,L^2 \geq x^2{}_{df}(\alpha),$$

where $x^2_{df}(\alpha)$ is the upper 1-α quantile of chi-squared distribution with degree freedom *df*. The power of the test at sample size *n* when $F \notin \mathrm{M}$ is

$$P_F\{2nL^2(\hat{F}, M) \geq x^2_{df}(\alpha) \mid F \notin M\}$$

The model credibility index *N** is defined as the sample size at which testing power for the alternative H_1: $F \notin M$ is 0.5, that is

$$P_F\{2N^*L^2(F, M) \geq x^2_{df}(\alpha) \mid F \notin M\} = 0.5,$$

Lindsay and Liu (2005)[5] used the fact that when the model is false, the centered likelihood ratio statistic has, asymptotically, a normal distribution with mean zero, and given approximation to *N** is

$$N^*_{asy} = \frac{x^2_{df}(\alpha)}{2L^2(F, M)}$$

The power takes 0.5 has enabled one to avoid calculating the asymptotic variance for the normal distribution. Clearly, the expression of the approximation *N** shows us, an inverse relationship to squared distance, the distance between the true sample distribution and the model plays a dominated role and α plays a role only in the numerator of the approximation to determinate *N**. The smaller the distance between the true sample distribution and the model is, the larger the *N** value will be. We also demonstrate this with examples in following chapters. In our examples, we take value 0.05 for α.

2.1.3 Application of bootstrap

Bootstrap techniques introduced by Efron (1993) [7], provided a simple and effective method to estimate the bias of the estimator, unknown distribution etc. To construct *N**, we use bootstrap procedure to simulate data sets, and expect to find the estimator of *N** under a targeted

power of rejecting the false model fitting and estimator of the variance of the power obtained at $\hat{N}^*$. Bootstrap approach provides a good tool to realize this goal. We employ different bootstrap resampling methods to estimate these two estimators, which will be discussed in the following paragraphs.

Let $X=(x_{1,\ldots,}x_n)$ iid from a distribution F, discrete or continuous. Because bootstrap procedure is to generate a set of samples for a data set $x_{1,\ldots\ldots,}x_n$, and each sample can be conducted the goodness-of-fit test for the models. To estimate N^*, we let the symbol $\hat{F}$ represents the empirical data distribution and use $\hat{F}$ to substitute F, then $\hat{N}^* = N^*(\hat{F}, M)$. We claim that null hypothesis is model M includes F, and alternative hypothesis is model M does not include F. The bootstrap generates samples under the alternative, so rejection probability is power of the test. We can then capture an estimator of N^* through this testing procedure by bootstrap simulation under various sample size. For any fixed sample size m, we simulate bootstrap samples of size m. Suppose we set simulation 1000 times for bootstrap procedure, we test goodness-of-fit for each bootstrap sample. If a significant result for each test is obtain, we say this sampling distribution is far from the model M. We count the total number of rejections in the 1000 bootstrap samples. When the proportion of rejection is about 500, we let this sample size be the estimator of N^*. If fail to get 500 times, we need to adjust the sample size until model be rejected by 50% of the time. We could carry out this analysis at any value of sample size. However, $\hat{N}^*$ is limited to be less than or equal to the total sample size n of the original data set. Otherwise, the process of seeking $\hat{N}^*$ would be noninformative. In order to save computation time by bootstrap, it is necessary to start an appropriate value based on the asymptotic approximation.

2.2 Consistency of bootstrap estimation of N*

After N^* is estimated, it is reasonable to evaluate the error of the estimation. However, it is difficult to discuss the continuity and consistency for integer N^*. Instead, we fix the sample size at m and consider the variance of the power. We will analyze the variance of the power to provide demonstration for consistency of bootstrap estimator $\hat{N}^*$. Lehmann's theory (1999) [8] guarantees the consistent of bootstrap estimator under some general regulations. Further, we will propose an improved subsampling estimator method.

Since consistency result of N^* is hard to be directly obtained, we need to explore the consistency of bootstrap estimator of power at certain sample size.

Assume that $X_1,\ldots,X_n$ are independent identically distributed random variables from the distribution F. We generate m bootstrap samples out of $X_1,\ldots,X_n$ for simplicity, denoted as $X_1,\ldots,X_m$. The test statistic is $T_m(X_1,\ldots,X_m)$. Our parameter is the power at sample size m, which is expressed as

$$\beta(m, F) = P_F\{T_m(X_1,\ldots, X_m) > t\}.$$

Here β is an estimator of power, and t is a critical value for the test statistic. In theory, the expression of the power estimator is an expectation function,

$$\beta(m, F) = E_F[\phi(X_1,\ldots, X_m)],$$

where $\phi(X_1,\ldots,X_m)$ is the indicator function $I\{T_m(X_1,\ldots, X_m) > t\}$. Without loss of generality we can assume ϕ to be symmetric in its m arguments. An unbiased estimator of β can be constructed as

$$U = \binom{n}{m}^{-1} \sum \ldots \sum_{1 \le i_1 < \ldots < i_m \le n} \phi(X_{i_1}, \ldots, X_{i_m}).$$

The following two theorems state the theory for asymptotic variance and the normal property of U in Lehmann (1999) [8].

Theorem 2.1. *If* $Var[\phi(x_1, \ldots, x_i, X_{i+1}, \ldots, X_m)] = \sigma_i^2$, *then*

(1) The variance of U-statistic equal to

$$Var(U) = \sum_{i=1}^{m} \binom{m}{i} \binom{n-m}{m-i} \sigma_i^2 \Big/ \binom{n}{m},$$

(2) If $\sigma_i^2 > 0$ *and* $\sigma_i^2 < \infty$ *for all i=1,...,m, then*

$$Var(\sqrt{n}U) \to m^2 \sigma_1^2$$

Theorem 2.2. (Lehmann Theorem)

(1) If $0 < \sigma_i^2 < \infty$, *then as* $n \to \infty$

$$\sqrt{n}(U - \beta) \xrightarrow{d} N(0, m^2 \sigma_1^2);$$

(2) If $\sigma_i^2 < \infty$ *for all i=1,...,m, then*

$$\frac{U - \beta}{\sqrt{Var(\sqrt{n}U)}} \xrightarrow{d} N(0,1).$$

Because ϕ is the indicator function, conditions $\sigma_i^2 < \infty$ for all i are obviously satisfied. The conditions for the consistency of bootstrap estimation of power at sample size m is, as long as $m^2/n \to 0$ and $n \to \infty$.

In the following section, we will show a new method by which the conditions for the consistency are simpler than Lehmann theorem (1999) [8].

2.3 Theory of bootstrap without replacement

The basic theory of bootstrap is presented in the context of independent data, that is, sampling with replacement.

A general theory could be derived based on smaller subsets of data. For example, for iid data set $x_{1,\ldots,}x_{n,}$ a statistic is computed over entire data and is recomputed over all $\binom{n}{m}$ data set of size *m.* Politis and Romano (1999)[9] called it as subsampling. The use of subsampling values to approximate the variance of a statistic is well known. The jackknife estimate of bias and variance has been well studied with $m=n-1$.

The random subsampling method uses subsets of data to approximate variance of statistics. Subsampling shares some similar properties to the bootstrap, in more broad generality. $X_{1,\ldots,}X_n$ is a sample with *n* iid random variables. The test statistic is $T_n(X_1,...,X_n)$.

Theorem 2.3: Assume the limiting distribution of $T_n(X_1,...,X_n)$ exists as $n\to\infty$. Also assume m$\to\infty$ and $m/n\to 0$ as $n\to\infty$.

Now let Y_i be subset of $\{X_{1.....}X_n\}$ without overlapping with size m and $T_{n,m,i}$ be the statistic T_m evaluated at the data set $Y_{i.}$ $Y_{1,\ldots,}Y_{Nn}$ are the $N_n=\binom{n}{m}$ subsets. Then

$$U_n = \frac{1}{N_n}\sum_{i=1}^{N_n} I(T_{n,m,i} > t)$$

converges to the limiting distribution of T_n in probability.

Remark: The proof of consistency of the subsampling distribution proves the consistency of the related *U* statistic. Furthermore, the variance of U_n approaches to zero when $n\to\infty$ as in the below Theorem 2.4.

Theorem 2.4: Under regulation of Theorem 2.3, let k be the greatest integer less than or equal to n/m. Let $T_{n,m,j}$ be the statistic evaluated at the data set $X_{m(j-1)+1},\ldots,X_{m(j-1)+m,}$ and set $\overline{U}_n = \frac{1}{k}\sum_{j=1}^{k} I(T_{n,m,j} > t)$. Since $\overline{U}_n$ is the average of k iid variables (each is bounded between 0 and 1).

$$Var(\overline{U}_n) \le \frac{1}{4k} \to 0\,,$$

$$Var(U_n) \le Var(\overline{U}_n) \to 0\,, \text{ as } n \to \infty\,.$$

The Theorems by Politis and Romano (1999) [9] improve the conditions of consistency. It only requires $m/n \to \infty$ and $n \to \infty$ for carefully constructed estimator.

2.4 Improved bootstrap estimation method

Based on Theorems by Politis and Romano (1999) [9], we propose an improved bootstrap method (sampling without replacement) to estimate the variance of power β at size m. We draw samples of size m from data $X_{1,\ldots,}X_n$ without replacement. But we control the rate of overlapping through techniques states as the below:

Let the number of overlapping variables be fixed at r and divide data $X_{1,\ldots,}X_n$ up into k random disjoint blocks. Let variables in the i^{th} block be denoted as $X_{r(i-1)+1,\ldots,}X_{r(i-1)+r.}$

To construct a subset of size m, we then draw the remaining *(m-r)* variables without replacement from the data $X_{1,\ldots,}X_{r(i-1),\ldots,}X_{ri+1,\ldots,}X_{n.}$ Then carry out the statistical test for the subset of m variables, record 1 if the results id a rejection and 0 if the results is not a rejection.

Repeat generating the *(m-r)* variables combined i^{th} block B times and the proportion of rejection is the simulated power $g_{i.}$

The mean of $g_{i,}$ i=1…k is the estimated power and variance of $g_{i,}$ i=1…k is the unbiased variance estimate.

One might randomly chosen the number of overlapping $r=p.m$, where p is a random number in (0,1). One might take various r and average the variance estimators.

The idea is similar to the jackknife estimator (which fixes $r=m-1$).

2.5 Confidence interval of approximations

The estimator of N^* is a point estimator, the confidence interval of N^* can be constructed. The confidence interval of N^* is an interval $[N^*_l, N^*_h]$ which is a random interval based on the sample. As known in our context, the power is a function of N^*. Therefore, we first estimate the power and generate confidence interval of the power β. We take approximately $Z_{\alpha/2}.\sigma$ as the margin error of the estimated power. Confidence interval of the power can be expressed as

$$\hat{\beta} \pm z_{\alpha/2}\sqrt{\hat{Var}} \ .$$

Confidence interval is $[\beta_l, \beta_h]$.

Taking inverse of the power function, confidence interval of N^* can be obtained. In the examples of the following chapters, we will use the described method to generate confidence intervals to approximate the power and N^*.

2.6 Another variance estimation idea

In this section, we provide a rough but convenient approach to estimate the variances of power. It also has a fast rate of convergence. In this method we don't control the overlapping rate.

Let *m* be the fixed subsample size (such as with power 0.5) and *n* be total sample size. To obtain an unbiased estimator of variance of power obtained at *m*, the improved method we will propose only needs $m/n \to 0$ and $n \to \infty$. Take *2m* samples without replacement repeatedly from the data size *n*. The limitation for this procedure is $2m < n$, and $m/n \to 0$ as $n \to \infty$.

Let the b[th] sample of size *2m* be broken into two subsample size *m*, say S_b and S_b^*. Suppose there are a total of B bootstrap samples taken. Let U_b and U_b^* be the corresponding test indicators from those samples. From this data one creates two estimators:

$$J = \frac{1}{T}\left[\sum (U_i \times U_j) + \sum (U_i^* \times U_j) + \sum (U_i \times U_j^*) + \sum (U_i^* \times U_j^*)\right]$$

Here the sums are over all pairs of *i* and *j* satisfying $j > i$ and *T* is the total number of product terms. These terms are estimating the covariance between the *U*'s from independently drawn subsets. Hence we never use $i = j$.

The second estimator is

$$K = \sum (U_i * U_i^*) / B.$$

This estimates the covariance between dependent *U*'s. The estimated variance of the *U* statistic is *J-K*. This form for the estimator shows that for the estimated variance to go to zero, the difference between *J* and *K* must go to zero. This will occur as *m* becomes much smaller than *n*, as then a pair of samples coming from different S_b (the *2m* samples) will only rarely overlap in any of their selections of *X*'s, and so start to look more like the two samples from the same S_b.

To obtain estimator of power, we summarized the procedures of drawing bootstrap resample without replacement next.

Step1: Select m as fixed subsample size (e.g. $\hat{N}^*$ by bootstrap)

Step2: Resampling at size *2m* from original data without replacement

Step3: Separate the *2m* subjects into two subsamples, each one with *m* numbers

Step4: Use statistic to do the test in each subsample

Step5: Record as 1 if reject, otherwise as 0, separately cumulate number of rejection in two subsamples.

Step6: Repeat setp2-step6 B times and count final number of rejection $u_1(i)$ and $u_2(i)$ for two Subsamples.

Step7: Obtain estimate of power= $\sum_{i=1}^{B}(u_1(i)+u_2(i))/2B$.

Using the information of U_1 *and* U_2, we obtained the estimator of the power. Subsequently, we can generate the variance of estimated power. The procedures is shown following.

Step1: Sum product value *SS1* with *u1(i)*u1(j)*, for all items of $j>i$, count items number

Step2: Sum product value *SS2* with u1*(i)*u2(j)*, for all items of $u2>u1$,count items number

Step3: Sum product value *SS3* with *u2(i)*u1(j)*, for all items of $u1>u2$,count items Number

Step4: Sum product value *SS4* with *u2(i)*u2(j)*, for all items of $j>i$,count items number

Step5: Take average of all product items from step1 to step 4 to Obtain *J*,

Step6: Sum value KK with *u1(i)*u2(i)*, for all terms of u1 and u2, take average of *KK*.

Step7: Obtain variance by *J-K*

CHAPTER 3: NUMERICAL EXAMPLE (TWO-WAY INDEPENDENT MODEL)

In this chapter, we apply model credibility index on contingency tables. Consider a contingency table in which the frequencies $n(t)$ in the cell t=1,….,T be a random sample. Random variables $n(t)$ have a multinomial distribution with parameters cell proportion and total size n. The cell proportions will be denoted $d(t)=n(t)/n$, which represent the empirical distribution of the data. In a two-way contingency table, data are collected with two variables and interest is to test the hypothesis that these two variables are independent. Chi-square test at 0.05 significant level are used for the test. In this section, we present a two-way table example to explain our method.

3.1 Data set

Table 3.1 shows a 5x4 contingency table with 5 columns, 4 rows, and total 20 cells about cross classifying number of children by annual income levels, which were used by Diaconis and Efron(1985) [10]. The data set has the categories of number of children being the rows and the categories of annual income being the columns. The sample size is n=25263.

Table 3.1 Cross-classification of number of children by annual income

Number of children	Annual income				
	0-1	1-2	2-3	3+	total
0	2161	3577	2184	1636	9558
1	2755	5081	2222	1052	11110
2	936	1753	640	306	3635
3	225	419	96	38	778
4+	39	98	31	14	182
Total	6116	10928	5173	3046	25263

Table 3.2 Expected value for each cell

Number of children	The expected annual income			
	0-1	1-2	2-3	3+
0	2313.9	4134.5	1957.1	1152.4
1	2689.7	4805.8	2274.9	1339.6
2	880.0	1572.4	744.3	438.3
3	188.3	336.5	159.3	93.8
4+	44.1	78.7	37.3	21.9

Expected values for each cell are shown in the Table 3.2. The results for goodness-of-fit test statistics, Chi-square test X^2=568.5663 and Log likelihood test $2nL^2$=569.4205 with 12 degree of freedom. Obviously, a strong significant p-value for the x^2-statistic leads to rejection of the independent. In this case, we say that the observed table is extremely close to dependent.

3.2 Bootstrap Determination of *N**

Table 3.3 shows the cell probability corresponding to the Table 3.1.

Table 3.3 Cell probability

	0-1	1-2	2-3	3+	total
0	0.085540118	0.141590468	0.086450540	0.0647587381	0.378339865
1	0.109052765	0.201124174	0.087954716	0.0416419269	0.439773582
2	0.037050232	0.069390017	0.025333492	0.0121125757	0.143886316
3	0.008906306	0.016585520	0.003800024	0.0015041761	0.030796026
4+	0.001543760	0.003879191	0.001227091	0.0005541701	0.007204212
total	0.2420932	0.4325694	0.2047659	0.1205716	1

To conveniently compute, we set bootstrap resampling 1000 times for the observed sample by keeping the proportion on each cell to generate a fixed sample size m with 0.5 power value. Each bootstrap resampling data is tested at a 0.05 significant level by Chi-square test and record the total rejection times. When the number of rejection is 500 times out of 1000 times, this sample size m is the estimation of model credibility index $\hat{N}^*$. According to the method of

generating N^* described in section 2.1.2, following flowchart shows the process of bootstrap determination of N^*.

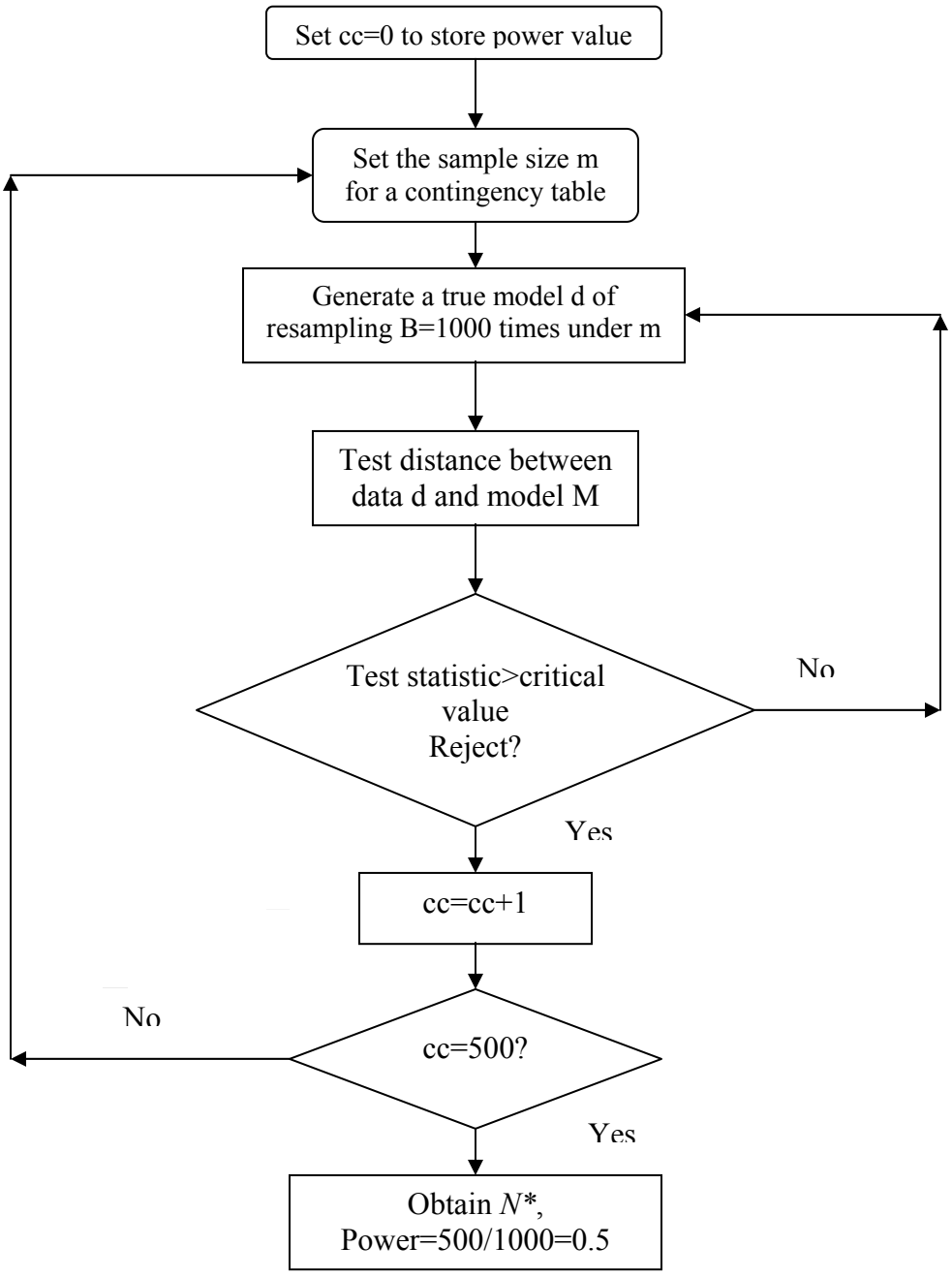

Figure 3.1 Flowchart to obtain N^*

The number of N* is used as a fixed sample size m to generate estimated power by bootstrap method. Using bootstrap resampling methods without replacement, we randomly resample without replacement for total data size into two groups both with size number *N**. If we do this many times, say 1000 times, we generate a statistic rejected the hypothesis with conventional levels of statistical significance under the null hypothesis of no difference between the observed distribution and empirical sampling distribution in each group. If the empirical sampling distribution includes about 500 samples out of 1000 observed sample (false model is our assumption, rejected probability is quite large), we conclude that the probability of such an outcome is about 0.5, which is the power we expected. We take the average from the two groups as estimated power. Randomization allows us to generate the sampling distribution without making any assumptions about the shape of the population distributions. The empirical sampling distribution (or reference distribution) emerges from the multiple randomizations of our observed data. We can determine the cumulated distribution for our observations on the sampling distribution to generate simulated sampling data. Therefore, large sample size is more practical to use randomization to generate the sampling distribution.

In the same way, following flowchart shows the process for estimator of power.

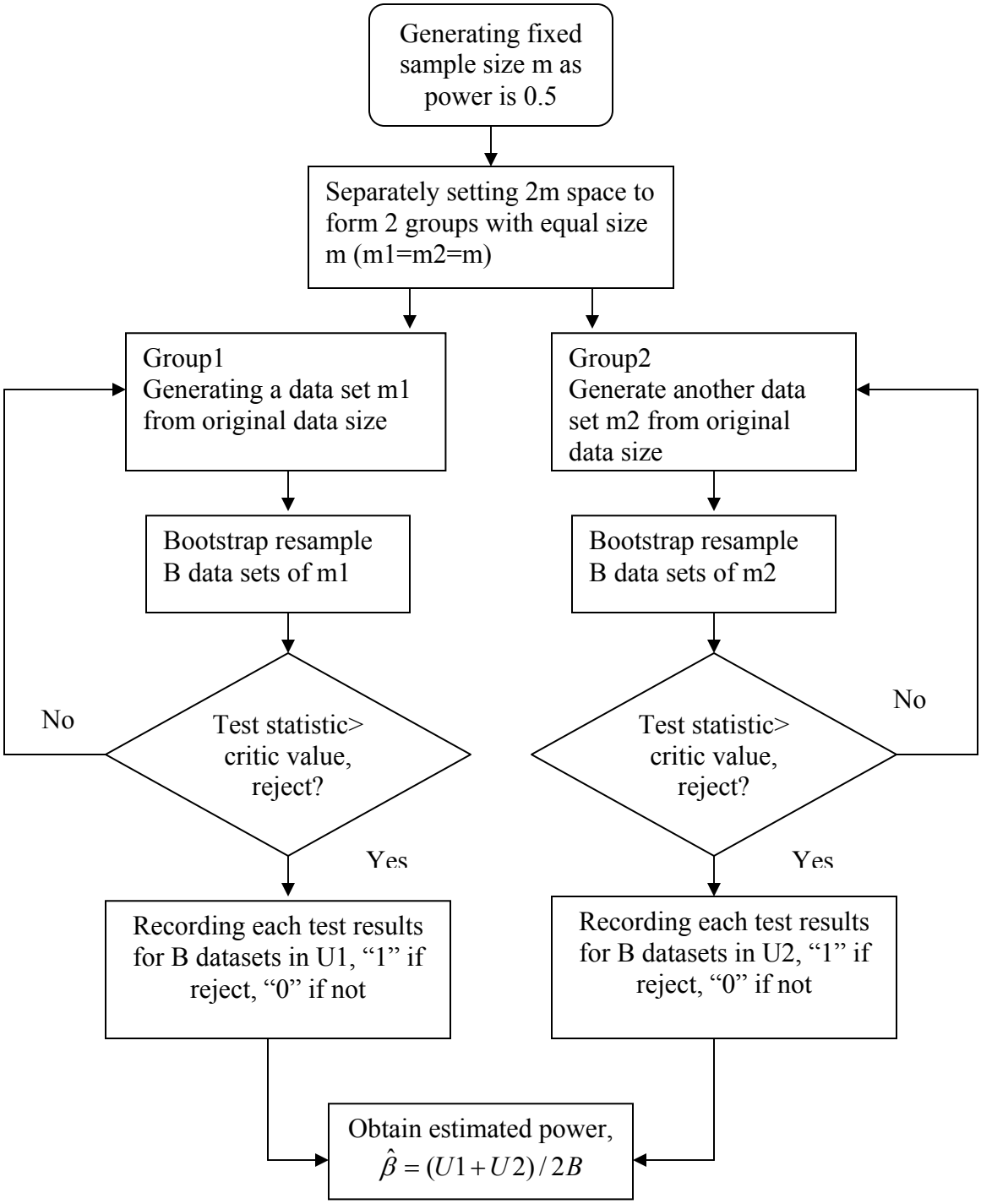

Figure 3.2 Flowchart to generate estimated power

Based on such calculation logic, we take different sample sizes m to observe the trend between fixed sample size m and generate estimator of power and to detect the model credibility

index, which is a sample size as power of rejection times is about 0.5. Table 3.4 shows their results.

Table 3.4 Sample size and corresponding power for Table 3.1

Sample size m	Power value	Sample size m	Power value
200	.356	550	.575
300	.375	560	.584
400	.423	570	.589
450	.446	580	.598
460	.457	600	.603
470	.468	700	.693
480	.485	800	.769
490	**.502**	900	.852
500	.513	1000	.976
510	.533	1200	.953
520	.536	1500	.991
530	.547	1800	.997
540	.549	2000	1

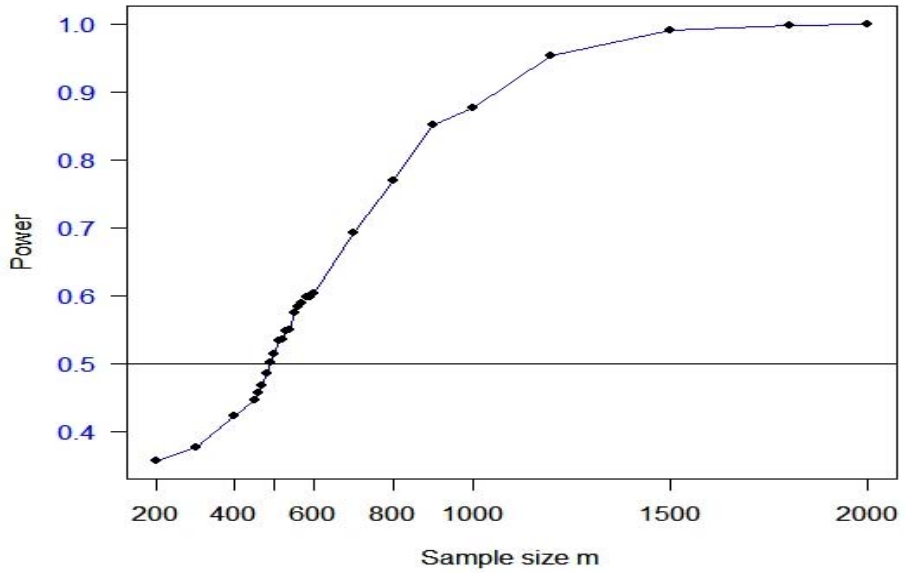

Figure 3.3 Power curve

Based on the values shown in the Table 3.4, we make Figure 3.3 to show relationship between sample size and power value.

Note two facts in Figure 3.3:

1. Model credibility index is the midpoint of the power curve.
2. Power is increasing in sample size m.

Under such facts, we may estimate model credibility index and justify model effectiveness. With a given data set, as sample size increases, the chance of model rejection rises, which indicates that the factors, number of children and income in the observed table are dependent.

3.3 Variance and confidence interval of power

To estimate the variance of the power obtained at any fixed sample size m, we use the subsampling method without replacement. We simulate $2m$ (m=490) data from original table without replacement and then randomly split these data into two groups of size m. The goodness-of-fit is tested separately within each group using Chi-square statistic. We mark "1" if rejection and mark "0" if not rejection, which constructs two "warehouses" where place the test results with "1" and "0". We count the rejection number and accumulate them respectively for each group. Finally, the estimator of the power is obtained by taking the average value for rejection number of two groups. Generally, the estimated power should be consistent with the fixed sample size that is obtained at 0.5 power value, but they are very different under some certain cases. We want to verify the variability of the power to demonstrate the behavior of the bootstrap resample without replacement method. Therefore we use information from these two "warehouses" by using the method described in section 2.6 to obtain first estimator J and second estimator K. Estimated variance $\hat{V}ar$ of the power is obtained by taking difference between J and K.

A flowchart of generating variance of estimated power also shows in following.

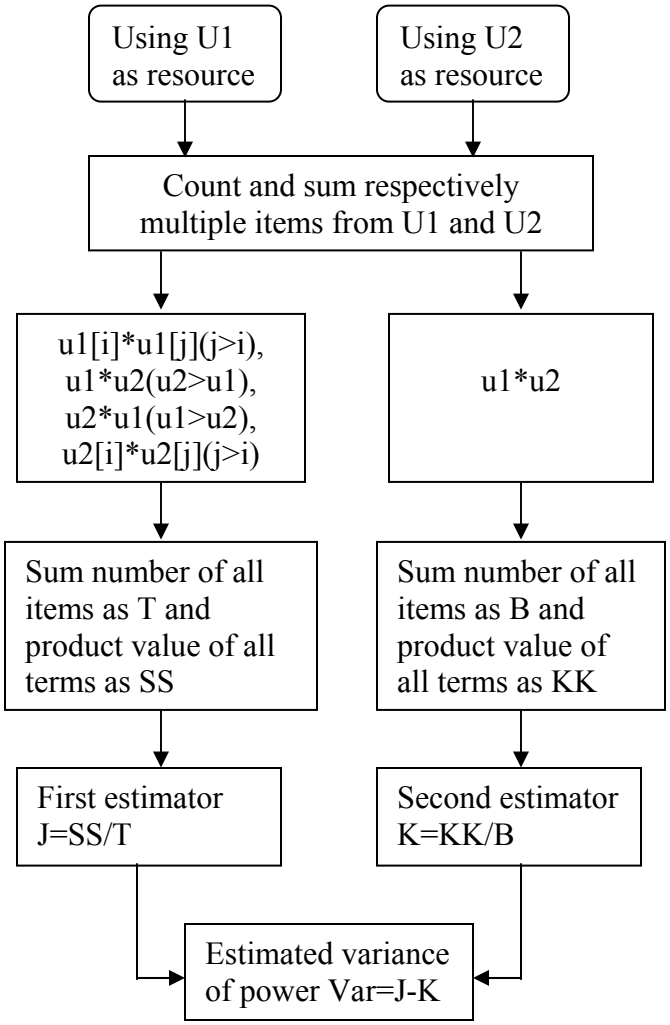

Figure 3.4 Flowchart to generate variance of estimated power

Based on the computing method, we obtain the results of estimated variance of power for data set in Table 3.1, which is shown in the Table 3.5.

Table 3.5 Summarization of the bootstrap resampling

J	0.251
K	0.247
$\hat{V}ar$ =J-K	0.004
$\hat{\sigma} = \sqrt{\hat{V}ar}$	0.063
Bootstrap estimator of Power, $\hat{\beta}$ =*(U1+U2)*/2	0.501

From the results of Table 3.5, the confidence interval of the bootstrap estimation of the power is shown below,

(1-α) % C.I. of the estimated power, $\hat{\beta} \pm z_{\alpha/2}\hat{\sigma}$

$$0.501 \pm 0.126 = [0.375, 0.627]$$

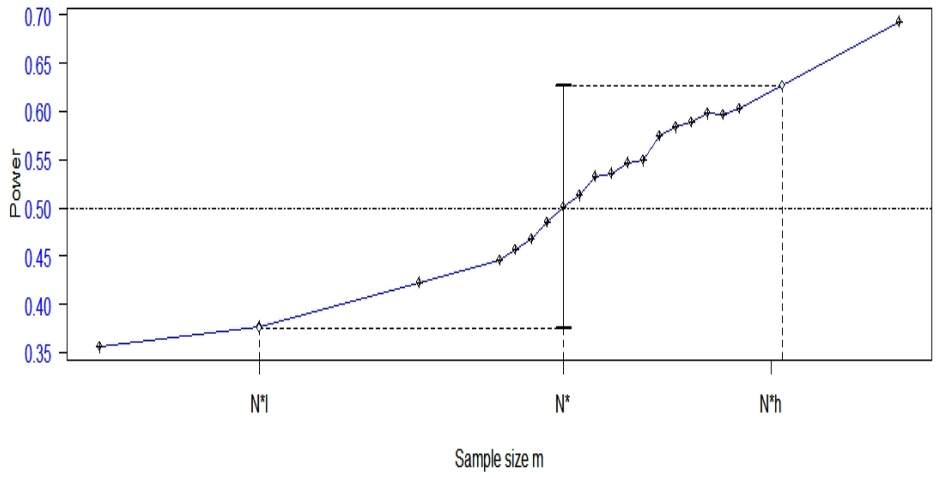

Figure 3. 5 Confidence interval of N*

Based on the above information, we make Figure 3.5 to find the relationship between power and model credibility index N*, as well as find their confidence intervals. From Figure 3.5, we can estimate C.I. of the estimator *N** is about [300, 627].

We resample from original huge sample size (n=25263), which leave the large space to generate samples similar with original sample. As a result, the estimated power with 0.501 value is very agreeable with the fixed sample size in this case. If the original sample size is small relatively, estimated power might differ with chosen 0.5 power value. To explore this phenomenon, we show process by simulation study in following section.

3.4 Simulation study

A simulation is an imitation of some real thing, state of affairs, or process. We use bootstrap resampling simulation technique generate observations from the distribution of the sample itself-the empirical distribution. In this section, simulation technique is used to generate the effect of the power and variance of power under varying sample size. Each simulation results in new two-m subjects from the original data without replacement. Without replacement means data are dependent, non-overlapping in a sample, but data are independent between samples so that we can use two estimators (non-overlapping covariance and square terms in variance) to obtain estimated variance of the power. We will interpret our method and show simulation results with analyzing model credibility indices for two-way table.

The consistency of power, hence N^*, is achieved as long as $m/n \to 0,\ n \to \infty$. As ratio m/n goes to zero, variance of power tends to convergence. To realize simulation study, we take a fixed number m=490 and different sample size n value, say 2000, 3000, 5000, 10000, 50000, and 100000 to make different ratio m/n. Using the computation methods in sections 2.5 and 2.6, we obtain the simulation results including estimated power, first estimator J, second estimator K and variance of power. Confidence interval of estimator power is computed at same time. All results are shown in Table 3.6 for each ratio m/n.

Table 3.6 Summarized results (*m*=490)

Sample size *n*	*m/n*	$\hat{\beta}$ =(U1+ U2)/2	*J*	*K*	$\hat{V}ar$ = J-K	$\hat{\sigma} = \sqrt{\hat{V}ar}$	C.I. of $\hat{\beta}$ $\hat{\beta} \pm 2\hat{\sigma}$
100000	0.0049	0.502	0.252	0.250	0.002	0.045	[0.412,0.592]
50000	0.0098	0.536	0.288	0.283	0.005	0.071	[0.394,0.678]
10000	0.049	0.646	0.418	0.411	0.007	0.084	[0.478,0.814]
5000	0.098	0.501	0.251	0.235	0.016	0.126	[0.249,0.753]
3000	0.163	0.625	0.391	0.364	0.027	0.164	[0.297,0.953]
2000	0.245	0.559	0.312	0.269	0.043	0.205	[0.145,0.973]

Based on the results in Table 3.6, we make graphs between sample size *n* and power variance, between ratio *m/n* and power variance, between sample size *n* and power confidence interval, between ratio *m/n* and power confidence interval.

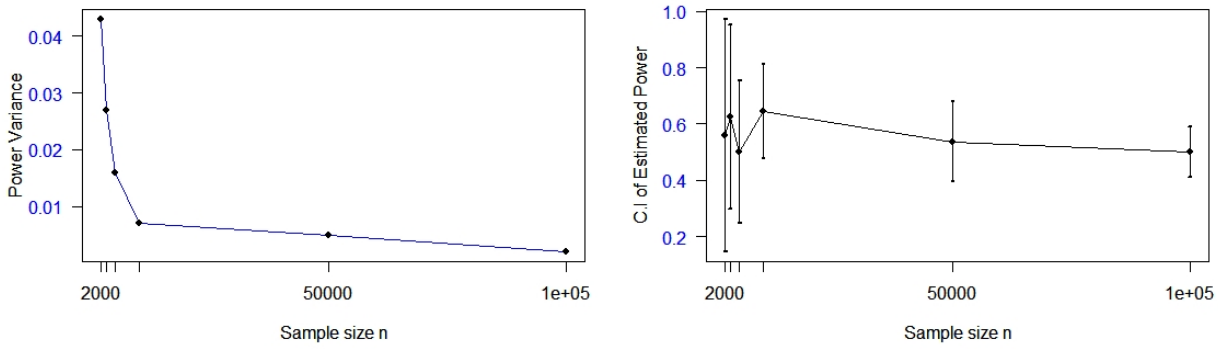

Figure 3.6 Sample size *n* vs power variance and vs C.I in simulation study

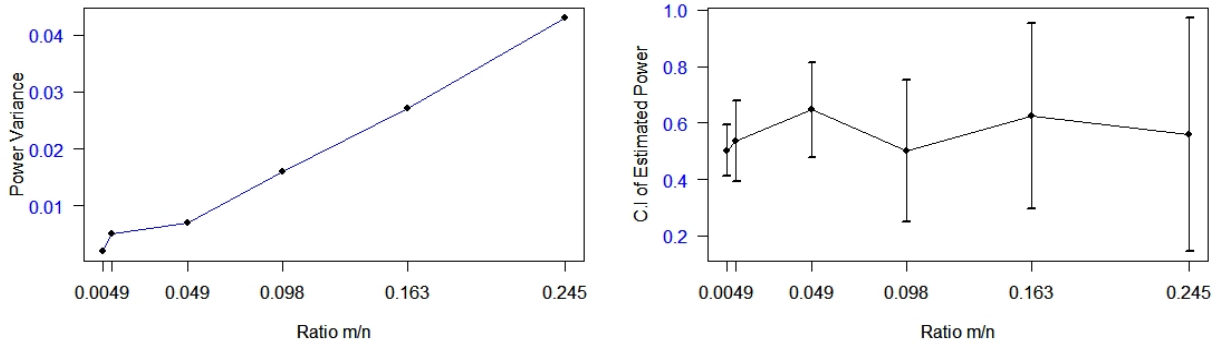

Figure 3.7 Ratio *m/n* vs power variance vs C.I. in simulation study

The relationships of *m/n* vs variance of the estimated power and *m/n* vs confidence interval of power are displayed in Figure 3.7.

As described in 2.4, as $m/n \rightarrow 0$, we can obtain an unbiased estimator of variance of power obtained at size m by using our improved bootstrap estimated method. From Figure 3.7, we can observe a phenomenon that power variance is decreasing, as the ratio *m/n* gets smaller, which is consistent with our method. For two-way table case, we estimate the model credibility index N^*, and evaluate the variance of its power, as well as construct the confident interval of the power.

We provide a real-life example in next chapter for logistic regression model.

CHAPTER 4: NUMERICAL EXAMPLE (LOGISTIC MODEL)

4.1 Data set

We present a four-way contingency table shown in Table4.1, which summarizes data of automobile accident that were used by Alan Agresti (2002) [11]. In this data set, 68,694 passengers in autos and light trucks involved in accidents in the state of Maine in 1991. The passengers are classified into four variables by gender, location of accident, seat-belt use, and injury situation. To simplify, we express factors gender by *G*, location by *L*, seat-belt use by *S*, and injury by *I*.

Table 4.1 Data of automobile accident

Gender	Location	Seat Belt	No Injury	Injury
Female	Urban	No	7287	996
Female	Urban	Yes	11587	759
Female	Rural	No	3246	973
Female	Rural	Yes	6134	757
Male	Urban	No	10381	812
Male	Urban	Yes	10969	380
Male	Rural	No	6123	1084
Male	Rural	Yes	6693	513

4.2 Models selection

For the dataset in the Table 4.1, we fit five kind logistic models (total 46 models) from 7 parameters to 11 parameters (k) with three 2-way interactions, four 2-way interactions, five 2-way interactions, six 2-way interactions, and one 3-way interaction. The fitted results, likelihood ratio test (L.R.T) and model quality statistic AIC are listed for each model.

For logistic model, likelihood ratio test statistic is used to test goodness-of-fit instead Chi-square statistic for two-way independent model. We follow the steps that are similarly for two-way independent model to generate $\hat{N}^*$ for models highlighted in Table 4.2. These models have the smallest LRT statistics in models with same number of parameters and hence are considered the best at its complexity.

Table 4.2 Process of model selection

Model	With one 3-way interaction	*k*	L.R.T	*N**
1	**S+L+G+I+GL+GS+GI+LS+LI+SI+S*L*G**	**11**	**7.46**	**60300**
2	S+L+G+I+GL+GS+GI+LS+LI+SI+S*L*I	11	20.63	
3	S+L+G+I+GL+GS+GI+LS+LI+SI+S*G*I	11	22.85	
4	S+L+G+I+GL+GS+GI+LS+LI+SI+L*G*I	11	18.57	

Model	With all (six) 2-way interaction	*k*	L.R.T	*N**
1	**S+L+G+I+GL+GS+GI+LS+LI+SI**	**10**	**23.35**	**20500**

Model	With five 2-way interaction	*k*	L.R.T	*N**
1	S+L+G+I+GL+GS+GI+LS+LI	9	921.7	
2	S+L+G+I+GL+GS+GI+LS+SI	9	806.5	
3	**S+L+G+I+GL+GS+GI+LI+SI**	**9**	**50.89**	**10560**
4	S+L+G+I+GL+GS+LS+LI+SI	9	421.7	
5	S+L+G+I+GL+GI+LS+LI+SI	9	891.7	
6	S+L+G+I+GS+GI+LS+LI+SI	9	193.5	

Model	With four 2-way interaction	*k*	L.R.T	*N**
1	S+L+G+I+GL+GS+GI+LS	8	1681	
2	S+L+G+I+GL+GS+GI+LI	8	925.8	
3	S+L+G+I+GL+GS+GI+SI	8	810.6	
4	S+L+G+I+GL+GS+LS+LI	8	1204	
5	S+L+G+I+GL+GS+LS+SI	8	1155	
6	S+L+G+I+GL+GS+LI+SI	8	446	
7	S+L+G+I+GL+GI+LS+LI	8	1674	
8	S+L+G+I+GL+GI+LS+SI	8	1665	
9	S+L+G+I+GL+GI +LI+SI	8	906	
10	S+L+G+I+GL+LS+LI+SI	8	1174	
11	S+L+G+I+GS+GI+LS+LI	8	1082	
12	S+L+G+I+GS+GI+LS+SI	8	926	
13	**S+L+G+I+GS+GI +LI+SI**	**8**	**208**	**2650**
14	S+L+G+I+GS+LS+LI+SI	8	542.1	
15	S+L+G+I+GI+LS+LI+SI	8	1049	

Model	With three 2-way interaction	*k*	L.R.T	*N**
1	S+L+G+I+GL+GS+GI	7	1686	
2	S+L+G+I+GL+GS+LS	7	1923	
3	S+L+G+I+GL+GS+LI	7	1208	
4	S+L+G+I+GL+GS+SI	7	1159	
5	S+L+G+I+GL+GI+LS	7	2433	
6	S+L+G+I+GL+GI+LI	7	1674	
7	S+L+G+I+GL+GI+SI	7	1666	
8	S+L+G+I+GL+LS+LI	7	1955	
9	S+L+G+I+GL+LS+SI	7	1907	
10	S+L+G+I+GL+LI+SI	7	1188	
11	S+L+G+I+GS+GI+LS	7	1802	
12	S+L+G+I+GS+GI+LI	7	1083	
13	S+L+G+I+GS+GI+SI	7	927.6	
14	S+L+G+I+GS+LS+LI	7	1324	
15	S+L+G+I+GS+LS+SI	7	1275	
16	**S+L+G+I+GS+LI+SI**	**7**	**556.6**	**1000**
17	S+L+G+I+GI+LS+LI	7	1831	
18	S+L+G+I+GI+LS+SI	7	1782	
19	S+L+G+I+GI+LI+SI	7	1063	
20	S+L+G+I+LS+LI+SI	7	1291	

Our interest is to obtain useful information from false model. Models listed in Table 4.3 are all rejected by conventional LRT testing, while the *N** index shows that they still have desirable goodness-of-fit.

Table 4.3 Summarized models

		k	L.R.T	AIC	*N**	*N**/*k*
Model 1	S+L+G+I +GS+LI+SI	7	556.6	722.1	1000	142.9
Model 2	S+L+G+I +GS+GI +LI+SI	8	208	379.5	2650	331.2
Model 3	S+L+G+I+GL+GS+GI+LI+SI	9	50.89	224.2	10560	1173.3
Model 4	S+L+G+I+GL+GS+GI+LS+LI+SI	10	23.35	198.8	20500	2050

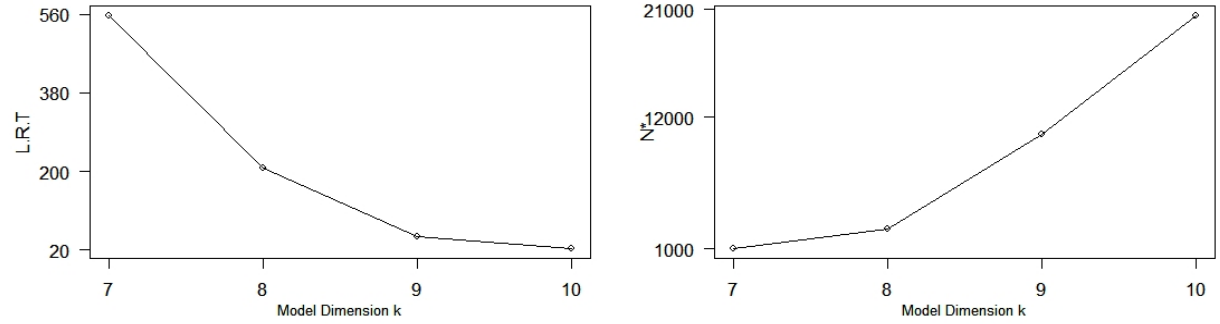

Figure4.1 Model dimension *k* vs likelihood ratio test (L.R.T) and *k* vs *N

4.3 Results

The power, the variance of the power, and confidence interval are calculated for each model below by bootstrap resampling without replacement method. We make inference to obtain confidence interval of the power.

Table 4.4 Results of bootstrap without replacement method

	k	N*	Power	Variance	$\hat{\sigma}$	C.I
Model 1	7	1000	0.492	0.0001	0.010	[0.472,0.512]
Model 2	8	2650	0.501	0.0140	0.118	[0.265,0.737]
Model 3	9	10560	0.478	0.0289	0.170	[0.138,0.818]
Model 4	10	20500	0.383	0.2365	0.486	[0,1]

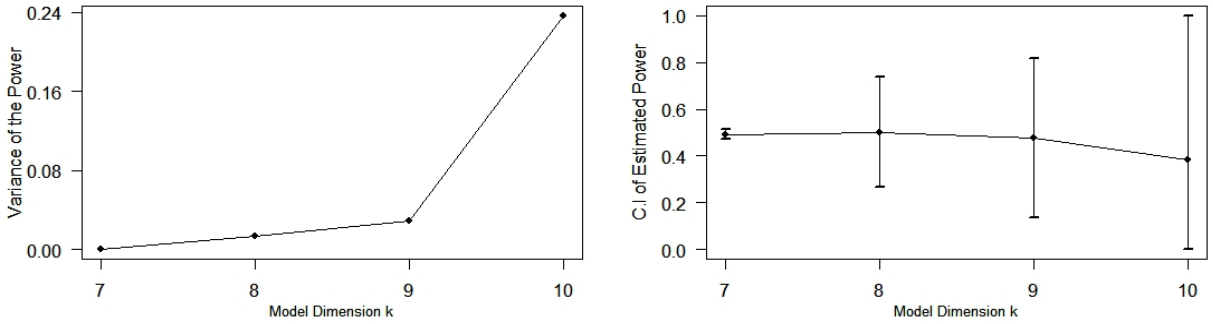

Figure4.2 Model dimension *k* vs variance of estimated power and vs C.I of power

4.4 Conclusion

From the results, we can conclude,

(1) Models with more parameters generate larger N^* value than model with few parameters. On the other hand, models with larger N^* fit better. This conclusion is consistence with likelihood ratio test value, that is, model with smaller LRT is better-fitted one. As seen in Figure 4.1, Models with more parameters accompany smaller likelihood ratio test value during selecting model**,** which indicates a good fitted model.

(2) Models with more parameters generate larger variance of power value than with few parameters' models.

(3) Models with more parameters are harder to obtain consistency estimate of the power, which generate wider confidence interval than with few parameters by bootstrap without replacement method.

For this example, model 1 in Table 4.4 should be one of the candidate models with smallest variance, smallest model credibility index N^*, smallest confidence interval. And model 2 is another candidate model with moderated values of N^*, power, variance, and wider confidence interval.

CHAPTER 5: DISCUSSION AND FUTURE WORK

Liu and Lindsay [3] introduced the model credit index method for model selection in a model false world with large data size. In this study, we extend the model credit index method in measuring consistency of estimation of model credit index by variance estimator of power using bootstrap without replacement.

By applying the methods to independent model and logistic model, the model characteristic can be identified using N^*. For a given total sample size n, fixed sample size m increase, $m/n \to$large, power $\to$large and its variance $\to$large for independent model. By simulation study, consistency of power is verified, that is, as $n \to \infty$, then $m/n \to 0$, and variance$\to 0$. For logistic model, model good fit with larger model parameter k, and generate large N^*, large variance of power and large confidence interval.

By the proven facts through out examples, the method of consistency of power proposed is demonstrated. Also, the method generates an unbiased variance estimator and converges fast with ratio $m/n \to 0$. The method in section 2.4 provides better estimation of variance. We would utilize this method in the future. Besides, trying to practice method in the general model like continuous data is also an issue for the further work.

REFERENCES

[1] Giudici, P. (2003) *Applied Data Mining*, Wiley.

[2] Hand, D.J. (2000) Methodological issues in data mining, in J.G.Bethlehem and P.G.M. van der Heijden (editors)*, Compstat 2000: Proceedings in Computational Statistics*, 77-85, Physica-Verlag

[3] Linhart.H. and Zucchini.W. (1986). *Model Selection*. John Wiley & Sons.

[4] Davies, P. L. (1995). *Data Features*. Statistica Neerlandica, 49 p185-245.

[5] Lindsay, Bruce. and Liu, Jiawei (2005). Model Assessment Tools for a Model False World. Statistical Science. To be published.

[6] Kullback, S. (1959), *Information Theory and Statistics*, New York: Wiley.

[7] Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall Ltd.

[8] Lehmann, E.L. (1999), *Elements of large-sample Theory*. Springer.

[9] Politis, D.N., Romano, J.P., Wolf, M. (1999).*Subsampling.* Springer, New York.

[10] Diaconis, P. and Efron, B. (1985). Reply to comments on "testing for independence in a two-way table: New interpretations of the chi-square statistic". *The Annals of Statistics*, 13 905-913.

[11] Alan Agresti (2002). *Categorical Data Analysis*. A John Wiley & Sons, Inc.

APPENDICES

Appendix A: R code to generate fixed sample size N* and estimated power for 2-way table

```
#produce data of table4

table4<-
c(2161,3577,2184,1636,2755,5081,2222,1052,936,1753,640,306,225,419,96,38,39,9
8,31,14)
n0<-sum(table4)

oc<-
matrix(table4,nrow=5,ncol=4,byrow=TRUE,dimname=list(c("0","1","2","3","4+"),c
("0-1","1-2","2-3","3+")))
oc #observed cell

rtot<-apply(oc,1,sum)
rtot #row total

ctot<-apply(oc,2,sum)
ctot #column total

ei<-outer(rtot,ctot,"*")/sum(oc)
ei #estimated cell

chi2.T<-sum((oc-ei)^2/ei)
chi2.T #5x4 table chi-square
pchisq(q=chi2.T,df=12, lower.tail=FALSE) #p-value

L<-2 * sum(oc * log(oc/ei))
L #likelihood ratio test


dt<-prop.table(oc)
dt  #obtain proportion table for each cell

drtot<-apply(dt,1,sum)
drtot #row total

dctot<-apply(dt,2,sum)
dctot #column total

dei<-outer(drtot,dctot,"*")/sum(dt)
dei #estimated cell

dL<-2 * sum(dt * log(dt/dei))
dL

chi.cri<-qchisq(0.05,12) #chi-square value
N.asy<-round(chi.cri/dL)
```

```
N.asy
chisq.test(oc)


#resampling table4(keep proportion for each cell)

#loop bt times, and record number of rejection for 1000 loops
#n is number of total observations in the table

cc<-0
n<-490 # n can be changed

bt<-1000 #bt can be changed

rsi<-rmultinom(bt, n, dt) #produce bt columns, each one can consist of a
table

for (i in 1:bt)
{
rs<-matrix(rsi[,i],nrow=5,ncol=4)

ck<-chisq.test(rs, rescale.p=TRUE)$p.value

if(is.na(ck))
{cc<-cc+1}

if(!is.na(ck))

if (ck<0.05)
{cc<-cc+1}
}

cc #rejected number in 1000 random two way tables(keep same proportion)
bata<-cc/bt
bata #estimated power when total sample size n=490

########
####generate 2 vectors as 2 groups for reject=1,not reject=0
#u1 vector for Group1
#u2 vector for Group2

m<-490 #reject number as power=0.5

n0<-25263 #initial observation's sum of table4

sam1<-matrix(0,20,bt)#assign 20 cell(5x4)space
sam2<-matrix(0,20,bt)
u1<-c(rep(0,1000))#assign 1000 space
u2<-c(rep(0,1000))


#generate 1000 tables(between these tables are independent; in each single
table,data follow the trend of initial one which is dependent)by 1000
bootstraps to test reject number

for (i in 1:bt)
{
```

```
twoid<-rep(0,2*m)#assign space for twoid
twoid<-sample(1:n0,size=2*m,replace=F)
id1<-twoid[1:m]#498 number, the value are less than 25263
id2<-twoid[-(1:m)]#another 498 number, values are less than 25263

br<-append(0,cumsum(table4))#cumulate initial data(20 cells)

sam1[,i]<-tabulate(cut(id1,br,include.lowest=T),length(br)-1)#generate 20
number which follow the trend of initial one

sam2[,i]<-tabulate(cut(id2,br,include.lowest=T),length(br)-1)#generate
another 20 number

rs1<-matrix(sam1[,i],nrow=5,ncol=4, byrow=T)#each 20 number consist to one
5x4 table

ck1<-chisq.test(rs1, rescale.p=TRUE)$p.value

if(is.na(ck1))
{u1[i]<-u1[i]
u1[i]<-u1[i]+1}

if(!is.na(ck1))

if (ck1<0.05)
{u1[i]<-u1[i]+1}

rs2<-matrix(sam2[,i],nrow=5,ncol=4,byrow=T)

#prop.table(rs2)

ck2<-chisq.test(rs2, rescale.p=TRUE)$p.value

if(is.na(ck2))
{u2[i]<-u2[i]
u2[i]<-u2[i]+1}

if(!is.na(ck2))

if (ck2<0.05)
{u2[i]<-u2[i]+1}

}
u1
table(u1) # number of reject in group1
u2
table(u2)#number of reject in group2


batahat<-sum(u1,u2)/(2*bt)

batahat #average value as estimated power
```

Appendix B: R coding for simulation of 2-way table

```
table4<-
c(2161,3577,2184,1636,2755,5081,2222,1052,936,1753,640,306,225,419,96,38,39,9
8,31,14)
n0<-sum(table4)

oc<-
matrix(table4,nrow=5,ncol=4,byrow=TRUE,dimname=list(c("0","1","2","3","4+"),c
("0-1","1-2","2-3","3+")))

dt<-prop.table(oc)
dt  #obtain proportion table for each cell

p<-prop.table(table4)



n<-100000
sim.size<-rmultinom(1, n, dt)
no<-sum(sim.size)
sim.oc<-
matrix(sim.size,nrow=5,ncol=4,dimname=list(c("0","1","2","3","4+"),c("0-
1","1-2","2-3","3+")))

sim.dt<-prop.table(sim.oc)


#n<-25263#original data size
# simulation various n
#n<-1000
#n<-5000
#n<-10000
#n<-50000
#n<-100000

cc<-0
N<-490
bt<-1000 #bt can be changed

rsi<-rmultinom(bt, N, sim.dt) #produce bt columns, each one can consist of a
table

for (i in 1:bt)
{
rs<-matrix(rsi[,i],nrow=5,ncol=4)

ck<-chisq.test(rs, rescale.p=TRUE)$p.value

if(is.na(ck))
{cc<-cc+1}

if(!is.na(ck))

if (ck<0.05)
{cc<-cc+1}
```

```
}

cc #rejected number is 505 in 1000 random two way tables(keep same
proportion)
bata<-cc/bt
bata

########
####generate 2 vectors as 2 groups for reject=1,not reject=0
#u1 vector for Group1
#u2 vector for Group2

m<-490 #reject number as power=0.5

n0<100000 #initial observation's sum of table4

sam1<-matrix(0,20,bt)#assign 20 cell(5x4)space
sam2<-matrix(0,20,bt)
u1<-c(rep(0,1000))#assign 1000 space
u2<-c(rep(0,1000))


#generate 1000 tables(between these tables are independent; in each single
table,data follow the trend of initial one which is dependent)by 1000
bootstraps to test reject number

for (i in 1:bt)
{

twoid<-rep(0,2*m)#assign space for twoid
twoid<-sample(1:n0,size=2*m,replace=F)
id1<-twoid[1:m]#490 number, the value are less than 25263
id2<-twoid[-(1:m)]#another 490 number, values are less than 25263

br<-append(0,cumsum(table4))#cumulate initial data(20 cells)

sam1[,i]<-tabulate(cut(id1,br,include.lowest=T),length(br)-1)#generate 20
number which follow the trend of initial one

sam2[,i]<-tabulate(cut(id2,br,include.lowest=T),length(br)-1)#generate
another 20 number

rs1<-matrix(sam1[,i],nrow=5,ncol=4, byrow=T)#each 20 number consist to one
5x4 table

ck1<-chisq.test(rs1, rescale.p=TRUE)$p.value

if(is.na(ck1))
{u1[i]<-u1[i]
u1[i]<-u1[i]+1}

if(!is.na(ck1))

if (ck1<0.05)
{u1[i]<-u1[i]+1}

rs2<-matrix(sam2[,i],nrow=5,ncol=4,byrow=T)
```

```
#prop.table(rs2)

ck2<-chisq.test(rs2, rescale.p=TRUE)$p.value

if(is.na(ck2))
{u2[i]<-u2[i]
u2[i]<-u2[i]+1}

if(!is.na(ck2))

if (ck2<0.05)
{u2[i]<-u2[i]+1}

}
u1
table(u1) # number of reject in group1
u2
table(u2)#number of reject in group2


batahat<-sum(u1,u2)/(2*bt)

batahat #average value as estimated power
```

Appendix C: R code for logistic model selection

```
################################
Generating N* and estimated power for 4-way contingency table
################################

d0<-c(7287,996,11587,759,3246,973,6134,757,10381,812,10969,380,6123,
      1084,6693,513)#dataset
n0<-sum(d0)
dt<-d0/n0 #proportion

#### models with different number of 2-way interactions

cc<-0
n<-1000#for model.16
n<-2650 #for model.13
n<-10560#for model3
n<-20500 #for model4
n<-60300#for model5

bt<-1000 #bt can be changed

rsi<-rmultinom(bt, n, dt) #produce bt columns, each one can consist of a
table

for (i in 1:bt)
{
d<-matrix(rsi[,i])
data.s<-cbind(expand.grid(inj=c("ni","yi"),
            sb=c("no","yes"), loc=c("urban","rural"),
            gen=c("female","male")), Fr=d)

#with 1 3-way interaction  model.0 BEST
model.0<-
glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+loc*inj+sb*inj+sb*loc*gen
,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


model.1<-
glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+loc*inj+sb*inj+sb*loc*inj
,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.2<-
glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+loc*inj+sb*inj+sb*gen*inj
,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))
model.3<-
glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+loc*inj+sb*inj+loc*gen*in
j,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))
```

```
#with 6 2-way interactions

model.0<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model with 5 2-factors,model.3 best
#model 1
model.1<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+loc*inj,

        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 2
model.2<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+sb*inj,
family=poisson, data=data.s,
         control=glm.control(epsilon=0.0000001,maxit=100))

#model 3 best one
model.3<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model 4 remove gen*inj+gen*loc
model.4<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+loc*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model 5
model.5<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*inj+loc*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 6
model.6<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+loc*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))




#15 models with 4 2-way interactions (15 models),model.13 best
#model 1
model.1<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb,

                        family=poisson, data=data.s,
```

```
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 2
model.2<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 3
model.3<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model 4 remove gen*inj+gen*loc
model.4<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+loc*sb+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 5
model.5<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+loc*sb+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 6
model.6<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 7
model.7<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*inj+loc*sb+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 8
model.8<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*inj+loc*sb+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 9
model.9<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*inj+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 10
model.10<-glm(Fr~gen+loc+sb+inj+gen*loc+loc*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))
```

```
#model 11
model.11<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+loc*sb+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 12
model.12<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+loc*sb+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model 13 AIC smallest best one
model.13<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 14
model.14<-glm(Fr~gen+loc+sb+inj+gen*sb+loc*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model 15
model.15<-glm(Fr~gen+loc+sb+inj+gen*inj+loc*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model with 3 2-way interactions (20 models),model1.16 best

model.1<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.2<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+loc*sb,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.3<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.4<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+sb*inj,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.5<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*inj+loc*sb,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))
```

```
model.6<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*inj+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.7<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.8<-glm(Fr~gen+loc+sb+inj+gen*loc+loc*sb+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.9<-glm(Fr~gen+loc+sb+inj+gen*loc+loc*sb+sb*inj,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.10<-glm(Fr~gen+loc+sb+inj+gen*loc+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.11<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+loc*sb,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.12<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.13<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.14<-glm(Fr~gen+loc+sb+inj+gen*sb+loc*sb+loc*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.15<-glm(Fr~gen+loc+sb+inj+gen*sb+loc*sb+sb*inj,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#best one
model.16<-glm(Fr~gen+loc+sb+inj+gen*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.17<-glm(Fr~gen+loc+sb+inj+gen*inj+loc*sb+loc*inj,

                        family=poisson, data=data.s,
```

```
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.18<-glm(Fr~gen+loc+sb+inj+gen*inj+loc*sb+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.19<-glm(Fr~gen+loc+sb+inj+gen*inj+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

model.20<-glm(Fr~gen+loc+sb+inj+loc*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


ndf<-model.16$df.residual
dev<-model.16$deviance

ndf<-model.13$df.residual
dev<-model.13$deviance

ndf<-model.3$df.residual
dev<-model.3$deviance

ndf<-model.0$df.residual
dev<-model.0$deviance

ndf<-model.0$df.residual
dev<-model.0$deviance

ck<-dev
cr<-qchisq(1-0.05,ndf)

      if(ck>cr)
            {cc<-cc+1}
}

bata<-cc/bt #obtain power for rejection
bata


###################################
#results of n as power=0.5
model.0 n=60300 bata=0.5 #1 3-way interaction
model.0 n=20500  bata=0.5 #all 6 2-way interactions
model.3  n=10560 bata=0.5 #5 2-way interactions
model.13 n=2650  bata=0.5 #4 2-way interactions
model.16 n=1000   bata=0.5 #3 2-way interactions

########################################


####generate 2 vectors as 2 groups for reject=1,not reject=0
```

```
#u1 vector for Group1
#u2 vector for Group2

#for model with 3 2-way interaction

m<-1000
m<-2650
m<-10560
m<-20500
m<-60300
n0<-68694 #initial table sum

sam1<-matrix(0,16,bt)#assign space
sam2<-matrix(0,16,bt)
u1<-c(rep(0,1000))#assign 1000 space
u2<-c(rep(0,1000))

#generate 1000 tables(between these tables are independent; in each single
table,data follow the trend of initial one which is dependent)by 1000
bootstraps to test reject number

for (i in 1:bt)
{

twoid<-rep(0,2*m)#assign space for twoid
twoid<-sample(1:n0,size=2*m,replace=F)
id1<-twoid[1:m]#20500 number, the value are less than 68694
id2<-twoid[-(1:m)]#another 20500 number, values are less than 68694

br<-append(0,cumsum(d0))#cumulate initial data(16 cells)

sam1[,i]<-tabulate(cut(id1,br,include.lowest=T),length(br)-1)#generate 16
number which follow the trend of initial one

sam2[,i]<-tabulate(cut(id2,br,include.lowest=T),length(br)-1)#generate
another 16 number

d<-sam1[,i]
data.s<-cbind(expand.grid(inj=c("ni","yi"),
            sb=c("no","yes"), loc=c("urban","rural"),
            gen=c("female","male")), Fr=d)

#using all best model for each kind dimention model

#model for 3 2-way interactions
model.16<-glm(Fr~gen+loc+sb+inj+gen*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model for 4 2-way interactions
model.13<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model for 5 2-way interaction
```

```
model.3<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model for 6 2-way interactions

model.0<-glm(Fr~gen+loc+sb+inj+gen*inj+inj*loc+inj*sb+gen*loc+gen*sb+loc*sb,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#with 1 3-way interaction model.0 BEST
model.0<-
glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+loc*inj+sb*inj+sb*loc*gen
,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))



ndf1<-model.16$df.residual
dev1<-model.16$deviance

ndf1<-model.13$df.residual
dev1<-model.13$deviance

ndf1<-model.3$df.residual
dev1<-model.3$deviance

ndf1<-model.0$df.residual
dev1<-model.0$deviance


ck1<-dev1
cr1<-qchisq(1-0.05,ndf1)

      if(ck1>cr1)
            {u1[i]<-u1[i]+1}

d<-sam2[,i]
data.s<-cbind(expand.grid(inj=c("ni","yi"),
            sb=c("no","yes"), loc=c("urban","rural"),
            gen=c("female","male")), Fr=d)

#model for 3 2-way interactions
model.16<-glm(Fr~gen+loc+sb+inj+gen*sb+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model for 4 2-way interactions
model.13<-glm(Fr~gen+loc+sb+inj+gen*sb+gen*inj+loc*inj+sb*inj,
```

```
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#model for 5 2-way interactions
model.3<-glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*inj+sb*inj,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))

#model for 6 2-way interactions
model.11<-glm(Fr~gen+loc+sb+inj+gen*inj+inj*loc+inj*sb+gen*loc+gen*sb+loc*sb,

                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))


#with 1 3-way interactions model.0 BEST
model.0<-
glm(Fr~gen+loc+sb+inj+gen*loc+gen*sb+gen*inj+loc*sb+loc*inj+sb*inj+sb*loc*gen
,
                        family=poisson, data=data.s,
                        control=glm.control(epsilon=0.0000001,maxit=100))



ndf2<-model.16$df.residual
dev2<-model.16$deviance

ndf2<-model.13$df.residual
dev2<-model.13$deviance

ndf2<-model.3$df.residual
dev2<-model.3$deviance

ndf2<-model.0$df.residual
dev2<-model.0$deviance

ck2<-dev2
cr2<-qchisq(1-0.05,ndf2)

      if(ck2>cr2)
            {u2[i]<-u2[i]+1}

}

u1
table(u1)# number of reject in group1
u2
table(u2)

batahat<-sum(u1,u2)/(2*bt)

batahat #average value as estimated power
```

Appendix D: R code for calculate variance of the estimated power

```
############
#find J

c1<-0;c2<-0;c3<-0;c4<-0
SS1<-0;SS2<-0;SS3<-0;SS4<-0

for ( i in (1:999))
   { for ( j in ((i+1):1000))

{
s1<-u1[i]*u1[j] # product u1*u1(j>i)
SS1<-SS1+s1 #sum in group1 u1*u1(j>i)
c1<-c1+1  # number of product in group1

s2<-u1[i]*u2[j]#product u1*u2(u2>u1)
SS2<-SS2+s2 #sum in group2
c2<-c2+1


s3<-u2[i]*u1[j] #product u1*u2(u1>u2)
SS3<-SS3+s3
c3<-c3+1

s4<-u2[i]*u2[j] #sum in group2 (j>i)
SS4<-SS4+s4
c4<-c4+1 #number of product in group2

}
}

c1;c2;c3;c4
T<-sum(c1,c2,c3,c4)
SS<-sum(SS1,SS2,SS3,SS4)
J<-SS/T #first estimator
J

#################
#Find K
# product terms for K(u1i*u2i)
K<-0
for (k in (1:1000))
{
k<-u1[k]*u2[k]
K<-K+k
}
K<-K/bt #second estimator
K

Var.bata<-abs(J-K)# variance of batahat
Var.bata

batahat  #sum(u1+u2)/(2*bt)
```