

**DOKUZ EYLÜL UNIVERSITY  
GRADUATE SCHOOL OF NATURAL AND APPLIED  
SCIENCES**

**SOME MODEL MISSPECIFICATIONS IN  
LOGISTIC REGRESSION MODEL**

**by  
Suay EREEŞ**

**December, 2013**

**İZMİR**

# **SOME MODEL MISSPECIFICATIONS IN LOGISTIC REGRESSION MODEL**

**A Thesis Submitted to the  
Graduate School of Natural and Applied Sciences of Dokuz Eylül  
University In Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy in Statistics Program**

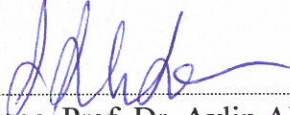
**by  
Suay EREEŞ**

**December, 2013**

**İZMİR**

## Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**SOME MODEL MISSPECIFICATIONS IN LOGISTIC REGRESSION MODEL**” completed by **SUAY EREEŞ** under supervision of **ASSOC. PROF. DR. AYLİN ALIN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.



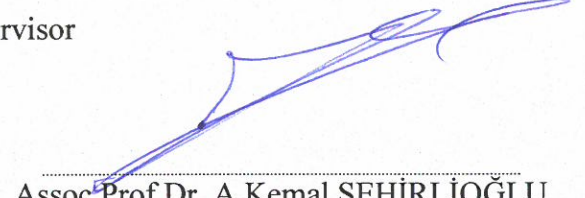
Assoc. Prof. Dr. Aylin ALIN

Supervisor



Prof. Dr. Serdar KURT

Thesis Committee Member



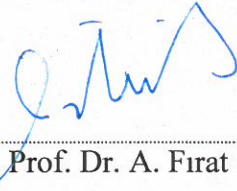
Assoc. Prof. Dr. A. Kemal ŞEHİRLİOĞLU

Thesis Committee Member



Prof. Dr. Onur KÖKSOY

Examining Committee Member



Assist. Prof. Dr. A. Fırat ÖZDEMİR

Examining Committee Member

Prof. Dr. Ayşe OKUR

Director

Graduate School of Natural and Applied Sciences

## ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor Assoc. Prof. Dr. Aylin ALIN for her guidance and valuable advices during my Ph.D. studies. This thesis has been more worthy thanks to her.

I would like to thank to my committee members Prof. Dr. Serdar KURT and Assoc. Prof. Dr. Ali Kemal ŞEHİRLİOĞLU for their significant and constructive suggestions. Their immense knowledge and inspirational discussions helped me improving my thesis.

I would like to thank to Dr. Charkaz AGHAYEVA for her generous helps by her extensive mathematical knowledge.

I will forever be thankful to my family, my parents, Sevgi and Abidin DÜNDAR for their endless love and encouragements in whole of my life, my older brothers Serdar DÜNDAR, Mustafa DÜNDAR and Aşkın DÜNDAR for always being lovely supporter to me and also for helps in my education life. Finally, I would like to thank to my dearest husband Erşans EREEŞ. He has inspirited me during my graduating studies being so lovely, encouraging and patient. He also has studied with me really hard during the stage of collecting and explaining real world data. I have completed this study with his faithful and valuable supports.

Suay EREEŞ

# **SOME MODEL MISSPECIFICATIONS IN LOGISTIC REGRESSION MODEL**

## **ABSTRACT**

Correct specification of the model is the most important assumption for the logistic regression model, as for all models. It means that the model has the correct functional form, does not include irrelevant variables and has all the relevant variables. Previous studies show that misspecification may cause undesirable results such as biased logistic regression coefficients, inefficient estimates, invalid statistical inferences and less efficient test statistics.

In this thesis, the effects of misspecification on asymptotic relative efficiency of various coefficients of determination are investigated. Misspecification types include using wrong functional form of explanatory variable, categorizing continuous explanatory variable and omitting the covariate. Unlike linear regression model, there is not only one coefficient of determination in logistic regression, which makes the results of this thesis more important. Simulation studies using bootstrap method and an application on agricultural data about land consolidation have been carried out to examine the efficiencies of these measures.

**Keywords:** Asymptotic relative efficiency, coefficients of determination, land consolidation, logistic regression, misspecification.

# LOJİSTİK REGRESYON MODELİNDE BAZI YANLIŞ MODEL TANIMLAMALARI

## ÖZ

Modelin doğru tanımlanması, diğer modeller için olduğu gibi, lojistik regresyon modeli için de en önemli varsayımdır. Bu, modelin doğru fonksiyonel fonksiyona sahip olması, gereksiz değişkenleri içermemesi ve tüm gerekli değişkenleri içermesi anlamına gelir. Önceki çalışmalar yanlış tanımlamanın yanlış lojistik regresyon katsayıları, etkin olmayan kestirimler, geçersiz istatistiksel çıkarımlar ve daha az etkin test istatistikleri gibi istenmeyen sonuçlara neden olabildiğini göstermektedir.

Bu tezde, yanlış tanımlamaların bazı belirtme katsayılarının asimtotik göreceli etkinliği üzerindeki etkileri araştırılmaktadır. Yanlış tanımlama türleri, açıklayıcı değişkenin yanlış fonksiyonel formunun kullanılmasını, sürekli açıklayıcı değişkenin kategorik hale getirilmesini ve eşdeğişken faktörün modele dahil edilmemesini içermektedir. Doğrusal regresyon modelinden farklı olarak, lojistik regresyonda sadece bir belirtme katsayısı yoktur. Bu durum, bu çalışmanın sonuçlarını daha önemli hale getirmektedir. Bootstrap yöntemi kullanılarak simulasyon çalışmaları ve arazi toplulaştırması ile ilgili tarımsal veri üzerine bir uygulama ölçülerin etkinliklerini incelemek için gerçekleştirilmiştir.

**Anahtar kelimeler:** Asimtotik göreceli etkinlik, belirtme katsayıları, arazi toplulaştırma, lojistik regresyon, yanlış tanımlama

# CONTENTS

	<b>Page</b>
THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGEMENTS .....	iii
ABSTRACT .....	iv
ÖZ.....	v
LIST OF FIGURES .....	viii
LIST OF TABLES .....	ix
<b>CHAPTER ONE – INTRODUCTION.....</b>	<b>1</b>
<b>CHAPTER TWO – MISSPECIFICATION IN LOGISTIC REGRESSION..</b>	<b>5</b>
2.1 Logistic Regression Model .....	5
2.2 Asymptotic Relative Efficiency.....	10
2.2.1 Asymptotic Relative Efficiency in Estimation .....	10
2.2.2 Asymptotic Relative Efficiency in Testing.....	14
2.3 Misspecification.....	20
2.3.1 Categorizing a Continuous Explanatory Variable .....	20
2.3.2 Omission of a Covariate.....	32
2.3.3 Mismodelling a Continuous Explanatory Variable .....	34
<b>CHAPTER THREE – COEFFICIENT OF DETERMINATION .....</b>	<b>38</b>
3.1 $R^2$ Statistics.....	38
3.2 Alternative $R^2$ Statistics.....	42
3.2.1 The Ordinary Least Squared $R^2$ .....	42
3.2.2 Squared Pearson Correlation Coefficient.....	44
3.2.3 Gini’s Concentration Measure .....	45

3.2.4 The Wald $R^2$ .....	46
3.2.5 McKelvey and Zavoina's Measure .....	47
3.2.6 The Contingency Coefficient $R^2$ .....	47
3.2.7 Adjusted Contingency Coefficient $R^2$ .....	48
3.2.8 The Likelihood Ratio $R^2$ .....	49
3.2.9 Geometric Mean Squared Improvement .....	51
3.2.10 Adjusted Geometric Mean Squared Improvement .....	53
<b>CHAPTER FOUR – NUMERICAL RESULTS.....</b>	<b>55</b>
4.1 Simulation Studies .....	55
4.2 Application on Real Land Consolidation Data .....	69
4.2.1 Introduction to Land Consolidation .....	69
4.2.2 Land Consolidation in Turkey .....	70
4.2.3 Application Case .....	71
<b>CHAPTER FIVE – CONCLUSIONS.....</b>	<b>78</b>
<b>REFERENCES .....</b>	<b>84</b>
<b>APPENDIX .....</b>	<b>93</b>



## LIST OF FIGURES

	<b>Page</b>
Figure 4.1 ARE's of each $R^2$ statistics under both correct and misspecified models for $n = 50$ .....	63
Figure 4.2 ARE's of each $R^2$ statistics under both correct and misspecified models for $n = 100$ .....	63
Figure 4.3 ARE's of three $R^2$ statistics with each other for $n = 50$ .....	68
Figure 4.4 ARE's of three $R^2$ statistics with each other for $n = 100$ .....	68
Figure 4.5 Histogram of the explanatory variable AR.....	73
Figure 4.6 Density plot of the explanatory variable AR.....	73

## LIST OF TABLES

	<b>Page</b>
Table 2.1	ARE when categorizing an explanatory variable $X$ into $k$ intervals... 23
Table 2.2	ARE when mismodelling a continuous explanatory variable $X$ ..... 35
Table 4.1	Number of categories and location of cutpoints..... 57
Table 4.2	The real values of $R^2$ for original model and the medians of $R^2$ for other models for $n = 50$ ..... 59
Table 4.3	The real values of $R^2$ for original model and the medians of $R^2$ for other models for $n = 100$ ..... 59
Table 4.4	ARE's of each $R^2$ statistics under both correct and misspecified models when $X$ has been mismodelled ..... 61
Table 4.5	ARE's of each $R^2$ statistics under both correct and misspecified models when $X$ has been categorized..... 61
Table 4.6	ARE's of each $R^2$ statistics under both correct and misspecified models when omitting $Z$ ..... 62
Table 4.7	ARE's of three $R^2$ statistics under correct model..... 64
Table 4.8	ARE's of three $R^2$ statistics with each other when $X$ has been mismodelled ..... 66
Table 4.9	ARE's of three $R^2$ statistics with each other when categorizing $X$ .... 66
Table 4.10	ARE's of three $R^2$ statistics with each other when omitting $Z$ ..... 67
Table 4.11	Descriptions for land consolidation data ..... 72
Table 4.12	$R^2$ values associated with all models for land consolidation data..... 75
Table 4.13	ARE's of each $R^2$ statistics on the base of $\ln(AR)$ ..... 76
Table 4.14	ARE's of each $R^2$ statistics on the base of $\ln(AR)$ for categorizing .. 76
Table 4.15	ARE's between each $R^2$ statistics..... 77

## **CHAPTER ONE**

### **INTRODUCTION**

Model specification is the first and the most crucial stage of regression analysis. However, misspecification is a general problem of estimation and interpretation in research studies, since it is not possible all the time to build the model perfectly with all the relevant variables and also with their correct functional form. The model is only assumed to be correct or at least as closer to the correct than the others. In many situations, the model is determined without complete confidence. All other regression assumptions follow from the requirement that the model is correctly specified. A good knowledge of theory, an accurate understanding of what the model implies can help to avoid the model misspecification.

Misspecification has three aspects in general: (1) The omission of some variables that affect the dependent variable may cause an omitted variables bias. In linear regression models, if the omitted covariates are independent of the included variables, then model misspecification due to omission does not cause an omitted variable bias. However, as shown by Neuhaus (1998) in logistic regression models, omitting covariates associated with the dependent variable, even if they are independent of the included variables, causes seriously downward estimates of regression coefficients. (2) Functional form of an explanatory variable should be determined carefully as they affect the data analysis. Incorrect functional forms lead incorrect conclusions. Simple regression models do not always represent the complex structure of the data, sufficiently. Some transformations of the continuous explanatory variables may be required to improve the model's fit to the data. Otherwise the results of poor fit and biased estimates become unavoidable. Kay and Little (1987) studied on the transformations based on the distribution of explanatory variable in logistic models. Box and Cox (1964) studied on the analysis of transformations in linear regression. (3) In especially medical researches, with the intention of simplifying the interpretation of models, categorization or grouping may be preferred, frequently. However this is the most encountered misspecification type

causing some problems such as efficiency losses in test statistics. Therefore, before categorizing some issues should be remembered by the researcher. For example, the number of categories and the distribution of the explanatory variable have a big importance for removing or at least decreasing the efficiency losses. Various authors have paid attention on this subject in many years. Bofinger (1970) has recommended a method of maximizing the correlation of categorized observations to select the cutpoints. Jarque (1981) has studied on how to attain efficient estimates in regression analysis when an explanatory variable has been categorized. O'Brien (2004) has presented an approach based on a formula of an efficient nonparametric estimate of the regression function for cutpoint selection. Prais & Aitchison (1954) have noted that the estimators of a regression model become unbiased and also that there is an information loss because of categorization. Cox (1957) defined an information loss measure from categorizing for choosing cutpoints for different size of categories due to the concept of asymptotic relative efficiency (ARE). Connor (1972) and Lagakos (1988b) have investigated ARE of test statistics with categorized explanatory variable which has up to 6 optimal categories and which has the distributions of uniform, normal and exponential with parameter  $\lambda = 1$ . But, the explanatory variable may have an exponential distribution with parameter that differs from one. In this case, how to obtain the cutpoints and ARE values will be discussed in Chapter 2.

The decision of the appropriate statistic is important for involving to the analysis. The concept of ARE is a useful and most frequently used technique for the comparison of related statistics evaluating their performances. It provides a previous knowledge about information loss. The association between reducing the information loss and maximizing ARE will be explained in Chapter 2 in more detail. ARE is based on the ratio of variances of two associated statistics. Pitman (1949) introduced the earliest approach to ARE. Stuart (1954) studied asymptotic relative efficiencies of distribution free tests of randomness using Pitman's proposes. Amemiya & Powel (1983) and Efron (1975) compared logistic regression and discriminant analysis with ARE. Saikkonen (1989) examined the effect of the misspecification on the three classical test statistics that are likelihood ratio, Lagrange multiplier and Wald statistics in terms of ARE. Begg & Lagakos (1990, 1993), Lagakos (1988a) and

Tosteson & Tsiatis (1988) particularly studied on the ARE of tests of association when explanatory variables have been misspecified in logistic regression models. In this thesis, looking with different perspectives, we will investigate the effects of misspecification on the ARE of various coefficients of determination ( $R^2$ ) in logistic regression model.

In ordinary least squares (OLS),  $R^2$  statistic represents the proportion of variance explained in the dependent variable. It is not the valid interpretation for logistic regression, since logistic regression concerns about the probability of a given dependent variable. For the logistic regression model, so many derived  $R^2$  statistics in accordance with different perspectives have been proposed in recent years. In Chapter 3, some reasons of derivation of various  $R^2$  statistics will be presented, in more detail. Kvalseth (1985) described eight criteria for a good statistic (Menard, 2000). There are different  $R^2$  statistics proposed in the literature satisfying some of these properties. There are at least ten different  $R^2$  statistics (Mittlböck & Schemper, 1996). So analysts may face the difficulty of choosing the convenient  $R^2$  statistic among all. Hence, studying their performances becomes a very important issue especially under misspecification. It is well known that these statistics are utility to measure how well a model fits the data, however it should be remembered that to judge the usefulness of the model based solely on these values is dangerous. There are other analyses to be taken into consideration such as the values of goodness of fit statistics (likelihood ratio statistic, Pearson chi-square).

Binary logistic regression models where dependent variable has only two different values have been applied on many fields. For example, agricultural data sets have been studied by Battaglin & Goolsby (1996), Cimpoieş (2007), Lerman & Cimpoieş (2006), Minetos & Polyzos (2009), Msoffe et.al. (2011), Mueller et. al. (2005), Raut, Sitaula, Vatn, & Paudel (2011), Schroeder et.al. (2001) and Zhang & Zhao (2013). In this thesis, for the purpose of demonstrating the effects of misspecification on the ARE of  $R^2$  in logistic regression model, an application on land consolidation will be performed. Nowadays, consolidation activities have been carried out, extensively, in many countries around the world. In the beginning of the work, the opinions of the

peasants should be determined cautiously for planning parcels. To be able to predict willingness of peasants for consolidation will help the researcher to have an idea about the behaviors of peasants statistically. So that willingness of peasants will also be investigated using this method.

The thesis proceeds as follows. In Chapter 2, after giving a general overview to logistic regression model, the concept of ARE will be presented and general formula for ARE for the case of categorizing the explanatory variable  $X$  which has exponential or Weibull distribution will be introduced. Chapter 2 will also include the types of misspecification. To compare the behaviors in terms of efficiency under misspecification, three well-known and favorite  $R^2$  statistics will be explained in Chapter 3. These statistics are the ones already included in most logistic regression outputs in popular statistical software packages such as SPSS, SAS and STATA. The illustration of the effects of misspecification on the efficiency through simulation studies and the real data set of land consolidation will be given in Chapter 4. Finally, concluding remarks will be presented in Chapter 5.

## CHAPTER TWO

### MISSPECIFICATION IN LOGISTIC REGRESSION

For model building stage in logistic regression, the most important assumption is that the model is correctly specified. It means that the model has the correct functional form, does not include irrelevant variables and has all the relevant variables. Misspecification may cause undesirable results such as biased logistic regression coefficients, inefficient estimates, invalid statistical inferences and less efficient test statistics (Lagakos, 1988b; Menard, 2000). Nevertheless, misspecifying an explanatory variable is a common problem in logistic regression, particularly in research studies. Therefore, there are numerous studies in literature regarding this issue for both linear and logistic models such as Adewale & Wiens (2009), Schafer (1987), Stefanski and Carroll (1985), White (1982).

After introducing the logistic regression model in the subsequent section, asymptotic relative efficiency will be explained in detail as an introduction to misspecification in Section 2.2. Then, in Section 2.3, the reasons and consequences of various misspecification types will be described. Categorizing a continuous independent variable, omission of an explanatory variable from a regression and finally consequences of using incorrectly specified model will be given in separate subsections. In this thesis, we are only interested in binary logistic regression where response takes only two different values. The term “logistic regression” will refer only to the binary case.

#### 2.1 Logistic Regression Model

In simple linear regression analysis, we accept that variables are linearly related and it is possible to calculate the strength of the linear relationship between variables as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n \quad (2.1)$$

where  $Y_i$  and  $X_i$  are , respectively, the dependent and explanatory variable for ith observation.  $X_i$  is assumed to be fixed.  $\beta_0$  and  $\beta_1$  are parameters whose values are being estimated and  $\varepsilon$  which is an independent random variable normally distributed with parameters 0 and  $\sigma^2$  is called the error term. Since  $E(\varepsilon_i) = 0$ ,

$$E(Y_i) = \beta_0 + \beta_1 X_i. \quad (2.2)$$

Considering  $Y_i$  is binary taking on the values of only 0 or 1, the probability that  $Y_i = 1$  is assumed to be  $\pi(X_i)$  ( $P(Y_i = 1) = \pi(X_i)$ ) and the probability that  $Y_i = 0$  is assumed to be  $1 - \pi(X_i)$  ( $P(Y_i = 0) = 1 - \pi(X_i)$ ).

In defining probabilities like  $\pi(X_i)$ ,  $X_i$  is used to emphasize that this probability is a function of the explanatory variables. For sake of simplicity,  $\pi_i$  will be used instead of  $\pi(X_i)$ , thereafter. For a binary random variable  $Y_i$ ,

$$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i. \quad (2.3)$$

Hence, from Equation (2.2) and Equation (2.3), the expected value of  $Y_i$  is

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i. \quad (2.4)$$

Therefore, the expected value of response always represents the probability that response is equal to 1 for all given levels of explanatory variables.

When response  $Y_i$  is binary, linear regression assumptions are violated and some important differences between linear and logistic regressions arise. First of all, for binary responses, the condition that the errors follow normal distribution is not satisfied, because the error  $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - \pi_i$  takes on only two values. If  $Y_i = 1$ , then  $\varepsilon_i = 1 - \pi_i$  with probability  $\pi_i$  and if  $Y_i = 0$ , then  $\varepsilon_i = -\pi_i$  with



probability  $1 - \pi_i$ . Therefore, it is clear that the error does not follow a normal distribution, but follows a distribution with zero mean and variance  $\pi_i(1 - \pi_i)$  which is a sign of a violation of linear regression assumption which requires the constancy of the error variance. Since  $\pi_i$  depends on  $X_i$  and  $\varepsilon_i$  depends on  $\pi_i$ ,  $\sigma^2(\varepsilon_i)$  varies by different levels of explanatory variables and so is not a constant. The most important difference between linear and logistic regression models is the range for the response's expected value. In linear regression, this expected value takes on any value within the range from  $-\infty$  to  $+\infty$ . On the other hand, since the response function represents the probabilities in logistic regression, its expected value should take on the values of only greater than or equal to zero or less than or equal to 1. However, using the linear function given in Equation (2.4) may give values outside of this range. To solve this problem, several transformations may be used. The most popular one among these is the logistic function.

The logistic function has the following form:

$$E(Y_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}, \quad i = 1, \dots, n \quad (2.5)$$

which is a nonlinear model in parameters.

Using Equation (2.5), the formula for the odds of the success ( $Y_i = 1$ ) is obtained as below.

$$\pi_i [1 + \exp(\beta_0 + \beta_1 X_i)] = \exp(\beta_0 + \beta_1 X_i)$$

So,

$$\begin{aligned} \pi_i &= [\exp(\beta_0 + \beta_1 X_i)] - \pi_i [\exp(\beta_0 + \beta_1 X_i)] \\ &= [\exp(\beta_0 + \beta_1 X_i)] [1 - \pi_i] \end{aligned} \quad (2.6)$$

Therefore, the odds that  $Y_i = 1$  is expressed as

$$\frac{\pi_i}{1 - \pi_i} = \exp(\beta_0 + \beta_1 X_i). \quad (2.7)$$

Taking the logarithm of Equation (2.7), we obtain a model linear in parameters and may take any values within the range of  $(-\infty, \infty)$  and define as

$$\begin{aligned} \text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1 - \pi_i}\right) \\ &= \beta_0 + \beta_1 X_i \end{aligned} \quad (2.8)$$

where  $i=1, \dots, n$ . This expression is called as logit function. Thus, the logit transformation helps linearize the nature of the nonlinear relationship between explanatory variable and the probability of dependent variable.

Maximum likelihood estimation is the mostly used technique to estimate the parameters for the logistic regression model. Since each  $Y_i$  observation is an independent Bernoulli random variable, their joint distribution function equals

$$f(Y_1, \dots, Y_n) = \prod_{i=1}^n f_i(Y_i) = \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1 - Y_i}, \quad (2.9)$$

which is also the likelihood function of the parameters  $\beta$  represented as  $L(\beta)$ . It would be easier to work with the logarithm of the likelihood function.

$$\begin{aligned}
\log_e L(\beta) &= \log_e \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} = \sum_{i=1}^n Y_i \log_e (\pi_i) + \sum_{i=1}^n (1 - Y_i) \log_e (1 - \pi_i) \\
&= \sum_{i=1}^n Y_i \log_e (\pi_i) + \sum_{i=1}^n \log_e (1 - \pi_i) - \sum_{i=1}^n Y_i \log_e (1 - \pi_i) \\
&= \sum_{i=1}^n Y_i [\log_e (\pi_i) - \log_e (1 - \pi_i)] + \sum_{i=1}^n \log_e (1 - \pi_i)
\end{aligned}$$

Finally, log-likelihood function is

$$\begin{aligned}
\log_e L(\beta) &= \sum_{i=1}^n \left[ Y_i \log_e \left( \frac{\pi_i}{1 - \pi_i} \right) \right] + \sum_{i=1}^n \log_e (1 - \pi_i) \\
&= \sum_{i=1}^n Y_i (\beta_0 + \beta_1 X_i) - \sum_{i=1}^n \log_e [1 + \exp(\beta_0 + \beta_1 X_i)]
\end{aligned} \tag{2.10}$$

To find the value of  $\beta$  that maximizes  $L(\beta)$ , we differentiate Equation (2.10) with respect to  $\beta_0$  and  $\beta_1$  then set the resulting expressions equal to zero. But since the equations do not have closed form, iterative methods are used to obtain estimates. When we have more than one explanatory variable, the model in Equation (2.8) takes the following form.

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \tag{2.11}$$

The log-likelihood function for this multiple binary logistic regression model becomes as

$$\log_e L(\beta) = \sum_{i=1}^n Y_i \left( \beta_0 + \sum_{j=1}^k \beta_j X_{ij} \right) - \sum_{i=1}^n \log_e \left[ 1 + \exp \left( \beta_0 + \sum_{j=1}^k \beta_j X_{ij} \right) \right] \tag{2.12}$$

## 2.2 Asymptotic Relative Efficiency

“For two competing statistical procedures  $A$  and  $B$ , suppose that a desired performance criterion is specified and let  $n_1$  and  $n_2$  be the respective sample sizes at which the two procedures ‘perform’ equivalently with respect to the adopted criterion.” (Serfling, 1980, p. 50). The ratio of these sample sizes is called relative efficiency of procedures. If this ratio approaches to some limit, then this limit value is named as asymptotic relative efficiency (ARE).

There are two fields that ARE is taken into consideration: ARE in estimation and ARE in testing. At the following subsections, these issues will be discussed.

### 2.2.1 Asymptotic Relative Efficiency in Estimation

Properties of estimators are considered for finite samples and infinite samples. For finite sample the estimator with a smaller variance is generally said to be efficient. However, qualifying an estimator as efficient only on the basis of variance is not reasonable. Not only dispersion but also expected value of an estimator should be calculated because of considering the property of unbiasedness, since both bias and variance are important and need to be as small as possible to achieve good estimation performance. In this sense, it will be more convenient to use mean square error (MSE) as a combination of variance and bias. Let  $T$  be an estimator of  $\theta$ .

$$\begin{aligned}MSE &= E(T - \theta)^2 \\ &= \sigma^2(T) + (\text{Bias}(T))^2\end{aligned}\tag{2.13}$$

where  $\sigma^2$  represents the variance. It is clear that for an unbiased estimator  $E(T) = \theta$  and so the mean square error equals the variance. In such case, a judgment can be made in accordance with variance and therefore it is said that unbiased estimators with the smallest variance are called efficient.

The asymptotic property of efficiency is considered when sample size becomes infinitely large. In such cases, since evaluations are much easier than the ones for finite-samples and often possible only asymptotically, the properties of an estimator are examined asymptotically. In this regard, it is said that a maximum likelihood estimate is asymptotically efficient, if its limiting distribution is asymptotically normal around the parameter value with a variance which achieves the Cramér-Rao lower bound. In this sense, under some general mild conditions, maximum likelihood estimates are asymptotically efficient. Let  $X_1, \dots, X_n$  be a sample with probability density function  $f(X; \theta)$  and let  $T_n$  based on this sample with size  $n$  be a sequence of estimators for a parameter  $\tau(\theta)$ , then if  $\sqrt{n}[T_n - \tau(\theta)] \rightarrow N(0, \sigma_\theta^2(T_n))$  and

$$\sigma_\theta^2(T_n) = \frac{\left(\frac{d}{d\theta} \tau(\theta)\right)^2}{E_\theta \left\{ \left(\frac{\partial}{\partial \theta} \log f(X; \theta)\right)^2 \right\}} \quad (2.14)$$

so the asymptotic variance of  $T_n$  achieves the Cramér-Rao lower bound, then it satisfies the conditions of being asymptotically efficient (Casella & Berger, 2002; Cox & Hinkley, 1974).

An estimator is asymptotically unbiased if its asymptotic mean is equal to the true value that is  $\lim_{n \rightarrow \infty} E(T_n) = \theta$ . However this is not true for asymptotic variance. Since when sample size increases an estimator often accumulate to only one point and so  $\sigma^2(T_n)$  approaches to zero, asymptotic variance cannot be calculated by limiting variance of estimator as  $n \rightarrow \infty$ . Nevertheless, if it is required to calculate the limit of the variance, a constant  $k_n$  should be inserted to force it to a limit. In other words, if  $\lim_{n \rightarrow \infty} k_n \sigma^2(T_n) = \tau^2 < \infty$ , then  $\tau^2$  is said to be the limiting variance of  $T_n$ .

On the other hand, the asymptotic variance is defined as the variance of the limit distribution of the estimator. Therefore, if  $k_n(T_n - \tau(\theta)) \rightarrow N(0, \sigma_{T_n}^2)$ , then  $\sigma_{T_n}^2$  is said to be the asymptotic variance or variance of the limit distribution of  $T_n$  and is defined by Hanushek & Jackson (1977) as

$$\sigma_{T_n}^2 = \left(\frac{1}{n}\right) \lim_{n \rightarrow \infty} E \left\{ \sqrt{n} (T_n - \lim E(T_n)) \right\}^2 \quad (2.15)$$

So it is obvious that the asymptotic variance is the expected squared deviation of  $T_n$  about its asymptotic mean. If  $T_n$  is asymptotically unbiased and asymptotically normal with mean  $\theta$  and variance  $\sigma_{T_n}^2$ , then asymptotic efficiency of  $T_n$  is

$$\begin{aligned} e(T_n) &= \lim_{n \rightarrow \infty} (\sigma_{T_n}^2 i(\theta))^{-1} \\ &= \lim_{n \rightarrow \infty} \frac{(i(\theta))^{-1}}{\sigma_{T_n}^2} \end{aligned} \quad (2.16)$$

where  $i(\theta) = E \left\{ -\frac{\partial^2 \log f(y; \theta)}{\partial \theta^2}; \theta \right\}$  and is called the Fisher information about  $\theta$  (Cox & Hinkley, 1974).

“The efficiency of the MLE becomes important in calibrating what we are giving up if we use an alternative estimator” (Casella & Berger, 2002, p. 477). Because of simplicity and robustness, sometimes different alternative estimators are considered. It is important to find out which one is more convenient to use. In the sense that, for competing two estimators  $T_1$  and  $T_2$  with following limiting distributions

$$\sqrt{n}(T_{1n} - \tau(\theta)) \rightarrow N(0, \sigma_1^2)$$

$$\sqrt{n}(T_{2n} - \tau(\theta)) \rightarrow N(0, \sigma_2^2)$$

the asymptotic relative efficiency of  $T_2$  with respect to  $T_1$  is the ratio of their asymptotic efficiencies and denoted as

$$ARE(T_2, T_1) = \frac{e(T_2)}{e(T_1)} = \frac{\sigma_1^2}{\sigma_2^2} \quad (2.17)$$

ARE may take on the values between zero and infinity. The estimator  $T_1$  is preferred if this ratio is less than 1, on the other hand the ratio greater than 1 indicates that  $T_2$  is more efficient than  $T_1$ .

To better understand the ARE of two estimators, the comparison of mean ( $\bar{X}$ ) and median ( $\tilde{X}$ ) may be given as an example. Mean and median both tries to measure the central tendency so it is remarkable that these statistics are alternatives to each other. In this regard, ARE is a useful way of comparing performances in terms of efficiency. Central limit theorem states that the sample means of random samples from a population with mean  $\mu$  and finite standard deviation  $\sigma$  have mean  $\mu$  and finite standard deviation  $\sigma/\sqrt{n}$ , furthermore with sufficiently large sample sizes, the sampling distribution of mean will approximately be normal with the same parameters, regardless of how the population values are distributed. By the way, for the same population, the median has approximately normal distribution with  $\mu$  mean and  $\frac{1}{2f(\mu)\sqrt{n}}$  standard deviation, where  $f(\mu)$  is continuous density function

(Panik, 2005). Since  $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$ , the variance of median is equal to  $\frac{\pi\sigma^2}{2n}$ .

Hence, ARE of median versus to mean as the ratio of their variances from Equation (2.17) (Serfling, 2011)

$$ARE(\tilde{X}, \bar{X}) = \frac{\sigma^2(\bar{X})}{\sigma^2(\tilde{X})} = \frac{\sigma^2/n}{\pi\sigma^2/2n} = \frac{2}{\pi} = 0.64.$$

Since the value of ARE is less than 1, it is said that mean is more efficient than median. In other words, mean needs 64% as many observations as the median to estimate population mean with the same efficiency, according to the definition of relative efficiency given in Section 2.2.

### ***2.2.2 Asymptotic Relative Efficiency in Testing***

The concept of asymptotic relative efficiency is a useful technique for the comparison of test sequences and often called Pitman efficiency since calculations are based on his theorem. Pitman (1949) introduced the earliest approach to ARE in testing. Serfling (1980) mentioned Pitman approach is widely applicable since the only major requirement is the information about asymptotic distribution of the test statistic (Lachin, 2000).

“Given two tests of the same size of the same statistical hypothesis, the relative efficiency of the second test with respect to the first is given by the ratio  $n_1/n_2$  where  $n_2$  is the sample size of the second test required to achieve the same power for a given alternative as is achieved by the first test with respect to the same alternative when using a sample of size  $n_1$ ” (Noether, 1955, p. 64). Therefore, relative efficiency requires identical alternatives but does not require a limited or a specific alternative, so this approach can be applied, in any case (Serfling, 1980).

Consider a test for the null hypothesis  $H_0 : \theta = \theta_0$  against the alternatives  $H_1 : \theta > \theta_0$  based on  $n$  observations and based on the statistic  $T_n = T(x_1, \dots, x_n)$ . Let  $E(T_n) = \mu_n(\theta)$  and  $\sigma^2(T_n) = \sigma_n^2(\theta)$ . Consider the sequence of alternatives is  $H_1 : \theta_n = \theta_0 + k/n^\delta$ , where  $k$  is an arbitrary positive finite constant and  $\delta > 0$  (Eeden, 1963, Noether, 1955). Alternative  $\theta = \theta_n$  changes with the sample size  $n$  and  $\lim_{n \rightarrow \infty} \theta_n = \theta_0$ . Assume that the following conditions are satisfied:



A.  $\mu'_n(\theta_0) = \dots = \mu_n^{(m-1)}(\theta_0) = 0, \mu_n^{(m)}(\theta_0) > 0$

Suppose that the derivatives exist.

B.  $\lim_{n \rightarrow \infty} \frac{n^{-m\delta} \mu_n^{(m)}(\theta_0)}{\sigma_n(\theta_0)} = c > 0$

C.  $\lim_{n \rightarrow \infty} \frac{\mu_n^{(m)}(\theta_n)}{\mu_n^{(m)}(\theta_0)} = 1$

D.  $\lim_{n \rightarrow \infty} \frac{\sigma_n(\theta_n)}{\sigma_n(\theta_0)} = 1$

E. The distribution of  $[T_n - \mu_n(\theta)]/\sigma_n(\theta)$  tends to the standard normal distribution, uniform in  $\theta$ , with  $\theta_0 \leq \theta \leq \theta_0 + d$  for some  $d > 0$ .

The condition E can be replaced by the following.

E'. The distribution of  $[T_n - \mu_n(\theta_n)]/\sigma_n(\theta_n)$  tends to the standard normal distribution, both under the alternative  $H_1 : \theta_n = \theta_0 + k/n^\delta$  and the null hypothesis  $\theta_n = \theta_0$ .

*Pitman's Theorem:* (Pitman, 1949) The asymptotic relative efficiency of two tests satisfying the above conditions with  $\delta_1 = \delta_2$  and  $m_1 = m_2$  is equal to the limit of the ratio of the efficacies of the two tests.

Pitman proved this theorem by following calculations. Let  $T_{1n}$  and  $T_{2n}$  be two test statistics of tests with the same alternative  $H_1 : \theta_n = \theta_0 + k/n^\delta$ , since we assume that  $\delta_1 = \delta_2 = \delta$ . These two tests must have the same power with respect to the same alternatives, as mentioned in definition. So, the alternatives are the same if

$$\frac{k_1}{n_1^\delta} = \frac{k_2}{n_2^\delta} \quad (2.18)$$

From Equation (2.18) the ratio of the sample sizes is

$$\frac{n_1}{n_2} = \left( \frac{k_1}{k_2} \right)^{1/\delta} \quad (2.19)$$

Noether (1955) proved that the power of a test is asymptotically  $L_n(\theta_n) \approx \phi\left(\lambda_\alpha - \frac{k^m c}{m!}\right)$  where  $\phi(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} \exp\left(-\frac{1}{2}x^2\right) dx$  and  $\phi(\lambda_\alpha) = \alpha$ . So two tests have the same power if

$$\frac{k_1^{m_1} c_1}{m_1!} = \frac{k_2^{m_2} c_2}{m_2!} \quad (2.20)$$

If  $m_1 = m_2 = m$ , then from Equation (2.19)

$$\frac{n_1}{n_2} = \left( \frac{k_1}{k_2} \right)^{1/\delta} = \left( \frac{c_2}{c_1} \right)^{1/m\delta} \quad (2.21)$$

Substituting  $c_1$  and  $c_2$  with the one given in condition B with respect to two tests

$$\left( \frac{c_2}{c_1} \right)^{1/m\delta} = \lim_{n \rightarrow \infty} \frac{n^{-1} \left[ \frac{\mu_{2n}^{(m)}(\theta_0)}{\sigma_{2n}(\theta_0)} \right]^{1/m\delta}}{n^{-1} \left[ \frac{\mu_{1n}^{(m)}(\theta_0)}{\sigma_{1n}(\theta_0)} \right]^{1/m\delta}} \quad (2.22)$$

Pitman called the quantity  $R_{in}^{1/m\delta}(\theta_0)$  the efficacy of the  $i$ th test where  $R_{in}(\theta) = \frac{\mu_{in}^{(m)}(\theta)}{\sigma_{in}(\theta)}$ , so the limit of the ratio of the efficacies of the two tests is the asymptotic relative efficiency of these tests as

$$ARE(T_2, T_1) = \lim_{n \rightarrow \infty} \frac{R_{2n}^{1/m\delta}(\theta_0)}{R_{1n}^{1/m\delta}(\theta_0)} \quad (2.23)$$

If  $m\delta = 1/2$ , for  $m = 1$  and  $\delta = 1/2$ , then

$$\begin{aligned} ARE(T_2, T_1) &= \lim_{n \rightarrow \infty} \frac{R_{2n}^2(\theta_0)}{R_{1n}^2(\theta_0)} \\ &= \lim_{n \rightarrow \infty} \left\{ \left[ \frac{\mu'_{2n}(\theta_0)}{\sigma_{2n}(\theta_0)} \right]^2 \left[ \frac{\sigma_{1n}(\theta_0)}{\mu'_{1n}(\theta_0)} \right]^2 \right\} \end{aligned} \quad (2.24)$$

This is the general definition of asymptotic relative Pitman efficiency. In this regard, only if

$$\lim_{n \rightarrow \infty} \frac{\mu_{2n}^{(m)}(\theta_0)}{\mu_{1n}^{(m)}(\theta_0)} = 1 \quad (2.25)$$

then ARE reduces to

$$ARE(T_2, T_1) = \lim_{n \rightarrow \infty} \frac{\sigma_{1n}^2}{\sigma_{2n}^2} \quad (2.26)$$

Therefore, if Equation (2.26) satisfies, ARE of two test statistics equals the limit of their variances. Some of authors addressed the relation between Pitman's ARE of

a test versus another and the correlation coefficient of their test statistics, for example Hájek (1962) showed this relation for rank-orders tests (Eeden, 1963).

*Theorem:* Assume that  $\rho(\theta)$  is the asymptotic correlation coefficient between test sequences  $T_{1n}$  and  $T_{0n}$  satisfying all the Pitman's conditions and  $\rho(\theta_n) \rightarrow \rho(\theta_0)$ , so that  $ARE(T_1, T_0) = \rho^2$  (Eeden, 1963).

Proof of this theorem starts with considering tests of the form as  $T_{\lambda n} = (1 - \lambda)T_{0n} + \lambda T_{1n}$  satisfy the Pitman's conditions, where  $\lambda$  is a constant and  $0 \leq \lambda \leq 1$  (Eeden, 1963). From this point, Eeden (1963) and Serfling (1980) continued to the proof through two different ways. Serfling (1980) assumed that  $T_\gamma$  is a best test maximizing  $ARE(T_0, T_\lambda)$  for  $\gamma = \frac{c_0 - \rho c_1}{(1 - \rho)(c_0 + c_1)}$ . When both nominator and denominator are divided by  $c_1$ , it is obtained that

$$\gamma = \frac{[ARE(T_0, T_1)]^{1/2} - \rho}{(1 - \rho)\{1 + [ARE(T_0, T_1)]^{1/2}\}},$$

where  $[ARE(T_0, T_1)]^{1/2} = \frac{c_0}{c_1}$ .

Since  $\mu_{\lambda n}(\theta) = (1 - \lambda)\mu_{0n}(\theta) + \lambda\mu_{1n}(\theta)$  so the first order derivative of this mean is

$$\mu'_{\lambda n}(\theta_0) = (1 - \lambda)\mu'_{0n}(\theta_0) + \lambda\mu'_{1n}(\theta_0) \sim n^{1/2}((1 - \lambda)c_1 + \lambda c_0)$$

and the variance of test is

$$\begin{aligned} \sigma_{\lambda n}^2(\theta) &= (1 - \lambda)^2 \sigma_{0n}^2(\theta) + \lambda^2 \sigma_{1n}^2(\theta) + 2\lambda(1 - \lambda)\sigma_{0n}(\theta)\sigma_{1n}(\theta)\rho(\theta) \\ &= (1 - \lambda)^2 + \lambda^2 + 2\lambda(1 - \lambda)\rho \end{aligned}$$

Therefore, from Pitman's condition B,  $c_\lambda$  is

$$c_\lambda = \frac{n^{1/2} [(1-\lambda)c_1 + \lambda c_0]}{n^{1/2} [(1-\lambda)^2 + \lambda^2 + 2\lambda(1-\lambda)\rho]^{1/2}} \quad (2.27)$$

If  $\lambda$  is replaced with  $\gamma$  in Equation (2.27), then for the best test

$$ARE(T_0, T_\gamma) = 1 + \frac{\{[ARE(T_0, T_1)]^{1/2} - \rho\}^2}{1 - \rho^2} \quad (2.28)$$

If  $T_0$  is a best test, then  $ARE(T_0, T_1) = 1$ , so we have  $ARE(T_1, T_0) = \rho^2$ .

Eeden (1963), with a different perspective, in order to proof the theorem, implied

that if  $T_{0n}$  is a best test, then so as to maximize  $ARE(T_0, T_\lambda) = \left(\frac{c_0}{c_\lambda}\right)^2$ ,

$$c_\lambda \leq c_0 \quad \text{for every } \lambda \quad (2.29)$$

Substituting  $c_\lambda$  in (2.29)

$$\frac{[(1-\lambda)c_1 + \lambda c_0]^2}{(1-\lambda)^2 + \lambda^2 + 2\lambda(1-\lambda)\rho} - c_0^2 \leq 0. \quad (2.30)$$

It follows that

$$[\lambda c_0 + (1-\lambda)c_1]^2 - \lambda^2 c_0^2 - (1-\lambda)^2 c_0^2 - 2\lambda(1-\lambda)\rho c_0^2 \leq 0. \quad (2.31)$$

After some mathematical calculations,

$$\lambda^2 [c_1^2 - c_0^2 - 2c_0(c_1 - \rho c_0)] - 2\lambda [c_1^2 - c_0^2 - c_0(c_1 - \rho c_0)] + c_1^2 - c_0^2 \leq 0 \quad (2.32)$$

which is simplified with  $c_0^2(c_1 - \rho c_0)^2 \leq 0$ , since  $c_0$  is positive,

$$\rho = \frac{c_1}{c_0} = [ARE(T_1, T_0)]^{1/2}. \quad (2.33)$$

Begg and Lagakos (1990, 1993), Lagakos (1988a) and Tosteson and Tsiatis (1988) particularly have showed great interest in the asymptotic relative efficiency of tests of association when explanatory variables have been misspecified or omitted, in logistic regression models, using these findings.

### 2.3 Misspecification

Correct specification of the model is the most important assumption for the logistic regression model. The violation of this assumption can occur due to: omission of an important variable, using a wrong functional form, inclusion of irrelevant variables. Without correct specification we will have biased logistic regression coefficients and less efficient estimates as well as invalid statistical inferences. However, misspecification is not an uncommon problem in practice, since we never know what the correct model is in real and we only assume that the model is correctly specified.

The types of misspecification including the discretizing a continuous explanatory variable, omission of a covariate, using wrong functional form of an explanatory variable will be presented, at the following subsections.

#### 2.3.1 *Categorizing a Continuous Explanatory Variable*

In medical and agricultural economics researches, particularly, when multiple logistic regression models are built, categorizing seems useful for simplifying the interpretation of models or sometimes the only available information about the

explanatory variable is already categorized. The most common forms of categorization are dichotomization and trichotomization, such as categorizing general health as good and bad or categorizing blood pressure as low, medium and high. However, though its simplicity and preferableness, for whatever reason, categorizing causes some problems in the analysis, such as misspecification error and loss in efficiency for test statistics. Prais and Aitchison (1954) studied on grouping in regression analysis and mentioned that regression estimates based on the grouped data will be unbiased and their variances will always be larger than the ones based on the ungrouped observations and this is caused by manner of grouping. They noted that the correlation coefficient for categorized data is an “unsatisfactory estimator of the correlation in the population”. Cramer (1964) agreed with them and added that the correlation coefficient based on the categorized data have unreliable results since it leads larger values than the one based on the original observations. He also indicated that groups should be defined as the ones minimizing the “within group sum of squares” of the variable so the efficiency of the categorized estimator will be maximized. Jarque (1981) added that, as grouping, all information on the variables should be included to the regression analysis for efficient estimates. Consequently, it is clear that since categorizing causes some loss of information, it is worthwhile to determine categories in a way that reduces this loss.

It is important to decide the number of categories ( $k$ ) to choose and the place of the category cutpoints, when categorizing an explanatory variable  $X$ . The choice of a cutpoint may be based on expert’s knowledge about the issue or experience or the results of other similar studies. However, sometimes cutpoints are not readily available. In these cases, statistical methods should be used to determine them. An unduly broad or unduly narrow range of categories causes that individuals with different levels of risk are in the same category. Thereby, there is quite likely loss of information. So, the researcher should be careful so as to determine the cutpoints that make this loss as small as possible. Connor (1972), particularly, revealed some problems on defining the correct cutpoints and mentioned that the effect of increasing the number of categories, especially of more than four categories, is small and he also mentioned that the choosing optimal categories or classes depend on the

distribution of  $X$ . Begg and Lagakos (1990) and Lagakos (1988a) investigate categorizing for  $k = 2, \dots, 6$  and also compared optimal intervals with equiprobable intervals. They concluded that if distribution of  $X$  is almost symmetric, then equiprobable intervals are allowed to use but if the distribution of  $X$  is quite skewed, then only optimal intervals should be used, instead of equiprobable intervals.

As a preliminary study, Cox (1957) explained a measure of information loss from grouping for choosing cutpoints for different size of categories. He suggested that efficiency of test may be used as a criterion for cutpoint selection and proposed the average information loss as follows

$$L_x = \sum_{i=1}^k p_i E\left[(X - E(X_i))^2 \mid X \text{ in the } i\text{th group}\right] / \sigma^2 \quad (2.34)$$

where  $p_i$  is the probability of an observation appearing in the  $i$ th group. This probability equals  $p_i = \int_{x_{i-1}}^{x_i} f(x) dx$  where  $x_i$  for  $i = 2, \dots, k$  are the class limits and the  $i$ th group is defined by  $x_{i-1} < X < x_i$ .  $E(X_i)$  is the mean of all observations in the  $i$ th group and  $\sigma$  is the standard deviation of  $X$  and each group have the same standard deviation.

In analysis of variance, as known, total sum of squares of all observations of the entire sample is equal to the sum of the sum of squares within groups and the sum of squares between groups as  $SSTO = SSBG + SSWG$ . The following equation expresses in more detail.

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - E(X))^2 = \sum_{i=1}^k n_i (E(X_i) - E(X))^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - E(X_i))^2 \quad (2.35)$$

where  $x_{ij}$  is the  $j$ th observation for  $i$ th variable.



Cramer (1964, p. 237) introduced the ratio

$$\frac{SSBG}{SSTO} = 1 - \frac{SSWG}{SSTO} \quad (2.36)$$

“as an indication of the relative efficiency of alternative methods of grouping a given set of observations”. If this ratio goes to unity, then the efficiency will be less reduced when grouping. Therefore, Cox’s formula in Equation (2.34) with respect to relative efficiency becomes

$$\begin{aligned} L_x &= 1 - \left[ \sum_{i=1}^k p_i (E(X_i) - E(X))^2 \right] / \sigma^2 \\ &= 1 - ARE \end{aligned} \quad (2.37)$$

It seems that ARE has to be maximized so as to reduce the loss of information. Connor (1972) investigated ARE of tests of the association between independent and dependent variables for up to 6 optimal intervals and for explanatory variable having the uniform, normal and exponential ( $\lambda = 1$ ) distributions. Lagakos (1988a) extended the results including the ARE values for equiprobable intervals that means intervals with equal frequencies of occurrence. He noted that ARE for equiprobable intervals can be much smaller, when the explanatory variable follows an exponential distribution. The results regarding test statistics from categorizing with optimal classes are reproduced in Table 2.1 by following Cox’s guidance. Related calculations are given below. Let  $X^*$  denotes misspecified version of  $X$ .

Table 2.1 ARE when categorizing an explanatory variable  $X$  into  $k$  intervals

$k$	Distribution of $X$	Class Probabilities	$ARE(X^*, X)$
2	Uniform	0.500, 0.500	0.75
	Normal	0.500, 0.500	0.65
	Exponential	0.797, 0.203	0.65
3	Uniform	0.333, 0.333, 0.333	0.89
	Normal	0.270, 0.459, 0.270	0.81
	Exponential	0.639, 0.288, 0.073	0.82
4	Uniform	0.250, 0.250, 0.250, 0.250	0.94
	Normal	0.164, 0.336, 0.336, 0.164	0.88
	Exponential	0.530, 0.300, 0.135, 0.035	0.89
5	Uniform	0.200, 0.200, 0.200, 0.200, 0.200	0.96
	Normal	0.109, 0.237, 0.307, 0.237, 0.109	0.92
	Exponential	0.451, 0.291, 0.165, 0.074, 0.019	0.93
6	Uniform	0.167, 0.167, 0.167, 0.167, 0.167, 0.167	0.97
	Normal	0.074, 0.181, 0.245, 0.245, 0.181, 0.074	0.94
	Exponential	0.393, 0.274, 0.176, 0.100, 0.045, 0.012	0.95

The results in Table 2.1 implies that if the explanatory variable follows normal distribution, then categorizing this variable into  $k = 2$  groups costs 35% loss in efficiency of test statistics, similarly, categorizing into  $k = 3$  groups causes 19% efficiency loss and so on. It is clear that the increasing the number of categories gives less loss in efficiency, for all three distribution types, as expected.

Suppose that  $X$  is standard normally distributed with the following probability density function

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

and distribution function

$$G(x) = \int_{-\infty}^x g(u) du.$$

Let the size of categories be 2,  $k = 2$ , if so the cutpoint is taken as the mean of  $X$ , zero, by symmetry conditions and the percentages of individuals for in the two groups being 50.0 and 50.0. For  $k = 3$ , a value that maximizes ARE should be chosen so we have to choose  $y > 0$  and the groups are  $(-\infty, -y), (-y, y), (y, \infty)$  by symmetry, again. The conditional mean of  $X$  given  $x_1 < X < x_2$  is

$$E(X|x_1 < X < x_2) = \frac{\int_{x_1}^{x_2} xg(x)dx}{\int_{x_1}^{x_2} g(x)dx} = \frac{1}{\sqrt{2\pi}} \frac{\left( e^{-\frac{1}{2}x_1^2} - e^{-\frac{1}{2}x_2^2} \right)}{G(x_2) - G(x_1)} \quad (2.38)$$

$$= \frac{g(x_1) - g(x_2)}{G(x_2) - G(x_1)}$$

The probabilities that  $X$  falls into the three different intervals are

$$p_1 = P\{-\infty < x < -y\} = G(-y) - G(-\infty) = G(-y)$$

$$p_2 = P\{-y < x < y\} = G(y) - G(-y)$$

$$p_3 = P\{y < x < \infty\} = G(\infty) - G(y) = 1 - G(y)$$

Therefore, since  $g(-y) = g(y)$  and  $1 - G(y) = G(-y)$  from symmetry, the asymptotic relative efficiency is

$$\begin{aligned}
ARE(X^*, X) &= \sum_{i=1}^k p_i \left( \frac{g(x_1) - g(x_2)}{G(x_2) - G(x_1)} \right)^2 \\
&= G(-y) \left( \frac{g(-\infty) - g(-y)}{G(-y) - G(-\infty)} \right)^2 + (G(y) - G(-y)) \left( \frac{g(-y) - g(y)}{G(y) - G(-y)} \right)^2 \\
&\quad + (1 - G(y)) \left( \frac{g(y) - g(\infty)}{G(\infty) - G(y)} \right)^2
\end{aligned}$$

After simplifying, it is obtained

$$\begin{aligned}
ARE(X^*, X) &= G(-y) \left( \frac{g(-y)}{G(-y)} \right)^2 + (1 - G(y)) \left( \frac{g(y)}{1 - G(y)} \right)^2 \\
&= \frac{2(g(y))^2}{G(-y)}
\end{aligned} \tag{2.39}$$

In order to find the value of  $y$ , the derivative of ARE is found and set to zero. After calculations, it seems that ARE has a maximum value of 0.8098 attained at  $y = 0.612$ . Therefore, the optimal cutpoint for standard normal distribution is 0.612. Besides, for general normal distribution with different parameters and for  $k = 3$ , the three groups should be in the intervals such as  $(-\infty, \mu - 0.612\sigma)$ ,  $(\mu - 0.612\sigma, \mu + 0.612\sigma)$ ,  $(\mu + 0.612\sigma, \infty)$ . The probabilities of observations being in the three groups are as follows.

$$\int_{-\infty}^{-0.612} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.27 \quad \int_{-0.612}^{0.612} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.46 \quad \int_{0.612}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.27$$

For example, if the normal distribution with parameters zero mean and 3 standard deviation is considered, then the intervals will be  $(-\infty, -1.836)$ ,  $(-1.836, 1.836)$ ,  $(1.836, +\infty)$  and the ARE will decrease the value of 0.0899.

The information loss formula, as seen from the Table 2.1, can be applied to other distributions such as exponential distribution. In literature, exponential distribution has been employed but only for the ones having parameter 1. We will extend the results for other values of parameter  $\lambda$  and make a generalization. Suppose that  $X$  is exponentially distributed with parameter  $\lambda$ . The conditional mean of  $X$  given  $x_1 < X < x_2$  is

$$\begin{aligned}
 E(X|x_1 < X < x_2) &= \frac{\int_{x_1}^{x_2} xg(x)dx}{\int_{x_1}^{x_2} g(x)dx} = \frac{\int_{x_1}^{x_2} x\lambda e^{-\lambda x} dx}{\int_{x_1}^{x_2} \lambda e^{-\lambda x} dx} \\
 &= \frac{e^{-\lambda x_1} \left( x_1 + \frac{1}{\lambda} \right) - e^{-\lambda x_2} \left( x_2 + \frac{1}{\lambda} \right)}{e^{-\lambda x_1} - e^{-\lambda x_2}} \\
 &= \frac{e^{-\lambda x_1} x_1 - e^{-\lambda x_2} x_2}{e^{-\lambda x_1} - e^{-\lambda x_2}} + \frac{1}{\lambda}
 \end{aligned} \tag{2.40}$$

When the number of categories  $k = 2$ , the probabilities that  $X$  falls into the first and second intervals are

$$p_1 = P\{0 < x < y\} = G(y) = 1 - e^{-\lambda y}$$

$$p_2 = P\{y < x < \infty\} = 1 - G(y) = e^{-\lambda y}$$

Therefore, ARE equals

$$ARE = \left(1 - e^{-\lambda y}\right) \left(\frac{-e^{-\lambda y} y}{1 - e^{-\lambda y}} + \frac{1}{\lambda}\right)^2 + e^{-\lambda y} \left(\frac{e^{-\lambda y} y}{e^{-\lambda y}} + \frac{1}{\lambda}\right)^2 \tag{2.41}$$

After simple calculations Equation (2.41) follows

$$ARE = \frac{e^{-\lambda y} y^2}{1 - e^{-\lambda y}}. \quad (2.42)$$

Taking the derivative of ARE and setting it to zero as follows

$$\frac{d}{dy} \left\{ \frac{e^{-\lambda y} y^2}{1 - e^{-\lambda y}} \right\} = \frac{-\lambda e^{-\lambda y} y^2 + 2ye^{-\lambda y} - 2ye^{-2\lambda y}}{(1 - e^{-\lambda y})^2} = 0. \quad (2.43)$$

Calculations show that after solving Equation (2.43) the result is

$$\frac{\text{lambertw}(-2e^{-2}) + 2}{\lambda} = \frac{1.5936}{\lambda}. \quad (2.44)$$

Therefore, the cutpoints that is the values maximizing ARE are calculated based on the parameter  $\lambda$ . So the cutpoint choice for exponential distribution with different parameters may be generalized. If we assume that  $\lambda = 1$ , then  $y = 1.5936$  and substituting it in ARE Equation (2.42) the following result given in Table 2.1 is found

$$ARE = \frac{e^{-1.5936}(1.5936)^2}{1 - e^{-1.5936}} = 0.6476.$$

The optimal probabilities due to  $y$  can be calculated as below.

$$\int_0^{1.5936} e^{-x} dx = 0.80 \quad \int_{1.5936}^{\infty} e^{-x} dx = 0.20$$

Hence, consequently it is clear that ARE has a maximum value of 0.6476 attained at  $y = 1.5936$ . The percentages of individuals in the two groups are 80.0 and 20.0.

Furthermore, for example, when  $\lambda = 3$  the new cutpoint value will be  $y = \frac{1.5936}{3} = 0.5312$  with the same class probabilities and ARE will reduce to 0.072.

When  $k = 3$ , the probabilities that  $X$  falls into the first, second and third intervals are as follows.

$$p_1 = G(y_1) = 1 - e^{-\lambda y_1}$$

$$p_2 = G(y_2) - G(y_1) = e^{-\lambda y_1} - e^{-\lambda y_2}$$

$$p_3 = 1 - G(y_2) = e^{-\lambda y_2}$$

Using these probabilities, ARE is calculated as follows.

$$\begin{aligned} ARE = & \left(1 - e^{-\lambda y_1}\right) \left(\frac{-e^{-\lambda y_1} y_1}{1 - e^{-\lambda y_1}} + \frac{1}{\lambda}\right)^2 + \left(e^{-\lambda y_1} - e^{-\lambda y_2}\right) \left(\frac{e^{-\lambda y_1} y_1 - e^{-\lambda y_2} y_2}{e^{-\lambda y_1} - e^{-\lambda y_2}} + \frac{1}{\lambda}\right)^2 \\ & + e^{-\lambda y_2} \left(\frac{e^{-\lambda y_2} y_2}{e^{-\lambda y_2}} + \frac{1}{\lambda}\right)^2 \end{aligned} \quad (2.45)$$

We may base our calculations on the exponential distribution with  $\lambda = 1$  and so Equation (2.45) follows,

$$\begin{aligned} ARE = & \left(1 - e^{-y_1}\right) \left(\frac{-e^{-y_1} y_1}{1 - e^{-y_1}}\right)^2 + \left(e^{-y_1} - e^{-y_2}\right) \left(\frac{e^{-y_1} y_1 - e^{-y_2} y_2}{e^{-y_1} - e^{-y_2}}\right)^2 \\ & + e^{-y_2} \left(\frac{e^{-y_2} y_2}{e^{-y_2}}\right)^2 \end{aligned} \quad (2.46)$$

Setting derivative of the above ARE is equal to zero, the cut points are calculated as  $y_1 = 1.0176$ ,  $y_2 = 2.6112$  and  $ARE = 0.8203$ . The percentages of individuals in the three groups are 63.9, 28.8 and 7.3 are found from following expressions,

$$\int_0^{1.0176} e^{-x} dx = 0.639 \quad \int_{1.0176}^{2.6112} e^{-x} dx = 0.288 \quad \int_{2.6112}^{\infty} e^{-x} dx = 0.073.$$

As a generalization of unit exponential distribution, for different  $\lambda$  parameters, the cutpoints may be calculated by  $\frac{1.0176}{\lambda}$  and  $\frac{2.6112}{\lambda}$ . So, if we take  $\lambda = 3$ , the new values of cutpoints will be  $y_1 = \frac{1.0176}{3} = 0.3392$  and  $y_2 = \frac{2.6112}{3} = 0.8704$  with the same class probabilities and ARE will reduce to 0.227.

As an extension, the information loss formula can be applied to Weibull distribution with the following probability density function

$$f(x) = \frac{\beta}{\delta} \left( \frac{x-\gamma}{\delta} \right)^{\beta-1} \exp \left\{ - \left( \frac{x-\gamma}{\delta} \right)^{\beta} \right\} \quad (2.47)$$

where  $-\infty < \gamma < \infty$  is location parameter,  $\delta > 0$  is scale parameter and  $\beta > 0$  is the shape parameter. The corresponding distribution function is

$$F(x) = 1 - \exp \left\{ - \left( \frac{x-\gamma}{\delta} \right)^{\beta} \right\} \quad (2.48)$$

Let  $\gamma = 0$ ,  $\delta = 1$  and  $\beta = 2$ . The conditional mean of  $X$  given  $x_1 < X < x_2$  is



$$\begin{aligned}
E(X|x_1 < X < x_2) &= \frac{\int_{x_1}^{x_2} xg(x)dx}{\int_{x_1}^{x_2} g(x)dx} = \frac{\int_{x_1}^{x_2} 2x^2 e^{-x^2} dx}{\int_{x_1}^{x_2} 2xe^{-x^2} dx} \\
&= \frac{x_1 e^{-x_1^2} - x_2 e^{-x_2^2} + \frac{\sqrt{\pi}}{2} (\operatorname{erf}(x_2) - \operatorname{erf}(x_1))}{e^{-x_1^2} - e^{-x_2^2}}
\end{aligned} \tag{2.49}$$

For  $k = 2$ , the probabilities that  $X$  falls into the first and second intervals are

$$p_1 = P\{0 < x < y\} = G(y) = 1 - e^{-y^2}$$

$$p_2 = P\{y < x < \infty\} = 1 - G(y) = e^{-y^2}$$

Calculations show that the maximum ARE is 0.8847 and the value of  $y$  that gives this maximized value is 1.26, with the percentages of individuals in the two groups are 0.80 and 0.20.

For  $k = 3$ , the probabilities that  $X$  falls into the first and second intervals are

$$p_1 = G(y_1) = 1 - e^{-y_1^2}$$

$$p_2 = G(y_2) - G(y_1) = e^{-y_1^2} - e^{-y_2^2}$$

$$p_3 = 1 - G(y_2) = e^{-y_2^2}$$

Thereby, the cutpoints are  $y_1 = 0.3278$ ,  $y_2 = 1.1692$  with  $\text{ARE} = 0.9897$ . The percentages of individuals in the three groups are 0.10, 0.64 and 0.26, respectively.

### 2.3.2 Omission of a Covariate

In observational studies, to attain an important explanatory variable is sometimes difficult or expensive and sometimes impossible to measure and therefore omitting it from the model may be preferred, easily. However, the omission of some variables that affect the dependent variable may cause an omitted variables bias. This bias depends on the correlation between the independent variables which are omitted and included in linear models. If the omitted variable is completely uncorrelated with the variables in the model, the coefficients may not be biased, but this is almost not possible in practice. The omitted variable bias has been widely studied for linear regression models as in Erees and Demirel (2012), Leightner and Inoue (2007). Besides, Gail, Wieand and Piantadosi (1984) have studied on the bias caused by omitted covariate in generalized linear models. Lagakos (1988a), Begg and Lagakos (1993) studied on omitted variables effect on the efficiency of test statistics used for significance of logistic regression parameters. Neuhaus (1998) and Neuhaus and Jewell (1993) have investigated the effects of omitting covariates on the parameter estimation in generalized linear models.

In linear regression models, if the omitted covariates are independent of the included variables, then model misspecification due to omission does not cause an omitted variable bias. However, in generalized linear models, so that in logistic regression models, omitting covariates associated with the dependent variable, even if they are independent of the included variables, causes seriously downward estimates of regression coefficients (Neuhaus, 1998).

Suppose that the true model has the following form with mean function given in Equation (2.51)

$$\text{logit } \Pr(Y = 1|X, Z) = \beta_0 + \beta_1 X + \beta_2 Z \quad (2.50)$$

with

$$\mu = E(Y|X, Z) = \frac{\exp\{\beta_0 + \beta_1 X + \beta_2 Z\}}{1 + \exp\{\beta_0 + \beta_1 X + \beta_2 Z\}} \quad (2.51)$$

Suppose that the covariate  $Z$  is omitted and then the misspecified model will be as

$$\text{logit } \Pr(Y = 1|X) = \beta_0^* + \beta_1^* X, \quad (2.52)$$

with

$$\mu^* = E(Y|X) = \frac{\exp\{\beta_0^* + \beta_1^* X\}}{1 + \exp\{\beta_0^* + \beta_1^* X\}}. \quad (2.53)$$

When the effect of an omitted covariate is investigated in terms of efficiency, the asymptotic Pitman efficiency of tests may be used as a criterion. If  $Z$  is independent of  $X$ , the variances of the estimates of true and misspecified regression coefficients and test statistics based on these coefficients may be compared to calculate ARE (Gail et al., 1984, Neuhaus and Jewell, 1993). In addition, if  $X$  and  $Z$  are correlated, the variances of the estimates of true and misspecified regression coefficients may be compared when  $Z$  is a nonconfounding covariate which means that  $Z$  does not confound the association of  $X$  and  $Y$  and so  $\beta_1^* = \beta_1$ . Therefore, the ARE with respect to omitted variable is

$$\begin{aligned} ARE[(X^*, 0), (X, Z)] &= 1 - \frac{\sigma_\mu^2}{E(\mu)[1 - E(\mu)]} \\ &= 1 - \frac{E([\mu - E(\mu)]^2)}{E(\mu)[1 - E(\mu)]} \end{aligned} \quad (2.54)$$

where  $X^*$  denotes a misspecified version of  $X$  and 0 indicates that  $Z$  has been omitted.

If the omitted covariate is independent from the included variables, then omitted covariate causes loss in efficiency and this loss increases depending on the association of the omitted covariate and the response. Nonetheless, omitting nonconfounding covariates provides a gain in efficiency and this gain in efficiency rises as the effect of the omitted covariate to dependent variable rises (Neuhaus, 1998).

### ***2.3.3 Mismodelling a Continuous Explanatory Variable***

The aim of regression analysis is to determine the best specified model to explain the data of interest. When the explanatory variable is continuous, it must be decided for its functional form, since using some transformations relating continuous variable may cause some problems. In medical studies, in particular, because of the complexity of relationships between variables, simple regression models may not represent the true relationships between these variables, exactly. It may not even be possible to detect whether a model is incorrectly specified, since for the sample sizes available in many applications, diagnostics of model fit have good power to detect only a limited number of the potential ways that a model may fail to be correctly specified (Keele, 2008). Therefore, it is important to know how much loss will occur and what the consequences will be and whether the results are reliable, in such cases. In agricultural research, adding the logarithm of the amount of forage consumption to the model as an explanatory variable whereas it is supposed to be added without taking logarithm is an example of mismodelling.

In linear regression, using wrong functional form and testing whether the model is linear are well established. Particularly, Ramsey RESET (Regression Specification Error Test) test is used as a general test so as to detect misspecification of functional form (Erees & Demirel, 2012). Its logic depends on the inclusion of the different powers of the fitted values to the original model. If the coefficients of associated with the added variables are significant, there is misspecification because of wrong functional form or omitting important variable (Brooks, 2008; Verbeek, 2004).

In literature, several transformations relating to explanatory variables are studied so as to find out how much efficiency may be lost when using the incorrect functional form of a continuous explanatory variable, in logistic regression. Lagakos (1988b) made a study about this kind of misspecification and found the results in Table 2.2. The uniform, two unimodal and symmetric beta distributions, a right-skewed beta distribution, and a U-shaped beta distribution were selected, for the distribution of  $X$ . Two convex functions,  $X^2$  and  $\exp(X)$ , two concave functions,  $\sqrt{X}$  and  $\ln(X)$  were examined as the functional forms of  $X$ .

**Table 2.2** ARE when mismodelling a continuous explanatory variable  $X$

Distribution of $X$	Shape of distribution	ARE when $X$ equals			
		$X^2$	$\exp(X)$	$\sqrt{X}$	$\ln(X)$
Uniform (0, 1)	Flat	0.94	0.98	0.96	0.75
Beta (2, 2)	Unimodal, symmetric	0.95	0.99	0.98	0.86
Beta (5, 5)	Unimodal, symmetric	0.97	1.00	1.00	0.93
Beta (1, 3)	Skewed right	0.90	0.99	0.95	0.69
Beta(0.5, 0.5)	U-shaped	0.95	0.97	0.93	0.68

It is understood from the table that if the distribution of the explanatory variable is beta with parameters (2, 2), for example, and if we use the form of the squared root, by mistake, the asymptotic efficiency of true version relative to misspecified will be 0.98. In other words, the loss in efficiency will be 2%.

The results in Table 2.2 are reproduced keeping that the asymptotic relative efficiency is equal to the square of the correlation in mind as given in Equation (2.33). Let misspecified version of  $X$  denote  $X^*$ . In this regards, since  $ARE(X^*, X) = \rho_{X^*X}^2$ , as given in Equation (2.33), some basic calculations are made using correlation formula. For example, suppose that the distribution of  $X^*$  is uniform (0, 1) and the relationship between  $X^*$  and  $X$  is  $X = (X^*)^2$ , if so  $ARE(X^*, X)$  is calculated as

$$\rho(X^*, (X^*)^2) = \frac{\text{cov}(X^*, (X^*)^2)}{\sqrt{\sigma^2(X^*)} \sqrt{\sigma^2((X^*)^2)}} \quad (2.55)$$

Since  $\text{cov}(X^*, (X^*)^2) = E(X^*(X^*)^2) - E(X^*)E((X^*)^2)$ , we should calculate these expected values

$$E(X^*(X^*)^2) = E((X^*)^3) = \int_0^1 (x^*)^3 dx^* = \frac{(x^*)^4}{4} \Big|_0^1 = \frac{1}{4}$$

and

$$E(X^*) = \int_0^1 x^* dx^* = \frac{(x^*)^2}{2} \Big|_0^1 = \frac{1}{2} \quad \text{and} \quad E((X^*)^2) = \int_0^1 (x^*)^2 dx^* = \frac{(x^*)^3}{3} \Big|_0^1 = \frac{1}{3}$$

Then the covariance

$$\text{cov}(X^*, (X^*)^2) = \frac{1}{4} - \frac{1}{2} \frac{1}{3} = \frac{1}{12}$$

and the variances are

$$\sigma^2(X^*) = E[(X^*)^2] - [E(X^*)]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

and

$$\begin{aligned} \sigma^2((X^*)^2) &= E\left[\left((X^*)^2\right)^2\right] - \left[E((X^*)^2)\right]^2 = E[(X^*)^4] - \left(\frac{1}{3}\right)^2 = \int_0^1 (x^*)^4 dx^* - \frac{1}{9} \\ &= \frac{4}{45} \end{aligned}$$

Therefore, substituting values in the correlation formula

$$\rho(X^*, (X^*)^2) = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}} \sqrt{\frac{4}{45}}} = \frac{\sqrt{15}}{4}$$

Then  $ARE(X^*, X)$  is equal to  $\rho^2(X^*, (X^*)^2) = \frac{15}{16} \cong 0,94$ . Hence, the result of  $ARE(X^*, X) = 0.94$  is found using the fact given by Equation (2.33) and is consistent with the result in Table 2.2.

## CHAPTER THREE

### COEFFICIENT OF DETERMINATION

In logistic regression analysis, in contrast to linear regression, there is no standard definition of coefficient of determination ( $R^2$ ). Different  $R^2$  statistics in accordance with different perspectives have been proposed in recent years. After giving the multiple correlation coefficient for general linear models, previous works and recommendations of some authors, the most frequently used and suggested  $R^2$  statistics for logistic models will be discussed in more detail, in section 3.2. Each  $R^2$  statistic will be given in separate subsections and three well-known  $R^2$  that are included in most logistic regression outputs in some underlying statistical software packages such as SPSS, SAS and STATA will be explained in last three subsections.

#### 3.1 $R^2$ Statistics

Coefficient of determination, also called explained variance, is well established in classical linear regression models (Draper & Smith, 1998, Helland, 1987, Ohtani & Tanizaki, 2004).  $R^2$  in linear regression is the square of multiple correlation which represents the total correlation between all the independent variables and the dependent variable. Since the square of a correlation is the same as a proportion of variance,  $R^2$  is said to be the proportion of variance about the mean explained by the regression model and also called explained variance by the model (Miles & Shevlin, 2001). It measures how well the regression model performs as a predictor of dependent variable. It is well known that if the only available information are the values of the dependent variable and there is no knowledge about independent variables, then we use the mean of  $Y$  as the best predicted value of  $Y$  for all cases and minimize the sum of squared errors based on prediction using the mean of  $Y$  which equals to  $\sum(Y_i - \bar{Y})^2$  which is called Total Sum of Squares (SSTO). On the other hand, if there is information about independent variables to predict  $Y$ , then we use the value of predicted  $Y$  from the regression equation  $\hat{Y}$  and minimize the sum of



squared errors based on this prediction. This quantity is called the Error Sum of Squares (SSE) and equals to  $\sum (Y_i - \hat{Y}_i)^2$ . Since this sum is expected to be smaller than the total sum of squares, the least squares method uses the minimization of SSE in order to find regression parameters. So as a proportion of variance, the idea of  $R^2$  is attributable to these sums of squares as

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \quad (3.1)$$

where SSR stands for Regression Sum of Squares. Since SSE is always less than or at most equal to SSTO and greater than 0, this ratio will be less than 1 or at least equal to 0 which means  $0 \leq R^2 \leq 1$ .

Although in linear regression there is only one  $R^2$  statistic, in logistic regression so many different measures of explained variation are reported on by different authors, throughout the years, because there is not one way to measure the strength of association between the dependent variable and all of the independent variables. As Efron (1978) indicated that linear regression models have only one error variation criterion for continuous dependent variables (SSE) while logistic regression models have several error variation criterion such as squared error, entropy etc. for binary dependent variable. Moreover, Menard (2000; p. 17) gives another reason for deriving so many  $R^2$  statistics and failing to agree on one statistic as “the existence of numerous mathematical equivalents to  $R^2$  in OLS, which are not necessarily mathematically (same formula) or conceptually (same meaning in the context of the model) equivalent to  $R^2$  in logistic regression”. Hence, analysts may face the difficulty of choosing the convenient  $R^2$  statistic among all of them and wonder which statistic is the most efficient and under what conditions the statistic should be included in the analysis.

Kvalseth (1985, p. 281) compared various  $R^2$  statistics and described eight criteria that they should possess to make a recommendation about the most convenient, the “good”  $R^2$  statistic. Some of them may be summarized as:

- (1)  $R^2$  must be useful as a goodness of fit measure and interpretable reasonably.
- (2)  $R^2$  should be generalized and applicable to any type of model, independent variable whatever their statistical properties are.
- (3) The potential limits of  $R^2$  should be defined in cases of perfect fit and complete lack of fit which are preferable to be 0 and 1.
- (4) “ $R^2$  should be such that its values for different models fitted to the same data set are directly comparable.”
- (5) “Relative values of  $R^2$  ought to be generally compatible with those derived from other acceptable measures of fit (e.g., standard error of prediction and root mean squared residual).”

There are different  $R^2$  statistics proposed in the literature satisfying some of these properties. However, Menard (2000) extended criterion 4 and 5. He suggested that “the coefficient of determination should be comparable across not only different predictors, but different dependent variables and different subsets of the dataset” as an extension of criterion 4. Moreover, he indicated that  $R^2$  is comparable with alternative coefficient of determination statistics but “some of the usual ‘other’ acceptable measures of fit (standard error of prediction, root mean squared residual)” given in criterion 5 “may not be appropriate” for logistic regression models (Menard, 2000, p. 18). So they cannot be comparable generally. Furthermore, Menard (2002) studied on the properties of six different statistics including the ordinary least squared  $R^2$  ( $R_{OLS}^2$ ), the likelihood ratio  $R^2$  ( $R_L^2$ ), geometric mean squared

improvement based  $R^2$  ( $R_M^2$ ), adjusted geometric mean squared improvement based  $R^2$  ( $R_N^2$ ), the contingency coefficient  $R^2$  ( $R_C^2$ ) and McKelvey and Zavoina  $R^2$  ( $R_{MZ}^2$ ). He recommended, after investigations and comparisons,  $R_L^2$  as the most convenient statistic for logistic regression.

Mittlböck and Schemper (1996) submitted four properties that R-squares should have and while reviewing twelve useful and suggested measures, checked whether they meet the requirements for being a “good” statistic. Two of these properties are the same with (1) and (3) proposed by Kvalseth. The other two include (a) consistency with the character of logistic regression that is there should be no underlying linear regression, no linearly transformation and (b) having consistent values with multiple correlation coefficient in OLS, numerically. After checking with simulation if the measures provide the conditions, they recommended the squared Pearson correlation coefficient ( $r^2$ ),  $R_{OLS}^2$  and Gini’s concentration measure ( $R_G^2$ ).

Hagle and Mitchell (1992) studied on four  $R^2$  type which are  $R_L^2$ ,  $R_C^2$ ,  $R_{MZ}^2$  and Achen pseudo  $R^2$ . They proposed an adjustment to  $R_C^2$  ( $R_{CA}^2$ ) and after using simulation methods and calculating error statistics of measures made comparisons with  $R^2$  in OLS and found that,  $R_{CA}^2$  and  $R_{MZ}^2$  are more preferable for being good approximations for  $R^2$  in linear regression. Veall and Zimmerman (1996) evaluated the performances of nine  $R^2$  statistics consisting five above and as a result of Monte Carlo simulations they recommended  $R_{MZ}^2$  as the most consistent measure for the OLS  $R^2$  when the binary dependent variable represents underlying continuous variable.

Hosmer & Lemeshow (2000) examined the performances of  $r^2$  and  $R_L^2$ , with an example and found that even if with good logistic models, they may have lower values than the generally experienced values of R-square in OLS with good linear relationship. However, although this is not unexceptional for logistic regression,

these low values do not sound very well when interpreting an analysis. So the authors suggested that it would be more helpful to include the statistics to the analysis during the model building process instead of after fitting the model.

### 3.2 Alternative $R^2$ Statistics

It's possible to examine the most frequently used statistics into two categories of likelihood (or entropy) based and variance based (Hu, Palta & Shao, 2006; Mittlböck & Schemper, 1996). In the following five subsections, we will discuss variance based measures of explained variation and in the other five subsections we will present measures based upon likelihood function.

#### 3.2.1 The Ordinary Least Squared $R^2$

In general linear models, we defined the error and total sum of squares in Section 3.1 and now we will use them for logistic regression as

$$SSE = \sum_{i=1}^n (Y_i - \hat{\pi})^2 \quad (3.2)$$

and

$$SSTO = \sum_{i=1}^n (Y_i - \bar{\pi})^2 \quad (3.3)$$

The ordinary least squared (OLS)  $R^2$  which is also called sum of squares  $R^2$  statistic is

$$R_{OLS}^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\pi})^2}{\sum_{i=1}^n (Y_i - \bar{\pi})^2} \quad (3.4)$$

$R_{OLS}^2$  corresponds to the coefficient of determination in linear regression when applied to a linear regression model (Hu, Palta and Shao, 2006). This statistic that has been studied by also Agresti (1990), Maddala (1983) varies from zero to one and so provides an advantage of being direct comparable between logistic models and OLS based models. But Menard (2000) mentioned that this is possible just technically and for illumination.

Kvalseth asserted that  $R_{OLS}^2$  satisfies nearly all the criteria except the end point requirement corresponding to complete lack of fit in some cases. He recommended  $R_{OLS}^2$  statistic after comparing eight different types of statistics for linear and nonlinear models, since he noted that if  $R_{OLS}^2$  is appropriate for linear models and nonlinear models which are inherently linear, it may be used and advisable for models which are inherently nonlinear. Menard agrees with him about  $R_{OLS}^2$  may be used as an analog of  $R^2$  in linear regression since they are equivalent mathematically. But he also noted that they are not “conceptually” equivalent to each other because of the difference of the quantity that is being minimized. In linear models this quantity that  $R^2$  based on is a squared error measure of variation and in logistic regression an entropy measure of variation, and so based on likelihood measure. Therefore  $R_{OLS}^2$  does not satisfy completely the first criteria of Kvalseth which says being interpretable reasonably. Because,  $R_{OLS}^2$  in logistic regression concerns about the numerical values of binary dependent variable instead of the probability of it. So, it is true that it has an intuitively interpretation but not in that is really concerned about.

While Mittlböck & Schemper (1996) preferred  $R_{OLS}^2$  since it provides all desirable properties that they determined, Cox & Wermuth (1992) showed and emphasized that with binary dependent variables  $R_{OLS}^2$  takes low values, in general 0.10, even if the model fits the data very well and so this causes lack of clear interpretation.

### 3.2.2 Squared Pearson Correlation Coefficient

Squared Pearson correlation coefficient is the squared correlation between observed dependent variable  $Y$  and its sample fitted value  $\hat{Y}$  in linear regression as known. Similarly, in logistic regression, squared Pearson correlation coefficient ( $r^2$ ) is the square of the sample correlation between the observed binary dependent variable  $Y_i$  and corresponding prediction  $\hat{\pi}_i$  and is defined as

$$r^2 = [\text{corr}(Y, \hat{\pi})]^2 \tag{3.5}$$

$$= \frac{\left[ \sum_{i=1}^n (Y_i - \bar{Y})(\hat{\pi}_i - \bar{\pi}) \right]^2}{\left[ \sum_{i=1}^n (Y_i - \bar{Y})^2 \right] \left[ \sum_{i=1}^n (\hat{\pi}_i - \bar{\pi})^2 \right]}$$

Since  $\sum_i \hat{\pi}_i = \sum_i Y_i = n\bar{\pi}$ , and  $\bar{Y} = \bar{\pi}$ , this formula may be written as

$$r^2 = \frac{\left( \sum_{i=1}^n Y_i \hat{\pi}_i - n\bar{\pi}^2 \right)^2}{\left( n\bar{\pi}(1 - \bar{\pi}) \sum (\hat{\pi}_i - \bar{\pi})^2 \right)} \tag{3.6}$$

Kvalseth (1985) explained that since this statistic is linear correlation based, it would not be effective as a goodness of fit measure for nonlinear models. Moreover, even though  $Y_i$  and  $\hat{\pi}_i$  are highly correlated which is an expected result, if their values have great deviations, then misleading results may be produced caused by using  $r^2$ . On the other hand, Mittlböck & Schemper (1996) revealed that  $r^2$  meets all requirements that they gave and recommended this statistic as one of the “good”  $R^2$  statistics.

### 3.2.3 Gini's Concentration Measure

The concentration measure of Gini  $C(\pi) = 1 - \sum \pi^2$  was used as a measure of dispersion of a nominal random variable  $Y$  by Haberman (1982). When logistic regression is discussed, Gini's concentration is used as measure of the expected variance of the binary dependent variable under the models with and without independent variables as  $\hat{\pi}(1 - \hat{\pi})$  and  $\bar{Y}(1 - \bar{Y})$ , respectively. Then  $R_G^2$  takes the form of

$$\begin{aligned}
 R_G^2 &= \frac{\sum_{i=1}^n \bar{Y}(1 - \bar{Y}) - \sum_{i=1}^n \hat{\pi}(1 - \hat{\pi})}{\sum_{i=1}^n \bar{Y}(1 - \bar{Y})} \\
 &= 1 - \frac{\sum_{i=1}^n \hat{\pi}(1 - \hat{\pi})}{\sum_{i=1}^n \bar{Y}(1 - \bar{Y})}
 \end{aligned} \tag{3.7}$$

It's clear from the equation that this statistic unusually involves no observed values, but only predicted values, in fact, their expected variances and assumes that the model is correct. Mittlböck & Schemper (1996) as a result of their comparison with simulation study showed that  $R_G^2$  meets all four requirements that they gave such as giving numerically consistent values with  $R^2$  in general linear models. Hu, Palta & Shao (2006) studied on the recommended statistics by Mittlböck & Schemper that are  $R_{OLS}^2$ ,  $r^2$  and  $R_G^2$  and found the following close relationship between them

$$\left( R_G^2 + R_{OLS}^2 \right)^2 = 4r^2 R_G^2.$$

### 3.2.4 The Wald $R^2$

Magee (1990) used the relation between the  $F$ , Wald and likelihood ratio statistics, all of which are used for testing the same hypothesis which states that at least one of the  $k - 1$   $\beta_i$ 's is equal to zero, in order to improve or select an  $R^2$  statistic from the existing statistics. Since  $F$  is equal to

$$F = \frac{(SSTO - SSE)/(k - 1)}{SSE/(n - k)} \quad (3.8)$$

and is a monotonic increasing function of  $R^2$  in OLS in Equation (3.1) as given in Johnston (1984, p:187),  $F$  can be written in terms of  $R^2$  as

$$F = \frac{R^2/(k - 1)}{(1 - R^2)/(n - k)} \quad (3.9)$$

Besides, Vandaele (1981) shows that Wald statistic which is equal to  $n(SSTO - SSE)/SSE$  can be written in terms of  $F$  as

$$W = (k - 1) \left[ 1 + \frac{k}{n - k} \right] F \quad (3.10)$$

If we combine Equation (3.9) and Equation (3.10) then Wald  $R^2$  statistic is obtained as below.

$$R_w^2 = \frac{W}{W + n} \quad (3.11)$$

The addition of sample size  $n$  to the denominator causes  $R_w^2$  cannot equal to one even if the model fit is perfect to data.



### 3.2.5 McKelvey and Zavoina's Measure

McKelvey & Zavoina (1975) proposed an  $R^2$  measure that may be employed for both probit and logit models when the dependent variable is an underlying continuous variable. The statistic is the proportion of the explained and unexplained variance of predicted values for the latent dependent variable.

$$R_{MZ}^2 = \frac{\text{var}(\hat{Y}_i)}{\text{var}(\hat{Y}_i) + \frac{\pi^2}{3}} \quad (3.12)$$

where  $\pi^2/3$  is the standard deviation for logistic distribution. Veall and Zimmerman (1996) recommended  $R_{MZ}^2$  since it is the most consistent measure and has a good approximation for the OLS  $R^2$  in linear regression when the binary dependent variable represents underlying continuous variable. However, because of not being a likelihood based measure, it is not applicable to polytomous models as well as dichotomous models (Menard, 2000).

### 3.2.6 The Contingency Coefficient $R^2$

Aldrich & Nelson (1985) proposed a measure named pseudo  $R^2$  or contingency coefficient  $R^2$ . Contingency coefficient (C) is a chi square based measure of association for two nominal variables in contingency tables and equal to

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad (3.13)$$

where  $\chi^2$  is the Pearson chi squared statistic (Blaikie, 2003). Aldrich & Nelson (1985) employed the well-known likelihood ratio statistic  $G_M$  which equals to

$-2\{\ln(L_0)-\ln(L_M)\}$  as  $\chi^2$  statistic with  $k$  degrees of freedom using this above equation and proposed  $R_C^2$  as

$$R_C^2 = \frac{-2\{\ln(L_0)-\ln(L_M)\}}{n-2\{\ln(L_0)-\ln(L_M)\}} = \frac{G_M}{G_M+n}, \quad (3.14)$$

where  $L_0$  is the likelihood function statistic for the model containing only the intercept and  $L_M$  is the likelihood function for the model containing all of the independent variables and the subscript  $C$  represents the dependency of the form of contingency coefficient. The values of contingency coefficient range between 0 and 1, when there is no relationship between two variables, it takes the value of 0 but it can't take the value of 1 even when two variables are perfectly related to each other. Similarly,  $R_C^2$  cannot have a maximum value of 1, since the sample size  $n$  is added to  $G_M$  in the denominator. Additionally, as a positive result of being based on  $G_M$  which is a likelihood derived statistic,  $R_C^2$  can be calculated not only dichotomous variables but also polytomous variables, too (Menard, 2000).

### 3.2.7 Adjusted Contingency Coefficient $R^2$

Hagle and Mitchell (1992) considered the case of model fit which is perfect with included independent variables, which is the case of  $\ln L_M = 0$ . Then, they rewrote the Equation (3.14) as

$$R_C^2 = \frac{-2\ln(L_0)}{n-2\ln(L_0)}. \quad (3.15)$$

Hence, they revealed that  $R_C^2$  takes the maximum value of 0.5809 for  $\bar{Y} = 0.5$  from the following equation.

$$R_{\max}^2 = \frac{-2(\bar{Y} \ln \bar{Y} + (1 - \bar{Y}) \ln(1 - \bar{Y}))}{1 + 2(\bar{Y} \ln \bar{Y} + (1 - \bar{Y}) \ln(1 - \bar{Y}))}$$

Hence, to make the statistic more reasonable by providing it to reach the value of one, they suggested a correction which remarks that  $R_C^2$  should be multiplied by  $1/0.5809$ , in other words divided by its maximum, to eliminate the effect of sample of size. So the adjusted  $R_C^2$  is obtained as follows.

$$R_{CA}^2 = \frac{R_C^2}{\max(R_C^2)} \quad (3.16)$$

As result, it is clear that  $R_{CA}^2$  varies from zero to one as preferred by most of authors.

### 3.2.8 The Likelihood Ratio $R^2$

In linear regression, the total sum of squares (SSTO) measures the uncertainty in predicting the dependent variable and does not take into account the independent variables. The error sum of squares (SSE) is the measure of the variation in the dependent variable and the independent variables are taken into account. The difference between them indicates the reduction in variation due to the independent variables. In logistic regression, as known, inferences are based on the log likelihood function.  $-2 \log$  likelihood ( $L_0$ ) represents the likelihood for the model without any independent variables and corresponds to the total sum of squares in OLS. The model  $-2 \log$  likelihood ( $L_M$ ) represents the likelihood for the model with independent variables and corresponds to the error sum of squares in linear regression. Therefore, the difference in the log likelihood models shows the improvement due to the independent variables, in logistic regression and already equals to the likelihood ratio statistic  $G_M$  (Pampel, 2000). Therefore, as  $R^2$  in linear regression is used for defining the proportional reduction in error sum of squares; an

analogue statistic can be used for defining the proportional reduction in these log likelihoods in logistic regression. Consequently, McFadden (1974) defined the likelihood based  $R^2$  statistic as

$$R_L^2 = \frac{-2[\ln(L_0) - \ln(L_M)]}{-2[\ln(L_0)]} = \frac{[-2\ln(L_0)] - [-2\ln(L_M)]}{-2[\ln(L_0)]}. \quad (3.17)$$

Since the zero deviance is  $D_0 = -2\ln(L_0)$  and model deviance is  $D_M = -2\ln(L_M)$  and the well-known likelihood ratio statistic is  $G_M = -2\ln\left(\frac{L_0}{L_M}\right) = D_0 - D_M$ , Equation (3.17) is equal to

$$R_L^2 = \frac{D_0 - D_M}{D_0} = \frac{G_M}{D_0}. \quad (3.18)$$

The values of  $R_L^2$  vary between zero and one. When all the coefficients are equal to zero,  $R_L^2$  takes the value of zero. If we fit the saturated model, then the value of log-likelihood from saturated model equals to zero ( $\ln(L_S) = 0$ ) and  $R_L^2$  takes the maximum value of 1.

Menard (2002) has suggested that  $R_L^2$  is the most proper measure for logistic regression. He explained the reasons of this suggestion under four considerations. (1)  $R_L^2$  is conceptually close to  $R_{OLS}^2$  since it depends only on the quantity that the model tries to minimize (-2 log likelihood) not also sample size. (2) It is sensitive to base rate. (3) It ranges between zero and one. (4) It is applicable not only to dichotomous dependent variables but also to polytomous nominal or ordinal dependent variables because of being dependent on likelihood function.

### 3.2.9 Geometric Mean Squared Improvement

In linear regression model when errors are normally distributed with zero mean and constant variance, standard multiple  $R^2$  is  $R^2 = 1 - \left(\frac{L_0}{L_M}\right)^{2/n}$  (DeMaris, 2002; Magee, 1990). It is clear that this expression can be extended for logistic regression models, because likelihood functions are already calculated for maximum likelihood estimation method in logistic regression. Therefore, the statistic for logistic regression is

$$R_M^2 = 1 - \exp\left\{-\frac{2}{n}[\ln(L_M) - \ln(L_0)]\right\} = 1 - \left(\frac{L_0}{L_M}\right)^{2/n}. \quad (3.19)$$

Cox and Snell (1989) defined the term  $\left(\frac{L_0}{L_M}\right)^{2/n}$  as the geometric mean improvement per observation produced by full model versus to intercept only model and so the subscript  $M$  indicates the use of geometric mean squared improvement.

As an undesirable property, this statistic cannot achieve an upper bound value of one even the model predicts the dependent variable perfectly. The maximum attainable value of  $R_M^2$  will be 0.75, if and only if  $Y = 1$  with 50% and  $Y = 0$  with 50%, in other words, each observation is predicted with a maximum probability of 1.00, for logistic model (Nagelkerke, 1991).

$R_M^2$  can be written in terms of likelihood ratio statistic  $G_M$  as following

$$\begin{aligned} R_M^2 &= 1 - \exp\left\{-\frac{1}{n}[-2\ln(L_0) - (-2\ln(L_M))]\right\} \\ &= 1 - \exp\left\{-\frac{G_M}{n}\right\}. \end{aligned} \quad (3.20)$$

Since  $R_M^2$  depends upon  $G_M$ , it may be applied to polytomous models, however because of being utilized as geometric mean squared improvement, it has the property of intuitively meaningful interpretation, only partially. Furthermore, we may interrelate with  $R_L^2$  using the likelihood ratio statistic that is common in both so that a theoretical expression may be provided between them. From Equation (3.19) and Equation (3.20),

$$R_M^2 = 1 - \exp\left\{-\frac{R_L^2 D_0}{n}\right\}. \quad (3.21)$$

Let  $\phi$  denote the ratio  $\frac{D_0}{n}$ , we have

$$R_M^2 = 1 - \exp\{-\phi R_L^2\}. \quad (3.22)$$

Since  $1 - \exp\{-\phi R_L^2\} \leq \phi R_L^2$ , it follows from Equation (3.22)

$$R_M^2 \approx \phi R_L^2 \quad (3.23)$$

Therefore, it is understood that when  $\phi \geq 1$ ,  $R_L^2 \leq R_M^2$  and otherwise  $R_L^2 > R_M^2$ .

### ***3.2.10 Adjusted Geometric Mean Squared Improvement***

As well-known, likelihood function  $L_M$  is the product of probabilities so it takes the values less than 1. Therefore, from Equation (3.19) the maximum of  $R_M^2$  may reach to the following

$$\max(R_M^2) = 1 - \exp\left\{\frac{2}{n} \ln(L_0)\right\} = 1 - (L_0)^{2/n}. \quad (3.24)$$

Nagelkerke (1991), to overcome the obstacle to achieve the property of having maximum 1, proposed to adjust  $R_M^2$  as

$$R_N^2 = \frac{1 - \exp\left\{-\frac{2}{n}[\ln(L_M) - \ln(L_0)]\right\}}{\max(R_M^2)} = \frac{1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}}{1 - (L_0)^{\frac{2}{n}}}. \quad (3.25)$$

This adjusted measure permits a value of one dividing  $R_M^2$  by its maximum possible value. We know that the maximum value of  $R_M^2$  is 0.75 so that  $R_N^2 = \frac{R_M^2}{0.75}$  and this means that  $R_N^2 > R_M^2$ , as expected.

Ryan (1997) examined  $R_M^2$ ,  $R_N^2$  and correct classification rate (CCR) which can be treated as a measure of the fit of a model, for assessing the quality of a logistic regression model and he suggested to use  $R_N^2$  as a supplementary statistic since  $R_N^2$  has meaningfully different values for different models. Hu, Shao & Palta (2006, p. 849), using entropy of the marginal distribution of  $Y$ , proved a theorem which assumes that “ $X_i$   $i = 1, \dots, n$ , are independent and identically distributed random p-vectors with finite second moment” and provides the asymptotic limits of  $R_M^2$  and  $R_N^2$ .

$$R_M^2 \xrightarrow{p} 1 - \exp\{2(H_2 - H_1)\}$$

$$R_N^2 \xrightarrow{p} \frac{1 - \exp\{2(H_2 - H_1)\}}{1 - \exp\{-2H_1\}}$$

where  $H_1 = -\sum_{j=1}^m E(\pi_{ij}) \log E(\pi_{ij})$  and is the entropy measure the marginal variation of  $Y_i$ ,  $H_2 = -\sum_{j=1}^m E(\pi_{ij} \log \pi_{ij})$  and is the conditional entropy measure the variation of  $Y_i$  given  $X_i$ , hence their difference gives the entropy explained by  $X_i$ ,  $\xrightarrow{p}$  denotes convergence in probability. Furthermore, they noted that larger the absolute value of regression coefficients larger these limits and also although the model has a strong relationship with dependent variable, the limits may have low values.

$R_N^2$  is also interrelated with  $R_L^2$  and this relation can be expressed theoretically using the likelihood ratio statistic. From Equation (3.22) and Equation (3.25), we may formulize  $R_N^2$  based on  $R_L^2$  and hence on  $\phi$

$$R_N^2 = \frac{1 - \exp\{-\phi R_L^2\}}{1 - \exp\{-\phi\}} \leq \frac{\phi}{1 - \exp\{-\phi\}} R_L^2. \quad (3.26)$$

If we denote the ratio  $\frac{\phi}{1 - \exp\{-\phi\}}$  as  $\phi^*$ , then  $\phi^*$  will be greater than or equal to 1, since  $1 - \exp\{-\phi\} \leq \phi$ . Therefore, the values of  $R_N^2$  are less than the values of

$$R_N^2 \approx \phi^* R_L^2. \quad (3.27)$$

Since  $\phi^* > 1$ , it is concluded that  $R_L^2 < R_N^2$ .



## CHAPTER FOUR

### NUMERICAL RESULTS

This chapter presents simulation studies and an application on real life addressing the effects of misspecification on the asymptotic relative efficiency of coefficients of determination in logistic regression analysis. We emphasized three well-known coefficients of determination that are given in last subsections in Section 3.2, the likelihood ratio  $R^2$ , geometric mean squared improvement and adjusted geometric mean squared improvement. We will compare them based on efficiency to see the influences of asymptotic results of misspecification explained in Chapter 2 on R-square statistics. Basing our analysis upon more reliable fundamentals is important for convenient inferences so we should attach a certain importance to this issue. Furthermore, we will perform an application to better understand the results of simulations and to show the asymptotic results in practice.

The following section will present the simulation results. In Section 4.2 we will give the numerical results of application and the comparisons with findings in simulation studies.

#### 4.1 Simulation Studies

Simulation studies are designed to show the influence of misspecification and distribution of continuous independent variable on various types of  $R^2$  for logistic regression. We examined the effects of misspecification in terms of asymptotic relative efficiency using bootstrap method.

Three R-square analogs, namely  $R_L^2$ ,  $R_M^2$  and  $R_N^2$  were considered for evaluating their performances. The reason they are chosen is that they are more popular and the most consulted measures for analysis. These statistics are already included in the logistic regression outputs in the popular software packages as SPSS, SAS and STATA. Besides, the simulation studies were repeated for different sizes of sample

to see whether the means and efficiencies of R-squares depend upon sample size. All calculations have been carried out using R programming language version of 3.0.1.

To examine the effects of misspecification on the asymptotic relative efficiency of  $R^2$  statistics, we studied on the population with size  $N = 100,000$ . From this population we have randomly selected 10,000 samples with sizes of  $n = 50$  and  $n = 100$ . Bootstrap sampling procedure with  $B = 500$  bootstrap replications has been used to estimate the variance of the corresponding statistics. The binary response variable  $Y$  was generated from the Bernoulli distribution with a success probability given with Equation (4.1), with a continuous explanatory variable ( $X$ ) and a discrete covariate ( $Z$ ). To be consistent with real life, we set the approximate correlations between  $X$  and  $Y$ ,  $X$  and  $Z$ ,  $Z$  and  $Y$  as 0.55, 0.15 and 0.40, respectively.

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 Z)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Z)} \quad (4.1)$$

$X$  has been assumed to have  $N(\mu, \sigma^2)$ , with  $\mu = 0$ ,  $\sigma^2 = 1$  and 9 to observe the effects of extreme rising of variance and exponential ( $\lambda$ ) with parameters  $\lambda = 1$  and 3 where  $\lambda = 1/\text{mean}$ . Covariate  $Z$  has been generated with a structure depending  $X$ . If  $X$  exceeds a value with a probability of almost 0.3, then  $Z$  has Binomial distribution with  $p = 0.6$  otherwise  $p = 0.4$ . For example, for an explanatory variable  $X$  having  $N(0,1)$  distribution, if  $X_i > 0.5$  for  $i = 1, \dots, n$  (since  $P(X_i > 0.5) = 0.3$ ),  $Z$  has Binomial distribution with  $p = 0.6$ . This specific value is 1.5 when  $X$  has  $N(0,9)$  distribution.

Misspecifications chosen for this study include: *i*) categorizing the continuous explanatory variable  $X$  into  $k = 2$  and  $k = 3$  categories, *ii*) using wrong functional form of  $X$ , *iii*) omitting the discrete covariate  $Z$  from the model. Wrong functional form involves taking the third power of  $X$  ( $X^3$ ), taking the natural logarithm of  $X$  ( $\ln(X)$ ) and taking the square root of  $X$  ( $\sqrt{X}$ ). For example, if we use  $X^3$  instead of  $X$ , the logistic function defined in Equation 4.1 will then take the form of

$$E(Y) = \frac{\exp(\beta_0^* + \beta_1^* X^3 + \beta_2^* Z)}{1 + \exp(\beta_0^* + \beta_1^* X^3 + \beta_2^* Z)}. \quad (4.2)$$

The models conducted with the population values without any misspecification are called original model and with the sample values without any misspecification are denoted by  $X$ . For categorizing  $X$ , the cutpoints given with Table 4.1 are selected as explained in Section 2.1.1.

For discretizing  $X$ , the corresponding cutpoints given with Table 4.1 are selected for each distribution. The techniques of choosing cutpoints were mentioned in detailed in Section 2.3.1. These values are based on the extension of the Cox's (1957) calculations and provide optimal intervals. The use of cutpoints for optimal intervals is attributed to the fact that optimal intervals are more preferable than equiprobable intervals. Connor (1972) and Lagakos (1988b) showed that the efficiencies of test statistics are much smaller for equiprobable intervals than optimal intervals especially when skewed distributions are used. Besides, the numbers of categories as two and three are found enough to see the effects of categorization since the efficiency losses are expected to be considerably low, after three categories.

Table 4.1 Number of categories and location of cutpoints

Distribution of $X$	Cutpoints of $X$ for	
	$k = 2$	$k = 3$
N(0,1)	Mean	-0.612 and 0.612
N(0,9)	Mean	-1.836 and 1.836
Exp(1)	1.594	1.018 and 2.611
Exp(3)	0.531	0.339 and 0.870

We report the results of  $R^2$  for original model built with population values without any misspecification and the medians of  $R^2$  statistics for other models in Table 4.2 and in Table 4.3 for  $n = 50$  and larger sample size  $n = 100$ , respectively.

For example, in the model type symbolized by  $X^3$  and in the second column, in Table 4.2, the value 0.408 is the median of  $R_L^2$  obtained from 10,000 simulations when  $X^3$  is used instead of  $X$  having standard normal distribution and when the sample of size is 50. Under normal distribution, changing the functional form of  $X$  with  $\ln(X)$  and  $\sqrt{X}$  has the most effect on  $R^2$  values for both sample sizes. For exponential distributions, on the other hand, omitting the covariate  $Z$  from the model significantly decreases the value of  $R^2$  statistics. These mentioned effects are more significant for  $N(0,9)$  and  $\text{Exp}(3)$ .

To have a general idea about asymptotic distributions of  $R^2$  statistics due to corresponding distributions of  $X$ , the density plots of  $R^2$  statistics have been drawn. From the figures presented in Appendix, we can observe that the R-squares have asymptotically normal distribution. There are some cases that show some minor departures from normality such as  $R_L^2$  and  $R_N^2$  in models with  $\ln(X)$  and  $\sqrt{X}$  when  $X$  having normal distribution (0,1) and  $R_L^2$  in models with  $X$  and  $X^3$  when  $X$  having normal distribution (0,3). However, as graphical display and Kolmogorov-Smirnov test results confirm, distributions of all  $R^2$  statistics are asymptotically normal.

Table 4.2 The real values of  $R^2$  for original model and the medians of  $R^2$  for other models for  $n = 50$ 

Model Type	N(0,1)			N(0,9)			Exp(1)			Exp(3)		
	$R_L^2$	$R_M^2$	$R_N^2$	$R_L^2$	$R_M^2$	$R_N^2$	$R_L^2$	$R_M^2$	$R_N^2$	$R_L^2$	$R_M^2$	$R_N^2$
Original	0.417	0.370	0.552	0.723	0.614	0.839	0.364	0.394	0.527	0.206	0.209	0.308
$X$	0.455	0.389	0.589	0.772	0.635	0.873	0.397	0.418	0.562	0.246	0.240	0.354
$X^3$	0.408	0.362	0.543	0.745	0.621	0.853	0.374	0.400	0.538	0.240	0.236	0.353
$\ln(X)$	0.230	0.222	0.335	<b>0.164</b>	<b>0.193</b>	<b>0.266</b>	0.352	0.378	0.512	0.233	0.229	0.339
$\sqrt{X}$	0.261	0.254	0.376	<b>0.205</b>	<b>0.234</b>	<b>0.320</b>	0.385	0.407	0.547	0.243	0.236	0.354
$k = 2$	0.338	0.306	0.465	0.558	0.521	0.713	0.289	0.324	0.437	0.220	0.218	0.328
$k = 3$	0.409	0.360	0.545	0.624	0.558	0.764	0.356	0.378	0.513	0.259	0.250	0.372
<i>Z omitted</i>	0.330	0.298	0.453	0.711	0.604	0.830	0.268	0.302	0.411	<b>0.088</b>	<b>0.094</b>	<b>0.138</b>

Table 4.3 The real values of  $R^2$  for original model and the medians of  $R^2$  for other models for  $n = 100$ 

Model Type	N(0,1)			N(0,9)			Exp(1)			Exp(3)		
	$R_L^2$	$R_M^2$	$R_N^2$	$R_L^2$	$R_M^2$	$R_N^2$	$R_L^2$	$R_M^2$	$R_N^2$	$R_L^2$	$R_M^2$	$R_N^2$
Original	0.408	0.364	0.543	0.729	0.617	0.843	0.365	0.394	0.528	0.213	0.215	0.317
$X$	0.429	0.370	0.562	0.752	0.629	0.859	0.381	0.404	0.544	0.236	0.232	0.346
$X^3$	0.374	0.334	0.507	0.713	0.607	0.831	0.351	0.379	0.510	0.221	0.217	0.322
$\ln(X)$	0.293	0.180	0.366	<b>0.516</b>	<b>0.263</b>	<b>0.591</b>	0.337	0.365	0.493	0.216	0.214	0.322
$\sqrt{X}$	0.319	0.191	0.394	<b>0.551</b>	<b>0.276</b>	<b>0.621</b>	0.371	0.397	0.533	0.232	0.228	0.338
$k = 2$	0.319	0.294	0.443	0.548	0.510	0.700	0.295	0.331	0.445	0.207	0.207	0.311
$k = 3$	0.384	0.340	0.516	0.609	0.548	0.752	0.331	0.364	0.489	0.233	0.230	0.344
<i>Z omitted</i>	0.319	0.292	0.442	0.692	0.598	0.817	0.260	0.297	0.398	<b>0.085</b>	<b>0.092</b>	<b>0.136</b>

As we addressed in Section 3.2.9 and in Section 3.2.10, from the interrelation between the R-squares, we concluded that  $R_N^2$  has the greatest values and the magnitudes of  $R_M^2$  and  $R_L^2$  changes with the ratio of the sample of size and the null deviance denoted by  $\phi$ . When  $\phi \geq 1$ ,  $R_L^2 \leq R_M^2$  and otherwise  $R_L^2 > R_M^2$ . Simulation results regarding these comparisons in terms of medians given with Table 4.2 and Table 4.3 confirm the theoretical findings about  $R_N^2$ , since overall  $R_N^2$  are bigger than  $R_M^2$  and  $R_L^2$  at all simulation models. In addition, since  $R_L^2 > R_M^2$  in general, it is said that  $\phi < 1$ .

It was explained in detail in Section 2.2 that ARE of two statistics is the ratio of their variances as given in Equation (2.17). Therefore, AREs of R-squares have been obtained using their sampling variances which are found by bootstrap method. The misspecified version of  $R^2$  will be denoted with  $R^{2*}$ . The asymptotic relative efficiency of  $R_j^{2*}$  with respect to  $R_j^2$  is

$$ARE \left( R_j^{2*}, R_j^2 \right) = \frac{\sigma^2 \left( R_j^2 \right)}{\sigma^2 \left( R_j^{2*} \right)} .$$

Corresponding values are presented with Tables 4.4 - 4.6.

All three cases for wrong functional form of continuous  $X$  are presented in Table 4.4. Increasing the standard deviation of normally distributed  $X$  has generally great effects on efficiency under misspecification. For all types of  $R^2$  statistics, using  $X^3$  instead of  $X$  causes significant efficiency loss. For instance, for  $R_N^2$  the efficiency loss is 100%. For exponentially distributed  $X$  variables, misspecification at the functional form of  $X$  does not significantly affect the variance of the statistics. The three  $R^2$  measures were found to result in almost identical efficiency values across all types of mismodelling. There is at most 5% loss in efficiency.

Table 4.4 ARE's of each  $R^2$  statistics under both correct and misspecified models when  $X$  has been mismodelled

	$R_L^{2*}, R_L^2$						$R_M^{2*}, R_M^2$						$R_N^{2*}, R_N^2$					
	$n = 50$			$n = 100$			$n = 50$			$n = 100$			$n = 50$			$n = 100$		
	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$
N(0,1)	0.81	0.96	1.03	0.83	0.65	0.68	0.80	0.84	0.92	0.79	1.18	1.24	0.76	0.70	0.78	0.75	0.57	0.61
N(0,9)	<b>0.07</b>	0.98	0.97	0.63	0.24	0.29	0.00	0.32	0.97	0.58	0.31	0.37	0.00	0.26	0.28	0.55	0.12	0.15
Exp(1)	0.91	0.88	0.98	0.89	0.88	0.98	0.86	0.79	0.96	0.83	0.78	0.96	0.86	0.79	0.96	0.82	0.78	0.96
Exp(3)	0.95	1.02	1.01	0.99	1.03	1.01	0.96	1.00	1.01	0.98	1.00	1.00	0.95	0.99	1.00	0.97	1.00	1.00

Table 4.5 ARE's of each  $R^2$  statistics under both correct and misspecified models when  $X$  has been categorized

	$R_L^{2*}, R_L^2$				$R_M^{2*}, R_M^2$				$R_N^{2*}, R_N^2$			
	$n = 50$		$n = 100$		$n = 50$		$n = 100$		$n = 50$		$n = 100$	
	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$
N(0,1)	1.25	1.15	1.15	1.08	0.97	1.02	0.93	0.98	0.97	1.04	0.91	0.97
N(0,9)	0.95	1.20	<b>0.73</b>	0.96	<b>0.51</b>	0.96	<b>0.40</b>	0.79	<b>0.53</b>	0.86	<b>0.41</b>	0.69
Exp(1)	1.14	1.08	1.09	1.09	0.88	0.98	0.88	0.98	0.88	0.98	0.87	0.97
Exp(3)	1.07	0.97	1.07	1.00	1.02	1.00	1.03	1.00	1.02	1.00	1.03	1.00

Table 4.5 shows the ARE values of each  $R^2$  statistics for categorizing continuous  $X$ . The ARE values are in general quite close to unity when  $X$  follows exponential distribution meaning there is a small loss of efficiency caused by categorization. For example, if  $X$  has an  $\text{Exp}(1)$ , categorizing into  $k = 2$  groups results with 12% loss in efficiencies of both  $R_M^2$  and  $R_N^2$ . On the other hand when  $X$  follows a  $N(0,9)$ ,  $R_M^2$  and  $R_N^2$  are significantly affected by categorization into  $k = 2$  groups, since efficiency losses reduce to almost 60%. In general, the efficiencies of R-squares, when misspecification means categorization, may alter with the number of groups and so much number of groups minimizes the increase in the variance of the R-squares.

Table 4.6 ARE's of each  $R^2$  statistics under both correct and misspecified models when omitting  $Z$

	$R_L^{2*}, R_L^2$		$R_M^{2*}, R_M^2$		$R_N^{2*}, R_N^2$	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$
N(0,1)	1.20	1.15	0.98	0.97	0.94	0.93
N(0,9)	1.12	0.99	0.98	0.88	0.97	0.86
Exp(1)	1.27	1.26	0.93	0.92	0.92	0.92
Exp(3)	1.97	2.00	1.54	1.57	1.48	1.51

Table 4.6 evaluates the AREs of each  $R^2$  statistics under correct model versus  $Z$  omitted model. All statistics seem robust against the omission of  $Z$  under larger sample size and normal distribution. Neuhaus (1998) showed that omitting non-confounding covariate which is correlated with  $X$  provides a gain in efficiency of the estimated effects and this gain in efficiency increases with the effect of the omitted covariate on dependent variable. When we compare the coefficients of correct and misspecified regression models, it is seen that  $\beta_1^* = \beta_1 = 2$ . Thereby  $Z$  is said to be a non-confounding covariate, i.e., it does not confound the association of  $X$  and  $Y$ . In this respect omitting a non-confounding covariate may provide a gain in efficiency especially for exponential distribution with mean value of  $1/3$ .



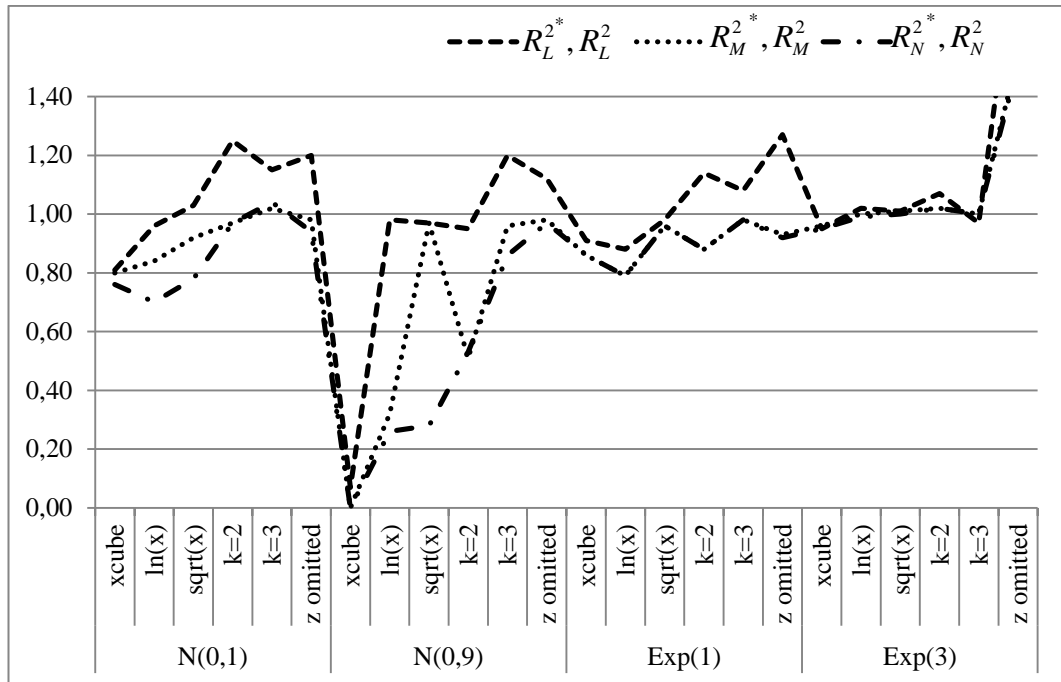


Figure 4.1 ARE's of each  $R^2$  statistics under both correct and misspecified models for  $n = 50$

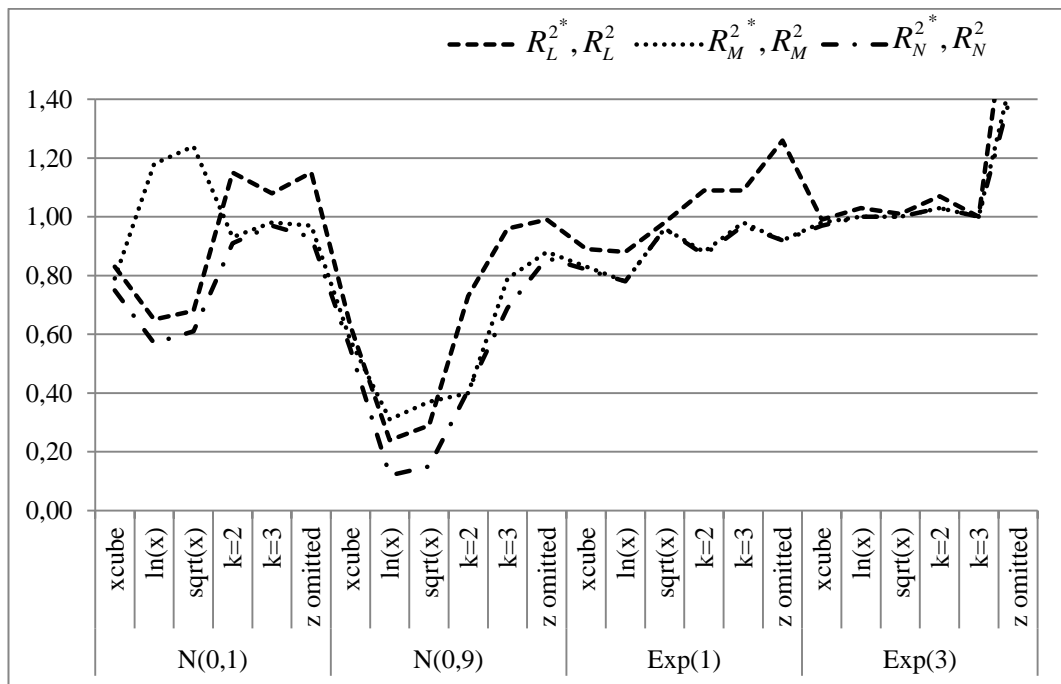


Figure 4.2 ARE's of each  $R^2$  statistics under both correct and misspecified models for  $n = 100$

As a summary, in Figure 4.1, a line graph is given to show the behaviors of ARE's of R-squares under misspecification for  $n = 50$  with visual perception. The

reduction in ARE of  $R_N^2$ , fluctuation in  $R_M^2$  and alteration in  $R_L^2$  when  $X$  having a normal distribution with larger variance is presented more obviously. This plot illustrates better that the efficiencies are affected seriously if  $X^3$  is used when  $X$  have normal distribution (0,3) and if  $Z$  is omitted when  $X$  have exponential distribution (3). Furthermore, once exponential distribution is considered, it makes no difference to use  $R_M^2$ ,  $R_N^2$  or  $R_L^2$ , since they give the same reaction to misspecification, regardless of parameter. On the other hand, the line graph given with Figure 4.2 shows that when  $n = 100$ , all three R-squares take very small values for  $\ln(X)$  and  $\sqrt{X}$ . It is clear that exponential distribution leads to the same efficiency loss regardless of not only parameter but also sample size.

For the second part of the simulation, the purpose is to examine the AREs of  $R^2$  statistics with respect to each other and to evaluate their performances. Table 4.7 shows the ARE values of  $R^2$  statistics when the model does not have any misspecification. It is understood that, if we use correctly specified model, considering  $R_M^2$  instead of  $R_L^2$  will be more reasonable, because the entire ARE values associated with them are notably small. Moreover, since not only  $ARE(R_L^2, R_M^2) < 1$  but also  $ARE(R_N^2, R_M^2) < 1$ ,  $R_M^2$  seems the most efficient  $R^2$  statistic among three. Differences in the distributions and the sample sizes do not change this result.

Table 4.7 ARE's of three  $R^2$  statistics under correct model

	$R_L^2, R_M^2$		$R_N^2, R_L^2$		$R_N^2, R_M^2$	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$
N(0,1)	0.49	0.55	1.02	0.93	0.50	0.52
N(0,9)	0.27	0.28	2.27	2.19	0.61	0.62
Exp(1)	0.60	0.65	0.92	0.85	0.55	0.56
Exp(3)	0.70	0.78	0.68	0.62	0.47	0.48

The consequences of using wrong functional form of  $X$  are presented in Table 4.8. It seems that as in the case of correctly specified model,  $R_M^2$  gives generally the most efficient results except the case of using  $X^3$  when  $X$  has  $N(0,9)$  distribution. From Table 4.2 we know that the variance of  $R_M^2$  and  $R_N^2$  for  $X^3$  model and  $n = 50$  is substantially larger since the ARE values are 0.00. This makes  $R_L^2$  is more efficient than  $R_M^2$  and  $R_L^2$ . In this case, ARE gets unacceptably large value which is presented with “-“ symbol in the table. Besides, we see that  $R_M^2$  is superior than  $R_N^2$  in all conditions. In addition, it is remarkable to note that the use of exponential distribution does not cause any unexpected result regardless of parameter value and sample size.

Categorization of  $X$  or omission of a covariate  $Z$  do not change the fact that  $R_M^2$  is the most efficient statistic, as it is obvious in Tables 4.9 and Table 4.10. The other common result for both tables is the behavior of  $R_N^2$  compared to  $R_L^2$  under  $N(0,9)$  where  $R_N^2$  is more efficient than  $R_L^2$ . Using  $R_N^2$  prevents at least 22% loss in efficiency without dependency of sample size when categorizing  $X$ . It is clear that the number of groups does not behave as a criterion in determining the superiority of coefficients of determination in logistic regression. Under omission case, the ARE of  $R_N^2$  versus  $R_L^2$  is 1.96 and 1.90, for  $n = 50$  and  $n = 100$ , respectively. This means that using  $R_L^2$  instead of  $R_N^2$  causes great loss with a 50 percent in efficiency.  $R_M^2$  tended to produce more efficient values. The variation in the dependent variable is explained by only the explanatory variable  $X$  included in the model using  $R_M^2$ , more efficiently.

Table 4.8 ARE's of three  $R^2$  statistics with each other when  $X$  has been mismodelled

	$R_L^2, R_M^2$						$R_N^2, R_L^2$						$R_N^2, R_M^2$					
	$n = 50$			$n = 100$			$n = 50$			$n = 100$			$n = 50$			$n = 100$		
	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$	$X^3$	$\ln(X)$	$\sqrt{X}$
N(0,1)	0.50	0.56	0.55	0.59	0.30	0.30	0.96	0.75	0.77	0.84	0.82	0.84	0.48	0.42	0.42	0.49	0.25	0.25
N(0,9)	-	0.80	0.74	0.30	0.22	0.23	0.00	0.62	0.66	1.91	1.07	1.11	0.55	0.49	0.49	0.58	0.23	0.25
Exp(1)	0.63	0.67	0.62	0.70	0.74	0.67	0.87	0.82	0.90	0.79	0.75	0.83	0.55	0.55	0.55	0.56	0.55	0.56
Exp(3)	0.70	0.71	0.70	0.79	0.80	0.78	0.68	0.66	0.67	0.61	0.60	0.61	0.47	0.47	0.47	0.48	0.48	0.48

99

Table 4.9 ARE's of three  $R^2$  statistics with each other when categorizing  $X$

	$R_L^2, R_M^2$				$R_N^2, R_L^2$				$R_N^2, R_M^2$			
	$n = 50$		$n = 100$		$n = 50$		$n = 100$		$n = 50$		$n = 100$	
	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$	$k = 2$	$k = 3$
N(0,1)	0.63	0.55	0.68	0.61	0.79	0.92	0.73	0.84	0.50	0.51	0.50	0.51
N(0,9)	0.50	0.33	0.51	0.34	1.26	1.63	1.22	1.58	0.63	0.54	0.63	0.54
Exp(1)	0.78	0.66	0.81	0.72	0.70	0.83	0.69	0.77	0.55	0.55	0.55	0.56
Exp(3)	0.73	0.68	0.81	0.77	0.65	0.70	0.59	0.62	0.47	0.47	0.48	0.48

Table 4.10 ARE's of three  $R^2$  statistics with each other when omitting  $Z$

	$R_L^2, R_M^2$		$R_N^2, R_L^2$		$R_N^2, R_M^2$	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$
N(0,1)	0.60	0.66	0.80	0.75	0.48	0.49
N(0,9)	0.30	0.32	1.96	1.90	0.60	0.61
Exp(1)	0.83	0.89	0.67	0.63	0.55	0.55
Exp(3)	0.89	0.99	0.51	0.47	0.46	0.46

The line graphs given with Figure 4.3 and Figure 4.4 reveal the changes in efficiencies due to sample size. For  $n = 50$ , when wrong functional form of  $X$  having N(0,9), especially  $X^3$ , is used, the reduction in efficiency of  $R_M^2$  and  $R_N^2$  is illustrated much better. In this case,  $R_L^2$  seems more preferable in terms of efficiency. Under these circumstances, as Menard (2002) noted,  $R_L^2$  is preferred over other  $R^2$  statistics. On the other hand, when  $n = 100$ , both  $R_M^2$  and  $R_N^2$  become more efficient than  $R_L^2$ . Therefore it is said that the variance of  $X$  having normal distribution lead to fundamental changes of efficiencies of both  $R_M^2$  and  $R_N^2$ , depending upon the sample size. For the other distributions, especially exponential distribution, three lines are quite close to each other in both figures. It means that misspecification does not affect the relationships of  $R^2$  statistics.

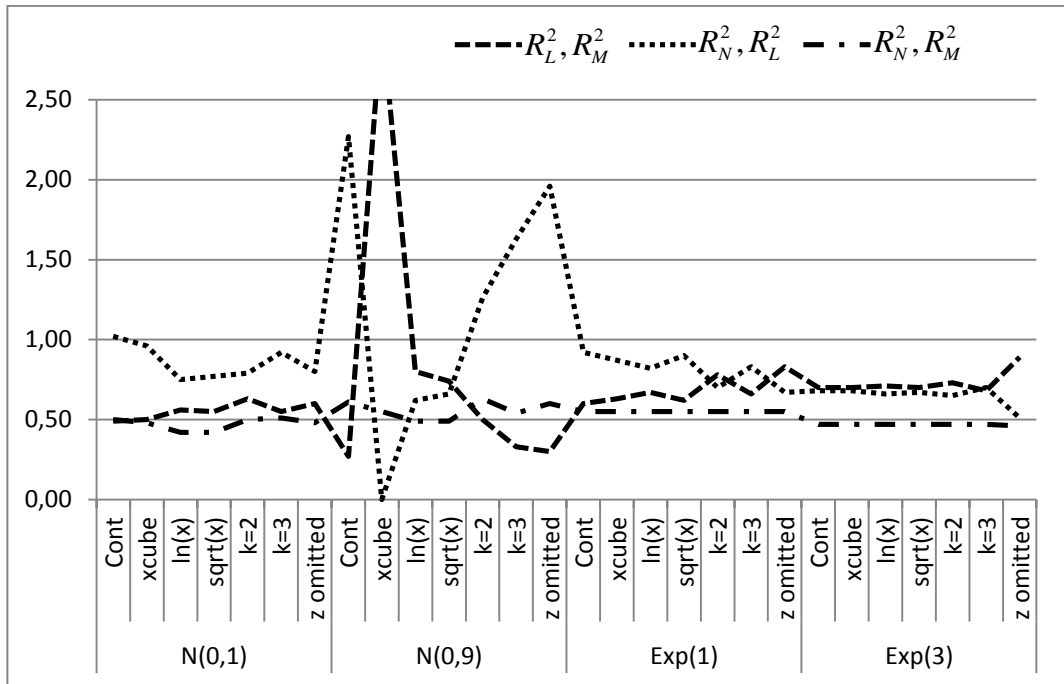


Figure 4.3 ARE's of three  $R^2$  statistics with each other for  $n = 50$

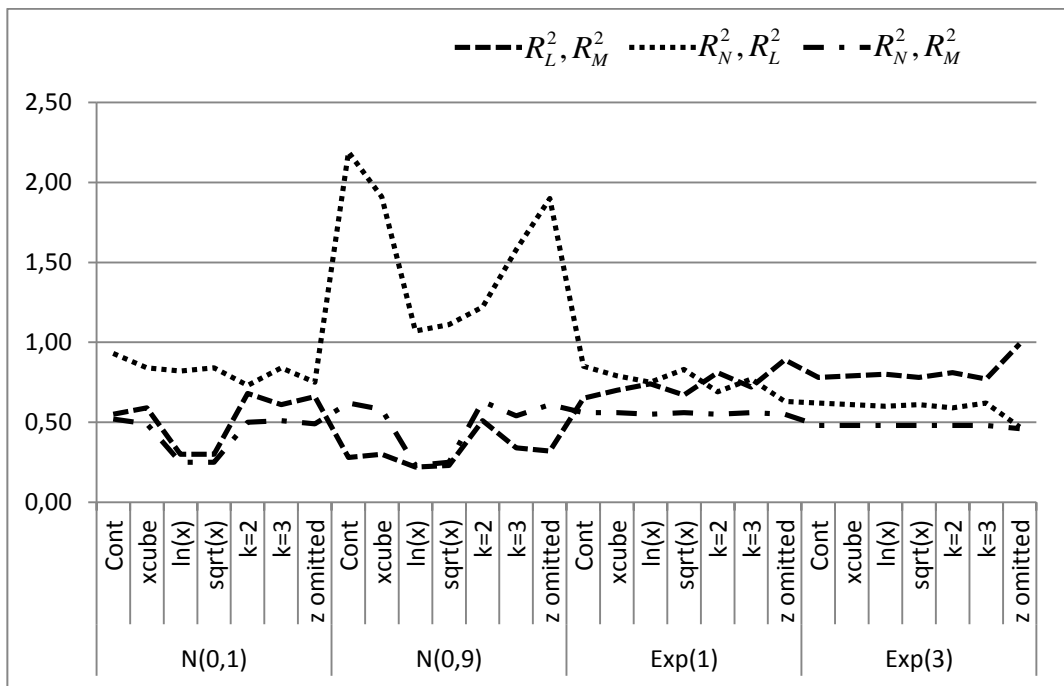


Figure 4.4 ARE's of three  $R^2$  statistics with each other for  $n = 100$

## **4.2 Application on Real Land Consolidation Data**

### ***4.2.1 Introduction to Land Consolidation***

Land consolidation may be described as “the planned readjustment of the pattern of the ownership of land parcels with the aim of forming larger and more rational land holdings” (Pašakarnis & Maliene, 2010, p: 546). Food and Agriculture Organization of the United Nations (FAO) whose one of the aims is to improve agricultural productivity is an intergovernmental organization and plays a very important role in supporting land consolidation activities. FAO implies that land consolidation is not only the simple reallocation of parcels to avoid adverse impacts of fragmentation but also it is associated with social and economic reforms (Food and Agriculture Organization of the United Nations, 2003).

The content and objective of land consolidation varies substantially from countries to counties. The contents may be based on the agricultural and forestry structure, other industries, sheltering and living environment, land use needs, attitude of landowners, society etc. The objectives may be considered as increasing productivity directing all parcels to roads and water access in parallel with lowering the costs of production and concerning ecological, social and cultural structures of the country (Vitikainen, 2004).

Land consolidation is the most favorable land management approach for avoiding land fragmentation improving agricultural productivity and has been applied in many countries around the world such as China, Cyprus, Armenia, Hungary, Lithuania and Serbia. In addition, land consolidation became part of the European Union’s new Rural Development Policy (Demetriou, Stillwell & See, 2012).

Land consolidation has a procedure of research, parcel planning, and evaluation for parcel design and finally consolidation application. During this process, receiving expert consultations, the intervention of governments, briefing of peasants and reaching a consensus are substantially necessary. The new parcel designs, especially,

should be determined taking peasants demands and technical expects into consideration as well as evaluation techniques. In the sense that, interviewing with peasants is an important part of consolidation since designs take form depending on the information taken this way.

In land consolidation projects, there are several reasons that make land owners avoid from joining the consolidation process which include the reasons such as:

- the parcels they already have are more productive than the others or
- the new parcels which are planned to take part are in the places they don't approve.

In despite of this, some land owners sustain the project since

- they will have individual parcel instead of shared one
- their land will connect to a path and a water source
- in different places fragmented small farms become an obstacle to development and sustainable farming,

So the land owners acknowledge the advantages of consolidation such as being economic and improving agricultural productivity.

#### ***4.2.2 Land Consolidation in Turkey***

The land consolidation performances in Turkey have begun in 1961 in Karkin village in Konya Province. Basing upon positive results new legislation about it has been enacted in 1966. With the enactment of this statute consolidation efforts have been applied to a wider range of area. It has continued in an area of total 2,943,000 hectares until 2012. Nowadays, consolidation activities have been carried on with success. From 2013 to 2017, the area of implemented is expected to be 5 million hectares (Boyraz & Üstündağ, 2008; Gün, 2003; Türker & Şaban, 2013).



### ***4.2.3 Application Case***

Binary logistic regression models where dependent variable has only two different values have been applied on many agricultural data sets. With the help of logistic regression analysis, Raut et al. (2011) studied on the influences of some variables such as irrigation facility and landholding size on the adoption of agricultural intensification, Minetos & Polyzos (2009) investigated the agricultural land use due to land use changes. Mueller et. al. (2005) used logistic regression method to improve models which map probability that soil erosion have been arisen before. Battaglin & Goolsby (1996) searched on the relations between different drainage basin variables and some chosen agricultural chemical concentrations in the rivers. There are numerous studies combining the logistic regression and agriculture (Msoffe et.al., 2011; Schroeder et.al., 2001; Zhang & Zhao, 2013). Apart from these studies, Cimpoieş (2007) and Lerman & Cimpoieş (2006) worked on determining the effects of consolidation on living standards of rural peasant families with logistic regression.

In this thesis, the willingness of peasants for consolidation will also be investigated using this method. The opinions of the peasants are very important to begin the consolidation works. So to be able to predict willingness provides a significant gain in vision. A researcher can comprehend what attributes affect the behaviors of peasants statistically and so that he/she can plan the preparatory works. This is the reason, for studying this particular case in this thesis.

This study which has an aim of predicting the willingness of peasants is a part of a larger study on land consolidation project which is carried out in Susuzköy Village, Ankara Province, in Turkey. During parcel planning stage, every peasant who is the owner of the parcel or leased the parcel was included to interview. Data were collected on 228 land owners, 176 of whom were willing to consolidation and 52 of whom were unwilling to the project.

Table 4.11 Descriptions for land consolidation data

Variable	Description	Codes/Values	Name
<i>Y</i>	Choice	0 : No, 1 : Yes	CHO
<i>X</i>	Area	Hectare	AR
<i>Z</i>	The ratio of number of shared parcel and number of individual parcel	0 : number of shared parcel > number of individual parcel , 1 : number of shared parcel ≤ number of individual parcel	SP

Land consolidation depends on many parameters as number and shape of parcels belonging to the peasant, the distance from water sources or road, productivity of land. In this sense that, experts apply an agricultural rating system before consolidation is undertaken. We, however, by getting experts' advice, included two of parameters in our study to be consistent with simulation studies in previous section such as one binary dependent variable, one continuous explanatory variable and a discrete covariate.

A functional logistic regression model was fitted to the land consolidation data. Table 4.11 gives the descriptions and codes of corresponding variables. The response indicates willingness of peasants to consolidation in Susuzköy Village, which is named by CHO attributing the choice of owners. The response has been of concern to survey and agriculture engineers for years. 0 represents the answer "no" and 1 represents the answer "yes". The continuous independent variable which is the size of area in hectares (AR) that peasants have, takes the values from 105 to 64,674. Finally, the discrete covariate represents the ratio of the number of land shared parcels and the number of individual parcels. The aim of selecting this variable for this study is to see the contribution of the comparison of having shared and individual parcels. 0 means the number of land shared parcels is greater than the number of individual parcels and 1 means majority of parcels are private, that is, belong to individuals. The covariate is named by SP. This is due to the fact that the

area size of owners and being a private owner can greatly alter the behavior of him and so the chances of inventing the consolidation.

Before analysis, the distribution of the continuous explanatory variable AR in data should be determined. In the sense that, a histogram and a density plot of the variable AR are drawn, as presented in Figure 4.5 and Figure 4.6. So the figure gave an idea about the distribution may be exponential. We applied Kolmogorov-Smirnov test to test for an exponential distribution and find p-value is approximately 0.7. Therefore, our data follows an exponential distribution with mean value of 7049.69 and  $\lambda = 0.00014$ .

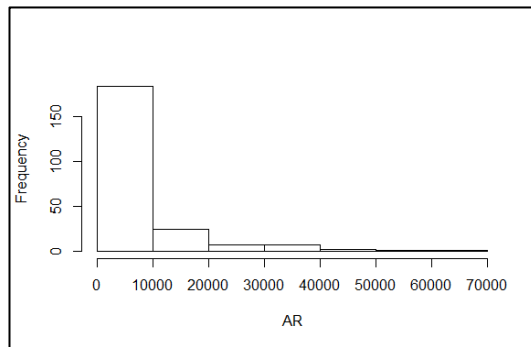


Figure 4.5 Histogram of the explanatory variable AR

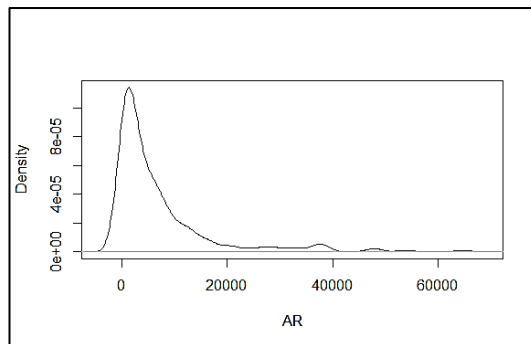


Figure 4.6 Density plot of the explanatory variable AR

In this thesis, we have fitted logistic regression model to land consolidation data to illustrate the behaviors of R-squares. After fitting the logistic regression model to the data, the estimated logit is obtained by the following expression

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

and so

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = 0.4852 + 0.0003AR - 0.5117SP$$

On the other hand, the model with other types of AR have been fitted and omitted SP from the model. To be consistent with simulation studies and real life as well, because these are more often used types, instead of AR,  $AR^3$ ,  $\ln(AR)$  and  $\sqrt{AR}$  have been used. In addition AR have been categorized into two and three categories considering AR having exponential distribution.

For categorizing AR with optimal intervals, the results of calculations about determining the cutpoints if the explanatory variable has exponential distribution with parameter  $\lambda$  in Section 2.3.1 are used. Corresponding the cutpoint to categorize AR into  $k = 2$  categories will be  $\frac{1.5936}{0.00014} = 11,382.86$ . For  $k = 3$ , the values are

$$\frac{1.0176}{0.00014} = 7268.57 \text{ and } \frac{2.6112}{0.00014} = 18,651.43.$$

$R^2$  values corresponding to all fitted types of models were calculated and presented in Table 4.12. The results are consistent with both theoretical and simulation findings about the comparisons of the magnitudes of R-squares. Overall  $R_N^2$  are bigger than  $R_M^2$  and  $R_L^2$  at all models. Furthermore,  $R_M^2$  and  $R_L^2$  tended to be so close values revealing very little changes in the null deviances of all models.

Some inferences may be drawn about what the correct model is, from the same table, since high values of R-squares are regarded as satisfactory for choosing the true model and low values as a sign of much remaining unexplained. In the sense that, considering that the low values indicate warning evidences that these models

may be misspecified, based on the Table 4.12, and having the likelihood ratio test with  $p$ -value = 0.000, deviance goodness of fit test with  $p$ -value = 0.984 and a percentage of 85.2 of concordant pairs which means a higher predicted probability, we assumed with the considered variables that  $\ln(AR)$  model is the correct model for describing the land consolidation data.

Table 4.12  $R^2$  values associated with all models for land consolidation data

Type of model	$R_L^2$	$R_M^2$	$R_N^2$
$AR$	0.174	0.171	0.259
$AR^3$	0.029	0.030	0.046
$\ln(AR)$	0.300	0.276	0.419
$\sqrt{AR}$	0.262	0.246	0.373
$k = 2$	0.077	0.079	0.120
$k = 3$	0.094	0.096	0.146
Omission SP	0.166	0.163	0.248

According to the results of  $\ln(AR)$  model given with Equation (4.3), with one unit increase in the area, the odds of accepting land consolidation increases  $\exp(1.0513) = 2.86$  times.

$$\hat{\pi} = \frac{\exp(-6.2985 + 1.0513\ln(AR) - 0.4411SP)}{1 + \exp(-6.2985 + 1.0513\ln(AR) - 0.4411SP)} \quad (4.3)$$

To examine the effects of misspecification on the ARE of R-square statistics, since their sampling variances are required, a bootstrap study with  $B = 2,000$  bootstrap replications to this real data have been performed. After finding sample variances using this way and fitting the logistic regression to the bootstrap data, the values of ARE are calculated. These calculations are based on  $\ln(AR)$  model since it is assumed to be as the correctly specified model, for this study. The ARE results are presented in Tables 4.13-4.15.

Table 4.13 ARE's of each  $R^2$  statistics on the base of  $\ln(AR)$

Type of model	ARE		
	$R_L^{2*}, R_L^2$	$R_M^{2*}, R_M^2$	$R_N^{2*}, R_N^2$
AR	0.45	0.42	0.38
$AR^3$	0.51	0.40	0.35
$\ln(AR)$	1.00	1.00	1.00
$\sqrt{AR}$	0.70	0.71	0.67
$k = 2$	4.92	3.11	2.96
$k = 3$	2.93	1.94	1.87
Omission SP	0.44	0.40	0.36

It is clear from the Table 4.13 that all  $R^2$  statistics seem influenced by misspecification substantially. The most effected is  $R_N^2$  with up to 65% loss in efficiency. Taking the third power of the explanatory variable causes great losses as conveniently with simulation results. In addition, using the variable AR without any transformation substantially influences the efficiencies, adversely. On the other hand, categorizing the values of AR has significant influence on efficiency and provides a substantial gain in efficiency. We may attribute this result to using of optimal intervals. To better understand, the data with equiprobable intervals have been regrouped, under equal conditions. As a result of this grouping, ARE values are calculated and given in Table 4.14. Reduction in ARE's is seen obviously from the table. So categorization type has significant influence on efficiency. But there is still no loss in efficiency of R-squares. Therefore, it may be said that these statistics are robust against categorization, for this study.

Table 4.14 ARE's of each  $R^2$  statistics on the base of  $\ln(AR)$  for categorizing

Type of model	ARE		
	$R_L^{2*}, R_L^2$	$R_M^{2*}, R_M^2$	$R_N^{2*}, R_N^2$
$k = 2$	1.62	1.25	1.30
$k = 3$	1.11	1.02	1.02

When we compare the efficiencies of all three coefficients of determination in Table 4.15, it is reasonable to infer that  $R_M^2$  is the most efficient statistic followed by  $R_L^2$  and  $R_N^2$  in row.  $R_M^2$  and  $R_L^2$  seem equal in terms of efficiency when AR has been categorized. They have almost the same efficiency and so they are not superior to each other.

Table 4.15 ARE's between each  $R^2$  statistics

<i>Type of model</i>	<i>ARE</i>		
	$R_L^2, R_M^2$	$R_N^2, R_L^2$	$R_N^2, R_M^2$
<i>AR</i>	0.71	0.61	0.44
$AR^3$	0.85	0.50	0.43
$\ln(AR)$	0.67	0.73	0.49
$\sqrt{AR}$	0.66	0.69	0.46
$k = 2$	1.05	0.44	0.46
$k = 3$	1.01	0.47	0.47
<i>Omission SP</i>	0.72	0.60	0.43

As a consequence, it is clear that these results are similar to the simulation results which imply that the most efficient  $R^2$  is geometric mean squared improvement  $R^2$ .

## CHAPTER FIVE

### CONCLUSIONS

Misspecification has great influence on many components of not only linear regression but also generalized linear models. It has effects particularly on the test statistics, the dependent and the independent variables and the estimates of parameters of regression, in terms of both biasedness and efficiency (Chao, Palta, Young, 1997; Gail et al., 1984; Neuhaus & Jewell, 1993). The aim of this thesis is to investigate the effects of misspecification on  $R^2$  statistics in logistic regression models due to ARE. These statistics are utility to measure how well a model fits the data, although they are alone not enough to judge the usefulness of the model. Some other analyses such as the values of goodness of fit statistics (likelihood ratio statistic, Pearson chi-square) should be taken into consideration.

In linear regression,  $R^2$  statistic which is also called explained variance is defined as the proportion of variance about the mean explained by the regression. This explained variance is measured based on error sum of squares. Linear regression models have only one error variation criterion for continuous dependent variables. In logistic regression, however, there are several error variation criteria such as squared error, entropy etc. for binary dependent variable. Therefore, there is not one way to measure the strength of association between the dependent variable and all of the independent variables. In this sense that, so many  $R^2$  analog statistics were derived by some authors failing to agree on one statistic. In this thesis, totally ten well-known  $R^2$  statistics have been explained separately in Chapter 3. Some suggestions that were made by authors about the most convenient  $R^2$  and eight criteria that Kvalseth (1985) described have been given. These most frequently used  $R^2$  statistics have been compared based upon significant contributions of some authors such as Hagle and Mitchell (1992), Hu, Palta and Shao (2006), Menard (2000), Mittlböck and Schemper (1996), Veall and Zimmerman (1996). At the end, a detailed discussion of the three most frequently used and suggested  $R^2$  statistics for logistic models which are likelihood ratio  $R^2$  ( $R_L^2$ ), geometric mean improvement



$(R_M^2)$  and adjusted geometric mean improvement  $(R_N^2)$  have been presented. In this thesis, simulation studies and the application have been designed to compare them.

The concept of ARE has been found to be appropriate for measuring the effects of  $R^2$  statistics under misspecification. ARE which is a useful technique for the comparison of related statistics have been examined in detail in Chapter 2 in two parts such as ARE in estimation and ARE in testing. In estimation, when there is at least one alternative for an estimator, it should be plausible to reveal the most efficient one for basing inferences and interpretations. ARE provides a perceptivity while comparing alternative estimators that are convenient to use. This concept attributes to the measures of performance of two estimators taking the ratio of their variances. Lagakos (1988a), Begg and Lagakos (1990, 1993), Noether (1955), Pitman (1949), Serfling (1980, 2011) studied ARE with different points of view. In testing, Pitman (1949) introduced the earliest approach to ARE in testing and the only major requirement is the information about asymptotic distribution of the test statistic. So among others Pitman approach is widely applicable approach. Section 2.2.2 has included the proof of Pitman's theorem with detailed calculations. In the end, it has been shown that when specific conditions are satisfied, ARE of two test statistics equals the limit of their variances. Using this approach, the performances of  $R^2$  statistics under misspecification have been measured, reasonably.

In this thesis, we have focused on three frequently encountered types of misspecification. They involve discretizing a continuous explanatory variable, omission of a covariate and using wrong functional form of an explanatory variable. Categorization may be the most frequently used technique in especially medical research, because of simplifying the interpretation of models. However, to encourage undesirable results is unavoidable. It is important to decide for the number of categories ( $k$ ) and correct category cutpoints. Cox (1957) suggested that efficiency of a test may be used as a criterion for cutpoint selection and proposed the average information loss caused by categorizing random variable  $X$ . It seems that maximizing ARE is needed to reduce the information loss. So the value maximizing the ARE will give the cutpoint value.

Cox (1957) made numerical recommendations for a normally distributed explanatory variable. His formula can be applied to other distributions such as exponential distribution. In literature, exponential distribution with parameter  $\lambda = 1$  has already been studied. Connor (1972) and Lagakos (1988a) studied on ARE of test statistics when categorizing for up to 6 optimal categories and for explanatory variable having the distributions of uniform, normal and exponential with parameter  $\lambda = 1$ . We have tried to extend the results for other values of  $\lambda$ , as mentioned elaborately in Section 2.3.1. For different  $\lambda$  parameters, it has been derived that the cutpoints may be calculated when the number of categories is two by  $1.5936/\lambda$  and when the number of categories is three by  $1.0176/\lambda$  and  $2.6112/\lambda$ . Larger the parameter  $\lambda$  gets the smaller cutpoints we have. Since  $\lambda$  is the inverse of the mean, the increased  $\lambda$  implies decreased mean. Moreover, it appears that ARE values are adversely affected by  $\lambda$  and become quite lower. Therefore, it has been concluded that, if the distribution of data is determined as exponential with large values of parameter, categorization causes inefficient results of test statistics, otherwise categorized models are reasonably safe. Another extension has been given in the same section applying the loss of information formula to Weibull distribution with  $\gamma = 0$ ,  $\delta = 1$  and  $\beta = 2$ . It has been found that categorizing an explanatory variable having Weibull distribution with these parameters does not make a destructive effect on efficiency. In general, as the number of categories increases, for all distributions with all parameters, the categorization becomes safer, as expected.

These results have provided a basis for misspecification effect on the efficiency of  $R^2$  statistics, since ARE calculations and considerations about the test statistics are applicable to the  $R^2$  statistics. In Chapter 4, some numerical results have been presented. Section 4.1 includes simulation studies to see how and how much change occurs in  $R^2$  statistics with respect to different types of misspecification. AREs for each  $R^2$  statistic under correct model versus under the misspecified model have been calculated. The efficiency comparisons for three  $R^2$  statistics with each other under both correct and misspecified models have also been made. In addition, the

influence of sample size has been investigated using 50 and 100 as sample sizes. According to the simulation results, for an explanatory variable having normal distribution, the increased variance generally leads to substantial losses in efficiency of all  $R^2$  statistics when  $X$  has been included in wrong functional form or when it has been categorized in two categories. However, if an explanatory variable has an exponential distribution, then misspecification does not cause any problem. In addition, omitting the other covariate tended to provide a gain in efficiency under misspecification, as Neuhaus (1998) mentioned. In generally, the less effected statistic is seen as  $R_L^2$ .  $R_M^2$  and  $R_N^2$  both give almost the same reaction to misspecification, especially for exponential distribution. It seems that, there is not much to worry if the explanatory variable has exponential distribution.

Even though  $R_L^2$  is recommended for some authors such as Menard (2002),  $R_M^2$  seems the most efficient  $R^2$  statistic among three statistics when it is accurate that there is no misspecification. Under misspecification,  $R_M^2$  is still the most efficient statistic except when  $X^3$  is used instead of  $X$  mistakenly. As simulation results show, using the third power of continuous variable in the model causes some problems for the coefficients of determination, such that  $R_M^2$  and  $R_N^2$  have great loss in efficiency, especially when  $X$  having normal distribution with increase in variance. Exponentially distributed  $X$  makes no difference in efficiency of these three statistics with regard to each other regardless of misspecification type. Simulation results show that  $R_M^2$  is more suggestible than  $R_L^2$  in terms of efficiency.

An application on land consolidation has been presented in Section 4.2 to see the behaviors of  $R^2$  statistics when a real data is used. Agriculture is a field that logistic regression is applied frequently. In this thesis, logistic regression has been fitted to land consolidation data which efforts a wide range of applications in around the world. Two cases have been shed light on by this application. Firstly, we provide a gain in information about the attributes that affect the peasants' behaviors by applying logistic regression. Using bootstrap technique, secondly, the efficiencies of

R-squared statistics with land consolidation data have been attained. To be able to predict the willingness of peasants, we have included two variables to the analysis. There is certainly a wide range of variables that influences the willingness. Among them, we have given preference to the size of area that peasants have and to the ratio of the number of land shared parcels and the number of individual parcels. According to the performed application, we have concluded that the model conducted with the explanatory variable taking the natural logarithmic is the correct model with the considered variables for describing the data and the area size has influence on the choice of peasants, positively.

Based on results of the bootstrap method, we see that  $R^2$  statistics seem influenced by misspecification substantially. The most effected by misspecification is  $R_N^2$ . It may be said that all these statistics are robust against categorization, for this study. This result has not changed regardless of using optimal and equiprobable intervals. The statistics have kept being robust against categorization for any number of categories. On the other hand, the efficiencies of  $R^2$  statistics with each other have been calculated under misspecification or under the model that assumed to be correct. Finally, there are clear indications that  $R_M^2$  is the most efficient statistic over the other two, due to application supporting the simulation study. Only in categorization case, it seems to have equivalent efficiencies with  $R_L^2$ .

The results should be interpreted considering that As a result of both simulation and application, it has been concluded that there are sufficient indications to believe that  $R_M^2$  is more efficient than the others even if we suspected that we would make a specification error. Our recommendation is to select the  $R^2$  statistic associated with the logistic regression analysis, carefully and when these R-squared statistics are calculated, it would be judicious decision to interpret the results considering  $R_M^2$ .

It should be noted that the inferences obtained from this study are limited with considered scenarios. For further researches, this study can be extended to find the ARE under different generalized linear model such as probit models. The distribution

of continuous explanatory variable can be chosen uniform and beta to observe their effects. The variance of the continuous variable can be designed moderately, for normal distribution. ARE formula for test statistics can be extended for coefficients of determinations, theoretically. Different correlation coefficients between the variables can be considered. Moreover, this study can be extended to measure the performances of other comparable statistics.

## REFERENCES

- Adewale, A. J., & Wiens, D. P. (2009). Robust designs for misspecified logistic models. *Journal of Statistical Planning and Inference*, 139, 3-15.
- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- Aldrich, J. H., & Nelson, F. D. (1985). *Linear probability, logit and probit models*. Beverly Hills: SAGE.
- Amemiya, T., & Powell, J. L. (1983). A comparison of the logit model and normal discriminant analysis when the independent variables are binary. In *Studies in Econometrics, Time Series and Multivariate Statistics* (3-30). New York: Academic Press.
- Battaglin, W. A., & Goolsby, D. A. (1996). Using GIS and logistic regression to estimate agricultural chemical concentrations in rivers of the Midwestern USA. In K. Kovar, & H. Natchnebel, (Eds.). *Application of Geographic Information Systems in Hydrology and Water Resources Management*, (253-260). Oxfordshire: IAHS Press.
- Begg, M. D., & Lagakos, S. (1990). On the consequences of model misspecification in logistic regression. *Environmental Health Perspectives*, 87, 69-75.
- Begg, M. D., & Lagakos, S. (1993). Loss in efficiency caused by omitting covariates and misspecifying exposure in logistic regression models. *Journal of the American Statistical Association*, 88 (421), 166-170.
- Blaikie, N. (2003). *Analyzing quantitative data: From description to explanation*. London: SAGE.
- Bofinger, E. (1970). Maximizing the correlation of grouped observations. *Journal of the American Statistical Association*, 65 (332), 1632-1638.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B (Methodological)*, 26 (2), 211-252.

- Boyrac, Z., & Üstündağ, Ö. (2008). Kırsal alanlarda arazi toplulaştırma çalışmalarının önemi. *e-Journal of New World Sciences Academy*, 3 (3), 563-578.
- Brooks, C. (2008). *Introductory econometrics for finance*. New York: Cambridge University Press.
- Casella, G., & Berger, R.L. (2002). *Statistical inference* (2nd ed.). Duxbury: Thomson Learning Inc.
- Chao, W., Palta, M., & Young, T. (1997). Effect of omitted confounders on the analysis of correlated binary data. *Biometrics*, 53 (2), 678-689.
- Cimpoieş, D. (2007). Lessons from developments strategies: socio-economic impacts of land policy in the republic of Moldova. *Buletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca*, 64 (1/2), 321-326.
- Connor, R. J. (1972). Grouping for testing trends in categorical data. *Journal of the American Statistical Association*, 67 (339), 601-604.
- Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, 53 (280), 543-547.
- Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. Chapman and Hall.
- Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data*. London: Chapman and Hall.
- Cox, D. R., & Wermuth, N. (1992). A comment on the coefficient of determination for binary responses. *The American Statistician*, 46 (1), 1-4.
- Cramer, J. S. (1964). Efficient grouping, regression and correlation in Engel curve analysis. *Journal of the American Statistical Association*, 59 (305), 233-250.
- DeMaris, A. (2002). Explained variance in logistic regression a Monte Carlo study of proposed measures. *Sociological Methods & Research*, 31 (1), 27-74.

- Demetriou, D., Stillwell, J., & See, L. (2012). Land consolidation in Cyprus: why is an integrated planning and decision support system required?. *Land Use Policy*, 29 (1), 131-142.
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.
- Eeden, C. V. (1963). The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. *The Annals of Mathematical Statistics*, 34 (4), 1442-1451.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70 (352), 892-898.
- Efron, B. (1978). Regression and ANOVA with zero-one data: Measures of residual variation. *Journal of the American Statistical Association* 73 (361), 113-121.
- Erees, S. & Demirel, D. (2012). Omitted variable bias and detection with RESET test in regression analysis. *Anadolu University Journal of Science and Technology – B Theoretical Sciences*, 2 (1), 1-19.
- Food and Agriculture Organization of the United Nations, (2003). *FAO Land Tenure Studies 6 The Design of Land Consolidation Pilot Projects in Central and Eastern Europe*. Publishing Management Service, Information Division, FAO, Viale delle Terme di Caracalla, Rome, Italy.
- Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71 (3), 431-444.
- Gün, S. (2003). Legal state of land consolidation in Turkey and problems in implementation. *Pakistan Journal of Biological Sciences*, 6 (15), 1380-1383.



- Haberman, S. J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77, 568-580.
- Hagle, T. M., & Mitchell, G. E. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36 (3), 762-784.
- Hájek, J. (1962). Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, 33, 1124-1147.
- Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists*. London: Academic Press, Inc.
- Helland, I. S. (1987). On the interpretation and use of  $R^2$  in regression analysis. *Biometrics*, 43 (1), 61-69.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.
- Hu, B., Palta, M., & Shao, J. (2006). Properties of  $R^2$  statistics for logistic regression. *Statistics in Medicine*, 25, 1383-1395.
- Jarque, C. M. (1981). Efficient grouping of observations in regression analysis. *International Economic Review*, 22 (3), 709-718.
- Johnston, J. (1984). *Econometric theory*. New York: McGraw-Hill.
- Kay, R., & Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, 74 (3), 495-501.
- Keele, L. J. (2008). *Semiparametric Regression for the Social Sciences*. New York: Wiley.
- Kvalseth, T. O. (1985). Cautionary note about  $R^2$ . *The American Statistician*, 39, 279-285.

- Lachin, J. M. (2000). *Biostatistical methods: The assessment of relative risks*. John Wiley & Sons.
- Lagakos, S. W. (1988a). The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika*, 75, 156-160.
- Lagakos, S. W. (1988b). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine*, 7, 257-274.
- Leightner, J. E., & Inoue, T. (2007). Tackling the omitted variables problem without the strong assumptions of proxies. *European Journal of Operational Research*, 178 (3), 819-840.
- Lerman, Z., & Cimpoieş, D. (2006). Land consolidation as a factor for rural development in Moldova. *Europe-Asia Studies*, 58 (3), 439-455.
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press.
- McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, 3, 303-328.
- McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4 (1), 103-120.
- Magee, L. (1990). R2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, 44 (3), 250-253.
- Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54 (1), 17-24.
- Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). SAGE.

- Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. SAGE.
- Minetos, D., & Polyzos, S. (2009). Analysis of agricultural land use transformations in Greece: a multinomial logistic regression model at the regional level. *International Journal of Sustainable Development and Planning*, 4 (3), 189-209.
- Mittlböck, M., & Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, 15, 1987-1997.
- Msoffe, F. U., Said, M. Y., Ogutu, J. O., Kifugo, S. C., Leeuw, J., Gardingen, P., et al. (2011). Spatial correlates of land-use changes in the Maasai-Steppe of Tanzania: Implications for conservation and environmental planning. *International Journal of Biodiversity and Conservation*, 3 (7), 280-290.
- Mueller, T. G., Cetin, H., Fleming, R. A., Dillon, C. R., Karathanasis, A. D., & Shearer, S. A. (2005). Erosion probability maps: calibrating precision agriculture data with soil surveys using logistic regression. *Journal of Soil and Water Conservation*, 60 (6), 462-468.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691-692.
- Neuhaus, J. M. (1998). Estimation efficiency with omitted covariates in generalized linear models. *Journal of the American Statistical Association*, 93 (443), 1124-1129.
- Neuhaus, J. M., & Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, 80 (4), 807-815.
- Noether, G.E. (1955). On a theorem of Pitman. *The Annals of Mathematical Statistics*, 26 (1), 64-68.
- O'Brien, S. M. (2004). Cutpoint selection for categorizing a continuous predictor. *Biometrics*, 60, 504-509.

- Ohtani, K., & Tanizaki, H. (2004). Exact distributions of  $R^2$  and adjusted  $R^2$  in a linear regression model with multivariate  $t$  error terms. *Journal of the Japan Statistical Society*, 34 (1), 101-109.
- Pampel, F. C. (2000). *Logistic regression a primer*. SAGE.
- Panik, M. J. (2005). *Advanced statistics from an elementary point of view*. London: Elsevier Academic Press.
- Pašakarnis, G., & Maliene, V. (2010). Towards sustainable rural development in Central and Eastern Europe: Applying land consolidation. *Land Use Policy*, 27, 545-549.
- Pitman, E. J. G. (1949). *Lecture notes on non-parametric statistical inference*. Columbia University.
- Prais, S. J., & Aitchison J. (1954). The grouping of observations in regression analysis. *Review of the International Statistical Institute*, 22 (1/3), 1-22.
- Raut, N., Sitaula, B. K., Vatn, A., & Paudel, G. S. (2011). Determinants of adoption and extend of agricultural intensification in the central mid-hills of Nepal. *Journal of Sustainable Development*, 4 (4), 47-60.
- Ryan, T. P. (1997). *Modern regression methods*. New York: Wiley.
- Saikkonen, P. (1989). Asymptotic relative efficiency of the classical test statistics under misspecification. *Journal of Econometrics*, 42, 351-369.
- Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, 74 (2), 385-391.
- Schroeder, J. C., Olshan, A. F., Baric, R., Dent, G. A., Weinberg, C. R., Yount, B., et al. (2001). Agricultural risk factors for t(14;18) subtypes of non-Hodgkin's lymphoma. *Epidemiology*, 12 (6), 701-709.

- Serfling, R. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.
- Serfling, R. (2011). Asymptotic relative efficiency in estimation. *International Encyclopedia of Statistical Sciences* (68-72).
- Stefanski, L. A. & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, 13 (4), 1335-1351.
- Stuart, A. (1954). Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. *Journal of the American Statistical Association*, 49 (265), 147-157.
- Tosteson, T.D. & Tsiatis, A.A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika*, 75 (3), 507-514.
- Türker, M., & Gülsever Şaban, F. T. Z. (2013). Land Consolidation in Turkey. Retrieved July 07, 2013, from [http://www.fao.org/fileadmin/user\\_upload/Europe/documents/Events\\_2013/TAIEX/5.2\\_Turkey\\_en.pdf](http://www.fao.org/fileadmin/user_upload/Europe/documents/Events_2013/TAIEX/5.2_Turkey_en.pdf).
- Vandaele, W. (1981). Wald, likelihood ratio and Lagrange multiplier tests as an F test. *Economics Letters*, 8 (4), 361-365.
- Veall, M. R., & Zimmermann, K. F. (1996). Pseudo- $R^2$  measures for some common limited dependent variable models. *Journal of Economic Surveys*, 10, 241-259.
- Verbeek, M. (2004). *A guide to modern econometrics* (2nd ed.). John Wiley & Sons.
- Vitikainen, A. (2004). An overview of land consolidation in Europe. *Nordic Journal of Surveying and Real Estate Research*, 1, 25-44.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50 (1), 1-25.

Zhang, Y., & Zhao, J. (2013). Analysis of factors influencing the satisfaction degree of leisure agricultural parks management based on binary logistic model. *Advanced Journal of Food Science and Technology*, 5 (3), 285-288.

## APPENDIX

### Asymptotic distributions of $R^2$ statistics due to corresponding distributions of $X$ for $n = 50$ and $n = 100$

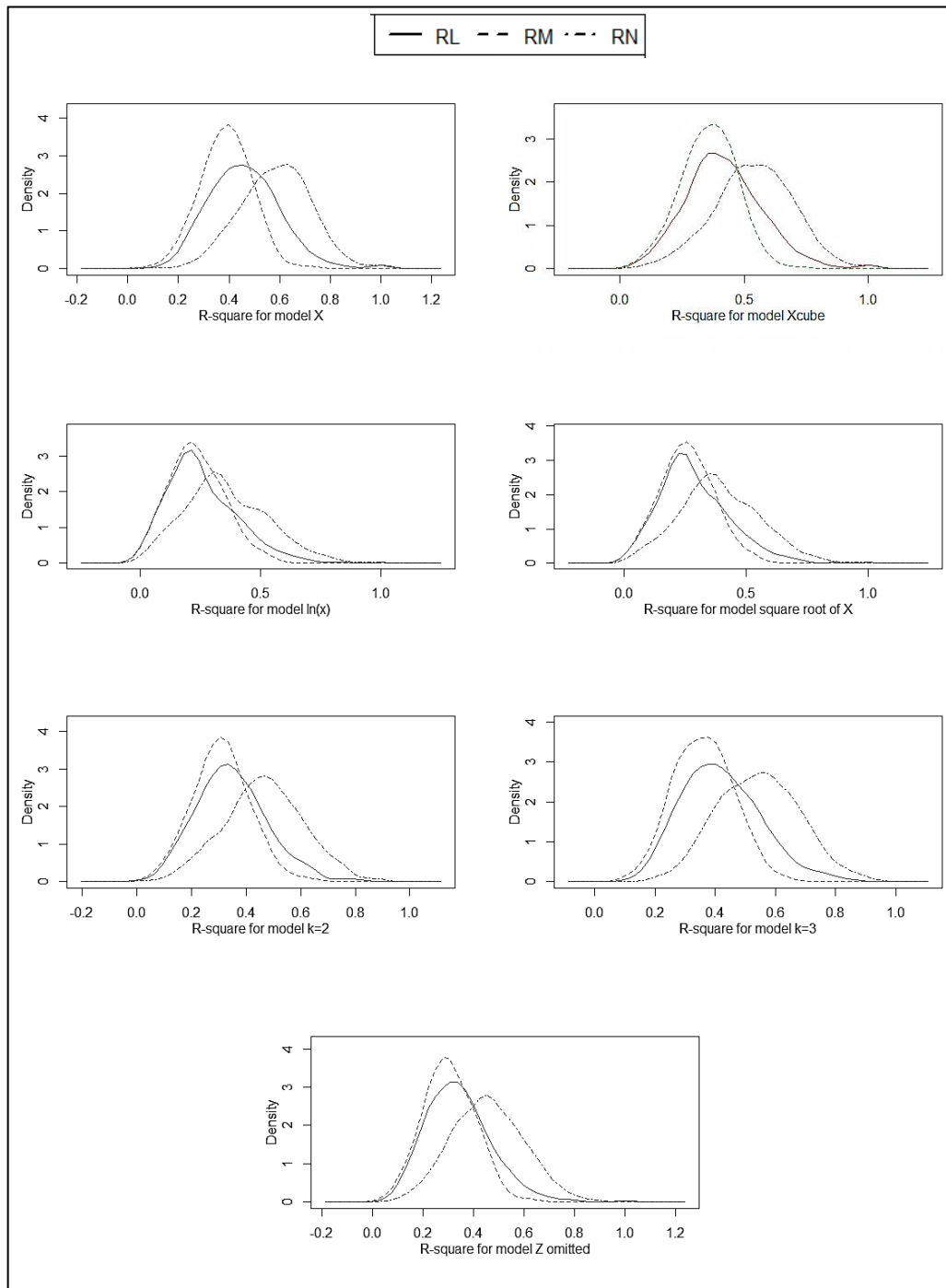


Figure A.1 Asymptotic distributions of R-squares when  $X$  have normal distribution  $(0,1)$  for  $n = 50$

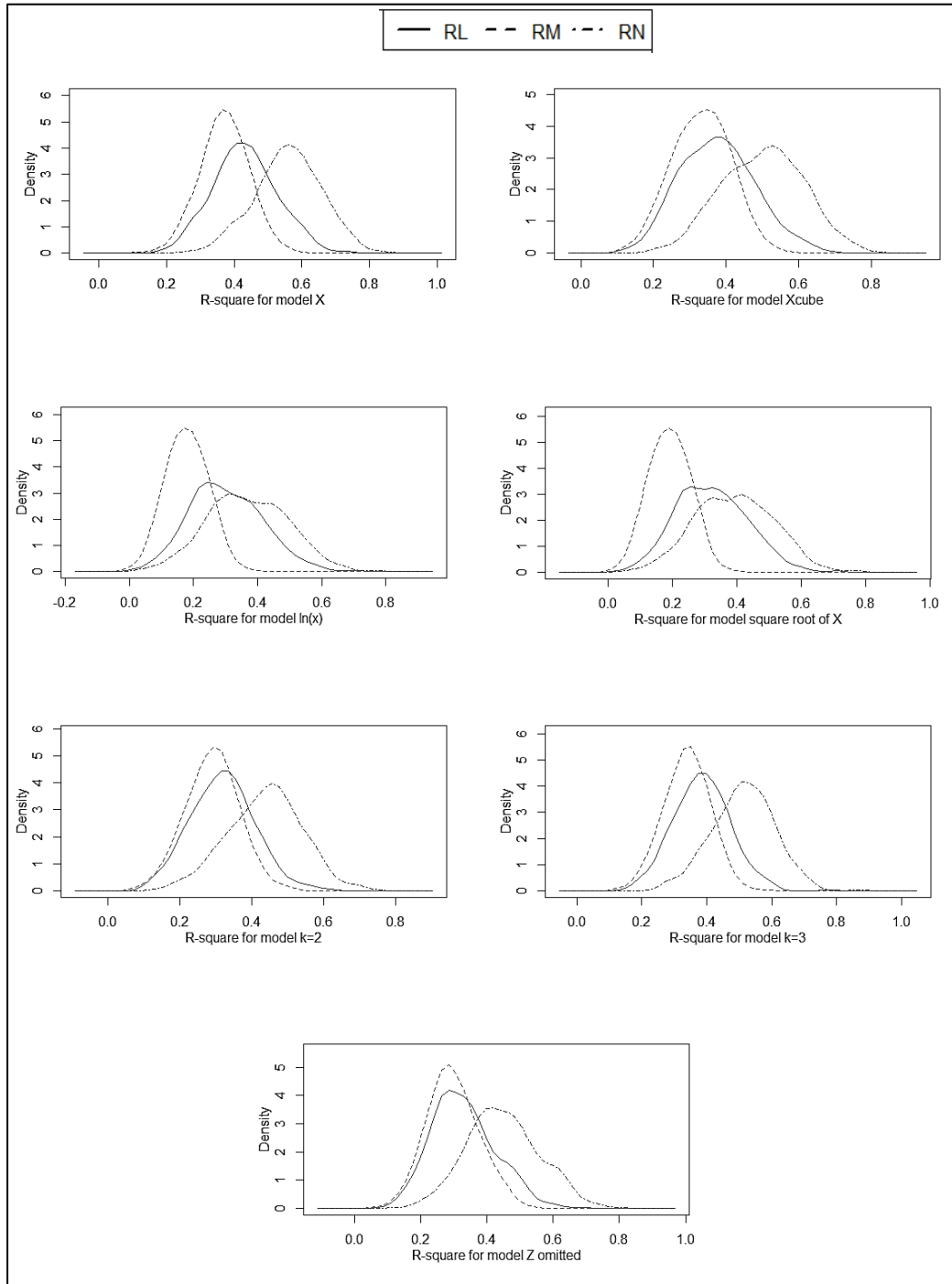


Figure A.2 Asymptotic distributions of R-squares when  $X$  have normal distribution  $(0,1)$  for  $n = 100$



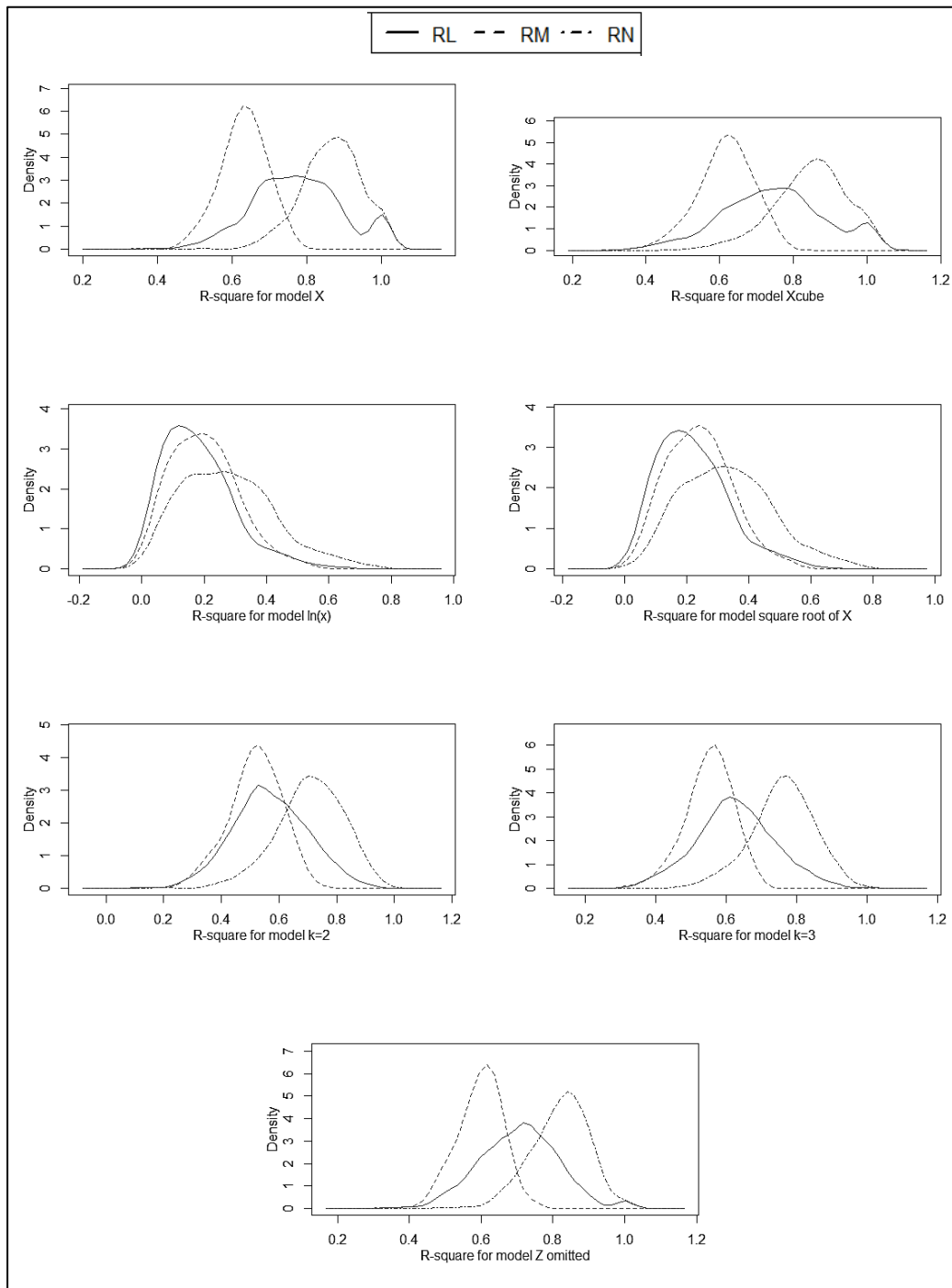


Figure A.3 Asymptotic distributions of R-squares when  $X$  is distributed normal  $(0,9)$  for  $n = 50$

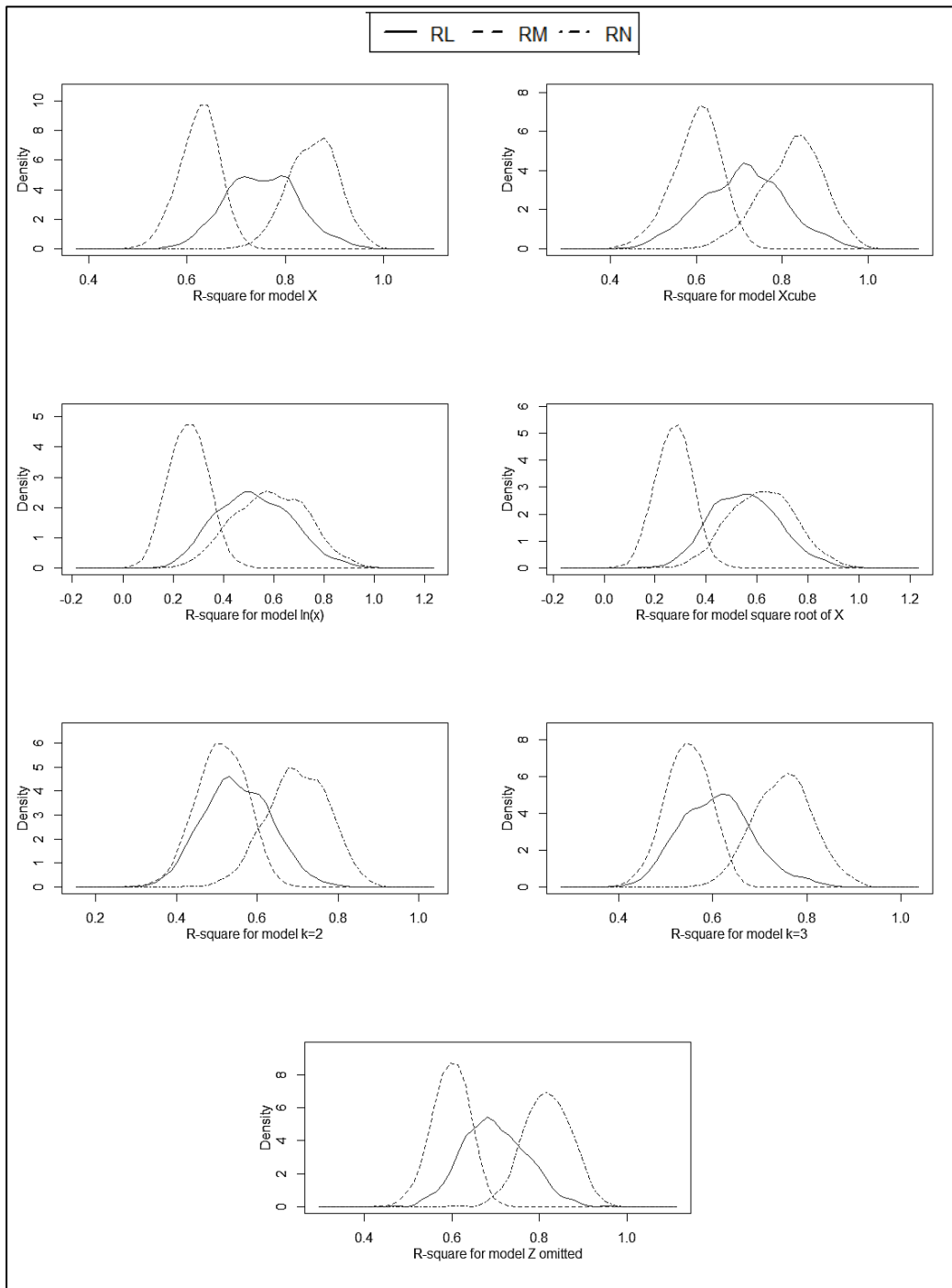


Figure A.4 Asymptotic distributions of R-squares when  $X$  have normal distribution  $(0,9)$  for  $n = 100$

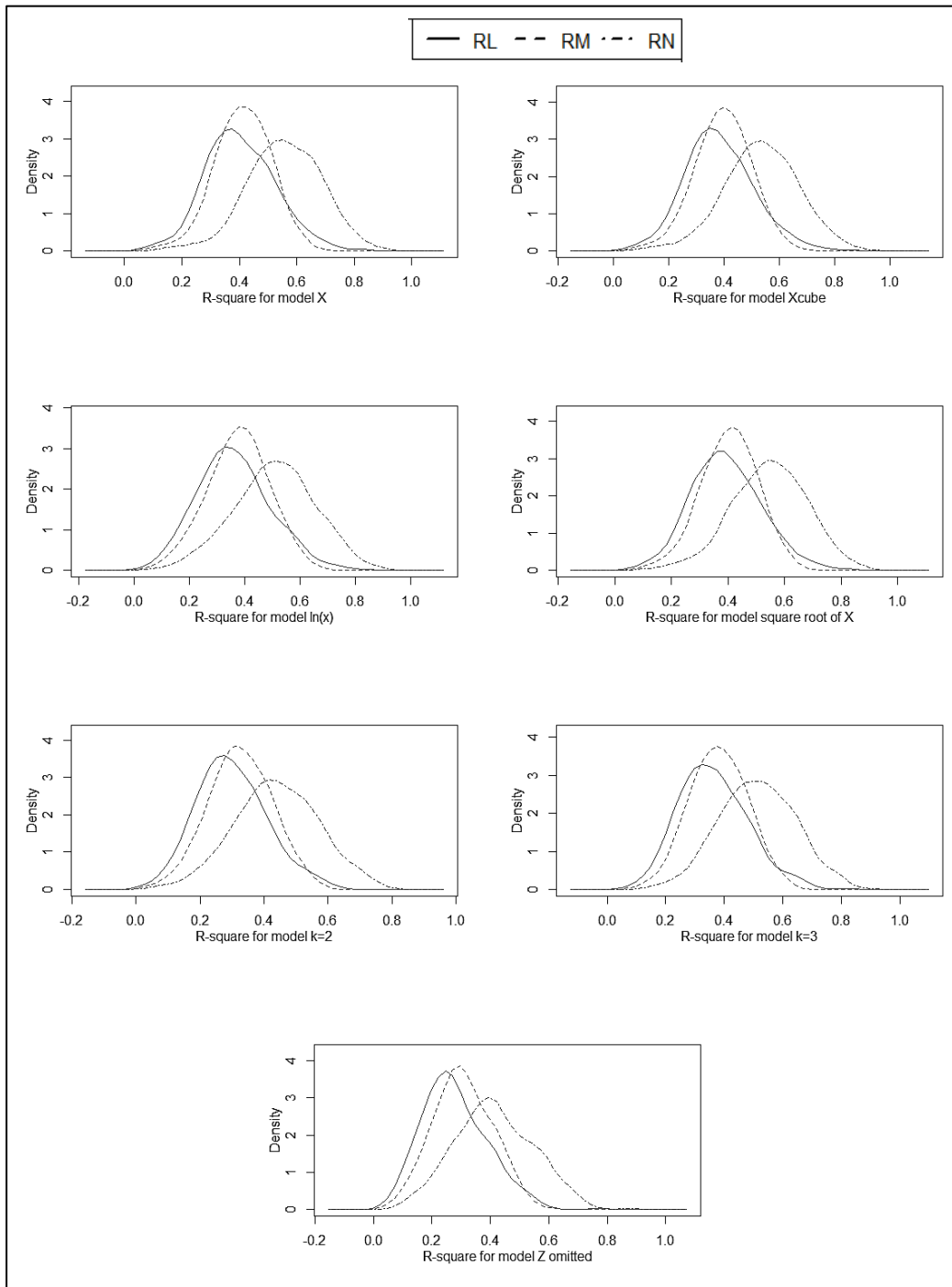


Figure A.5 Asymptotic distributions of R-squares when  $X$  have exponential distribution (1) for  $n = 50$

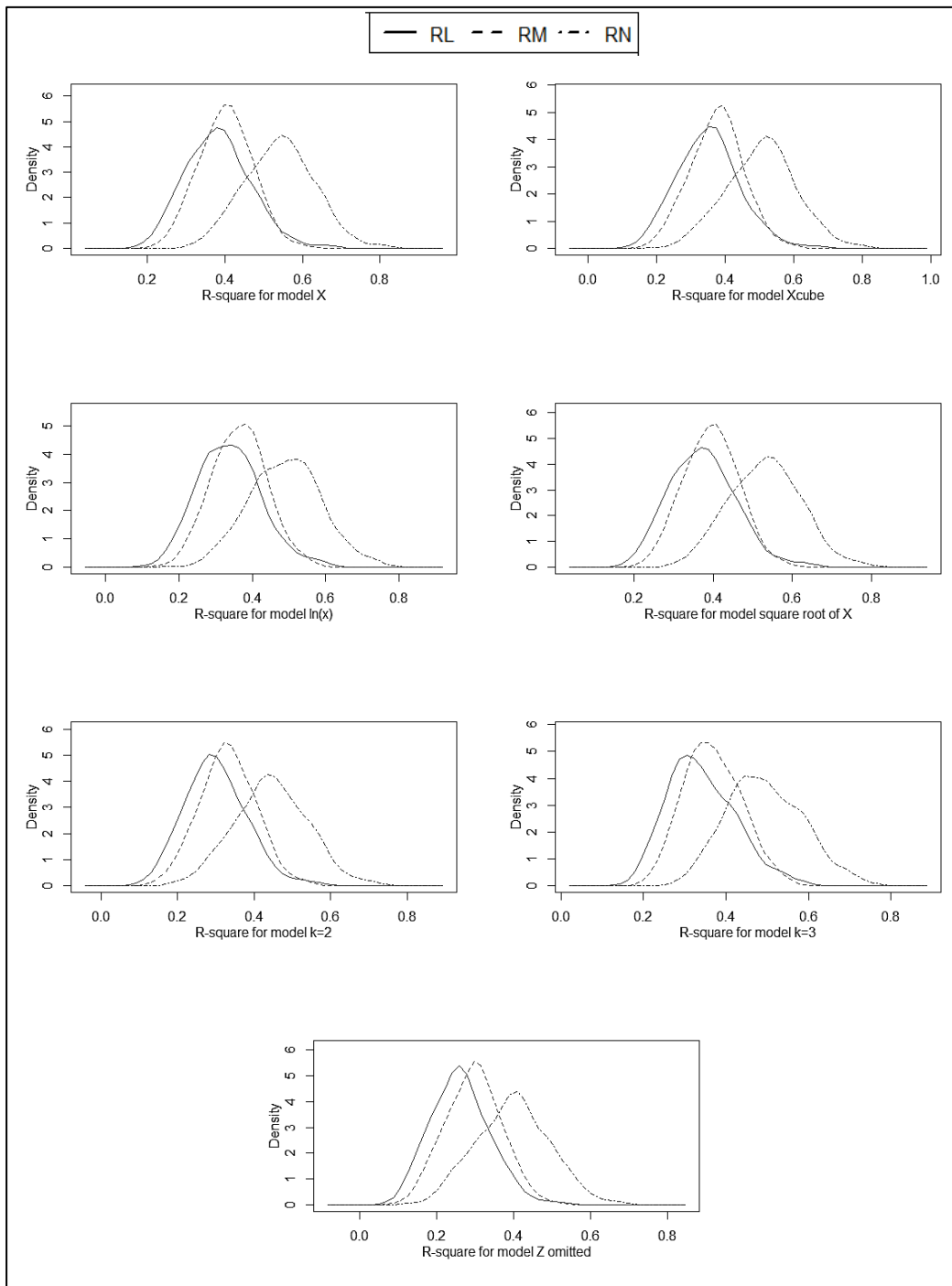


Figure A.6 Asymptotic distributions of R-squares when  $X$  have exponential distribution (1) for  $n = 100$

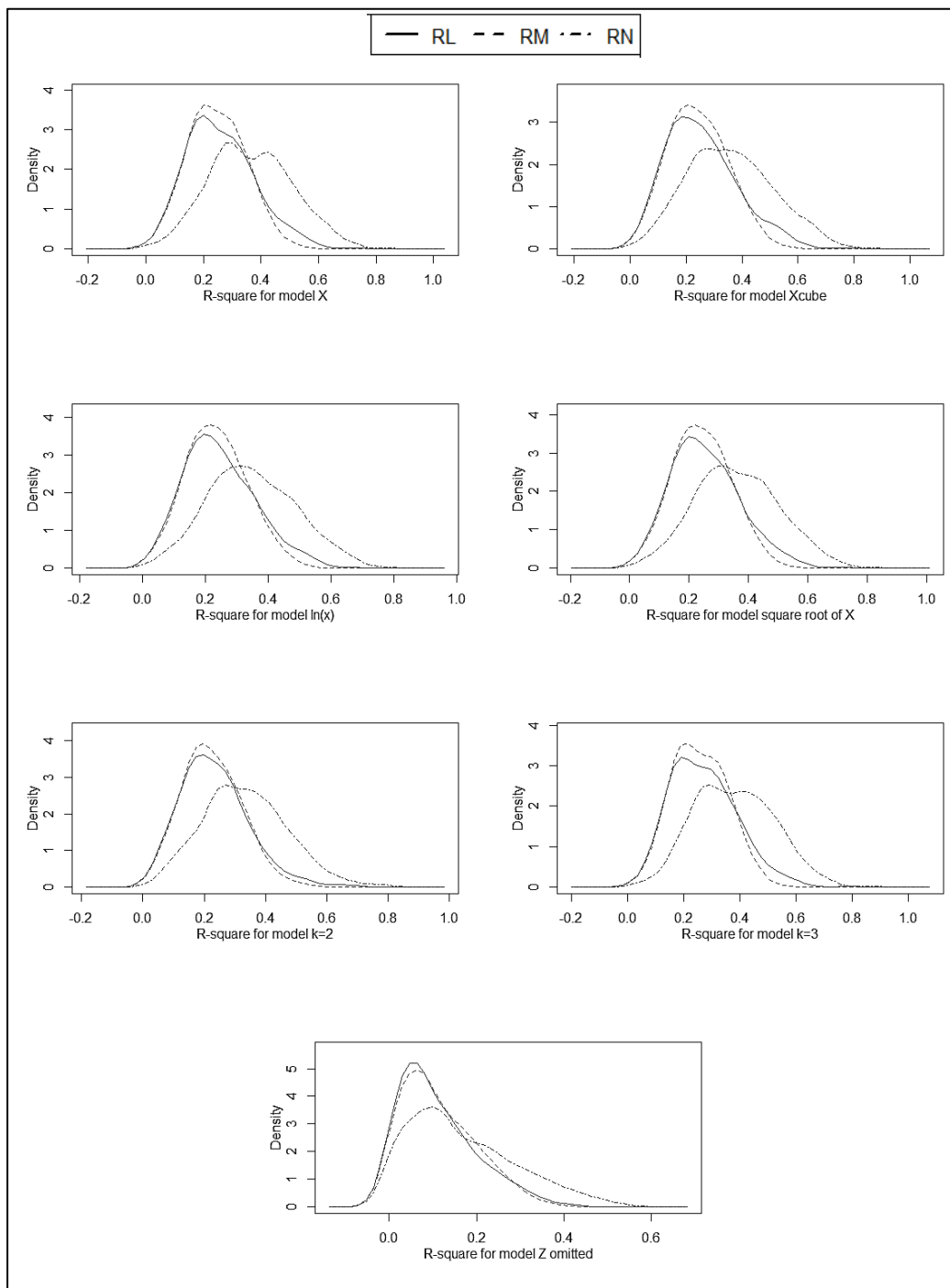


Figure A.7 Asymptotic distributions of R-squares when  $X$  have exponential distribution (3) for  $n = 50$

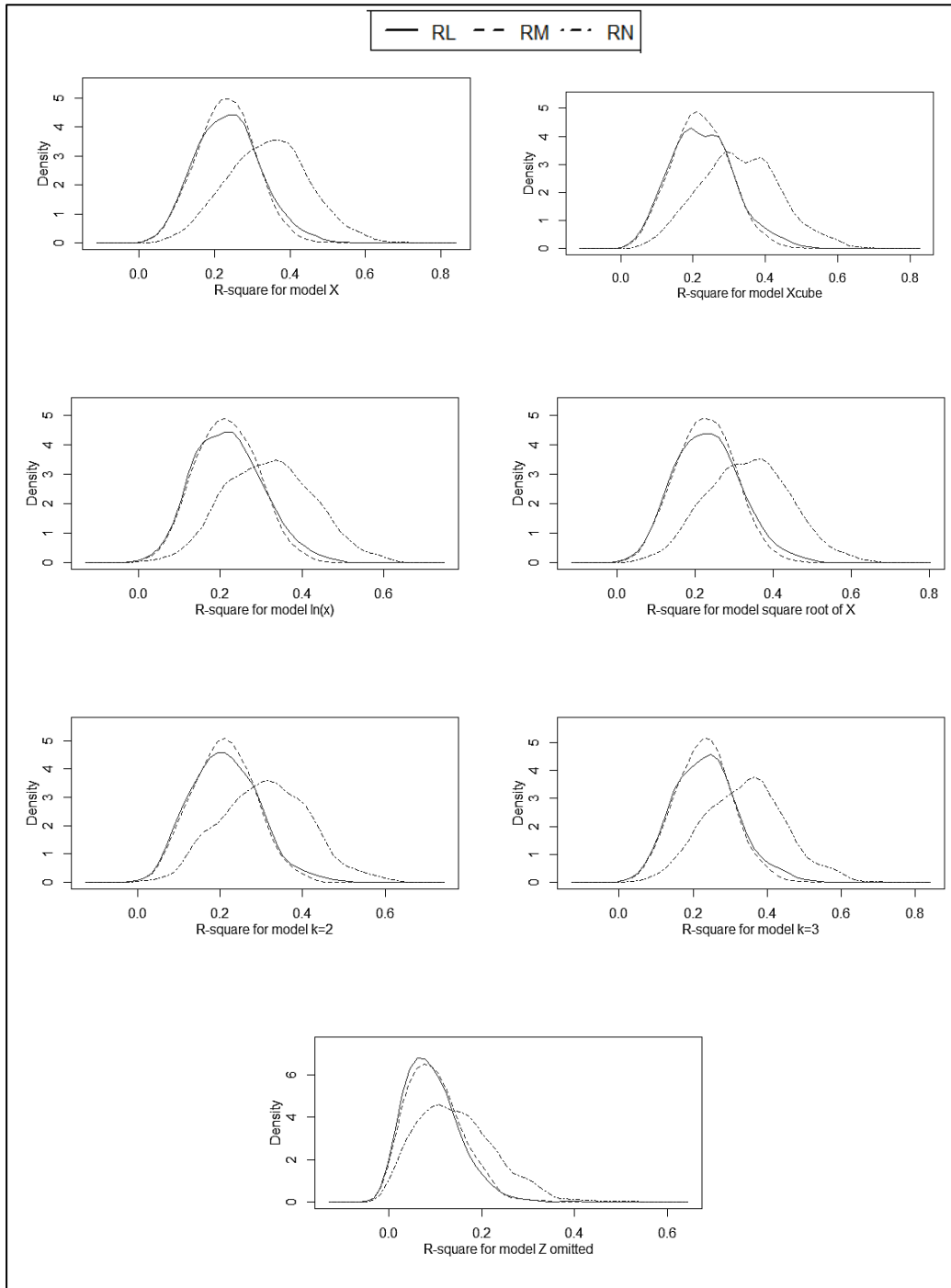


Figure A.8 Asymptotic distributions of R-squares when  $X$  have exponential distribution (3) for  $n = 100$