ANADOLU ÜNİVERSİTESİ

Bilim ve Teknoloji Dergisi B-Teorik Bilimler
Cilt: 3 Sayı: 1 2014
Sayfa: 1-10

ARAŞTIRMA MAKALESİ /**RESEARCH ARTICLE**

Suay EREEŞ [1], Aylin ALIN [2]

EFFECTS OF MISSPECIFYING AN EXPLANATORY VARIABLE ON COEFFICIENTS OF DETERMINATION IN LOGISTIC REGRESSION

ABSTRACT

Misspecifying an explanatory variable is a common problem in logistic regression as it is for all members of generalized linear models. Categorizing a continuous explanatory variable, using wrong functional form of an explanatory variable and omitting an explanatory variable from the model are commonly made misspecifications in logistic regression analysis. Studies show that all of these cases cause a loss in efficiency for test statistics. In this paper, the effects of these types of misspecification on asymptotic relative efficiency of different coefficients of determination are investigated. All calculations and comparisons are based on extensive simulation study using bootstrap methods.

Keywords: Misspecification, R-square, asymptotic relative efficiency, logistic regression, bootstrapping.

LOJİSTİK REGRESYONDA BİR AÇIKLAYICI DEĞİŞKENİN YANLIŞ TANIMLANMASININ BELİRTME KATSAYILARI ÜZERİNDEKİ ETKİLERİ

ÖZ

Açıklayıcı değişkenin yanlış tanımlanması genelleştirilmiş doğrusal modellerin tüm üyeleri için olduğu gibi lojistik regresyon modeli için de genel bir problemdir. Sürekli bir değişkeni kesikli hale getirme, açıklayıcı değişkenin yanlış foksiyonel formunun kullanılması ve bir değişkeni modelden dışlama lojistik regresyonda sıkça yapılan belirleme hatalarıdır. Çalışmalar göstermiştir ki bu gibi hatalar test istatistiklerinde etkinlik kaybına neden olmaktadır. Bu çalışmada, bu gibi belirleme hatalarının, lojistik regresyonda açıklanan değişimi ölçmek için kullanılan farklı determinasyon katsayılarının asimtotik göreli etkinliği üzerindeki etkileri araştırılmıştır. Bootstrap yöntemi kullanılarak detaylı benzetim çalışmaları ile hesaplamalar ve karşılaştırmalar yapılmıştır.

Anahtar kelimeler: Belirleme hatası, R-kare, asimtotik göreli etkinlik, lojistik regresyon, bootstrap

[1], Department of Statistics, Yaşar University, İzmir, Turkey.
E-mail: suay.erees@yasar.edu.tr
[2], Department of Statistics, Dokuz Eylül University, İzmir, Turkey.
E-mail: aylin.alin@deu.edu.tr

Recieved: 10 April 2013 **Revised:** 04 June 2013 **Accepted:** 06 June 2013

1. INTRODUCTION

Correct specification of the model is the most important assumption for the logistic regression model. It means that the model has the correct functional form, does not include irrelevant variables and has all the relevant variables. Without correct specification we will have biased coefficients and less efficient test statistics. Therefore, there are numerous studies in literature regarding this issue for both linear and logistic models. Lagakos (1988), Begg and Lagakos (1990, 1993) and Tosteson and Tsiatis (1988) particularly have showed great interest in the asymptotic relative efficiency (ARE) of tests of association when explanatory variables have been misspecified or omitted in logistic regression models. In this study we look this problem with different perspective. We will investigate the effect of misspecification on the asymptotic relative efficiency of coefficients of determination (R^2).

In logistic regression analysis, in contrast to linear regression analysis, there is no consensus on how R^2 will be calculated. Kvalseth (1985) described eight criteria for a good R^2 statistic (Menard, 2000). There are different R^2 statistics proposed in the literature satisfying some of these properties. For this study, we use three well known R^2 statistics including the ones proposed by McFadden (1974), Cox and Snell (1989) and Nagelkerke (1991). To examine the effects of misspecification on the asymptotic relative efficiency of these statistics, simulation studies are carried out using bootstrap method. These simulation experiments consider the logistic regression model given with (1), with binary outcome (y), a continuous explanatory variable (x) and a discrete covariate (z).

$$E(y) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 z)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 z)} \tag{1}$$

The article is organized as follows. In section 2, we present the types of misspecification and their forms used in this study. In section 3, we review the coefficients of determinations which are used frequently in software programs and in literature and present their asymptotic relative efficiencies. In section 4, we display and discuss the results of simulation study and finally we conclude with section 5.

2. MISSPECIFICATION

2.1.1. Categorizing A Continuous Explanatory Variable

In medical research, 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. However, for whatever reason, categorizing causes misspecification error and efficiency loss for test statistics. 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, the researcher should be careful to choose the cut points. For choosing cut points for different size of categories, Cox (1957) proposed a measure of information loss from grouping given in Eq.(2)

$$\begin{aligned} L_x &= \sum p_i E\left[(X - E(X_i))^2 \middle| X \ in\ the\ ith\ group\right] / \sigma^2 \\ &= 1 - \left[\sum p_i (E(X_i) - E(X))^2\right] / \sigma^2 \\ &= 1 - M \end{aligned} \tag{2}$$

where L_x is the loss of information, $E(X_i)$ is the mean of all observations in the *i*th group and σ is the standard deviation of *x*. Let the size of categories be 2, i.e., *k* = 2. The cut point is then taken as the mean by symmetry and the percentages of individuals for in the two groups being 50.0 and 50.0. For *k* = 3, the three groups should be $(-\infty, \mu - 0.612\sigma)$, $(\mu - 0.612\sigma, \mu + 0.612\sigma)$, and $(\mu + 0.612\sigma, \infty)$. The probabilities of individuals for each of these groups should be 27%, 46% and 27%, respectively. This information loss formula can be applied to other distributions.

2.1.2. Omission Of An Explanatory Variable

In observational studies, to attain an important explanatory variable is sometimes difficult or expensive, and sometimes impossible to measure such as socio-economic status. Therefore omitting this variable from the model may be preferred, easily. The omission of some variables from a regression that affect the dependent variable may cause an omitted variables bias. This bias depends on the correlation between the omitted and included independent variables. 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). Moreover, Begg and Lagakos (1992) have given much attention to the efficiency of tests for association between explanatory and response variables for logistic regression.

2.1.3. Mismodelling A Continuous Explanatory Variable

Using some transformations relating continuous variable may also cause some problems. In medical studies, in particular, because of the complexity of relationships between variables, regression models may not represent the true relationships between these variables, exactly. It may not even be possible to detect when 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, what the consequences will be and whether the results are reliable, in such cases.

3. COEFFICIENTS OF DETERMINATION AND ASYMPTOTIC RELATIVE EFFICIENCIES

In logistic regression analysis, in contrast to linear regression analysis, there is no consensus on how R^2 will be calculated. Kvalseth (1985) described eight criteria for a good R^2 statistic (Menard, 2000). There are different R^2 statistics proposed in the literature satisfying some of these properties. For this study, we use three well known R^2 statistics, one of which has been proposed by McFadden (1974) and denoted by R_L^2

$$\begin{aligned} R_L^2 &= \frac{-2[\ln(L_0) - \ln(L_M)]}{-2[\ln(L_0)]} = \frac{[-2\ln(L_0)] - [-2\ln(L_M)]}{-2[\ln(L_0)]} \\ &= \frac{G_M}{-2[\ln(L_0)]} \end{aligned} \tag{3}$$

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 explanatory variables and G_M is the well-known likelihood ratio statistic. This may simply be seen as the general version R^2 for generalized linear models. Under identity link function, Eq.(3) is equivalent to the ordinary R^2 used for linear regression models.

The second R^2 statistic is proposed by Cox and Snell (1989) and denoted by R_M^2

$$R_M^2 = 1 - \exp\left\{-\frac{2}{n}\left[\ln(L_M) - \ln(L_0)\right]\right\} = 1 - \left(\frac{L_0}{L_M}\right)^{2/n} \tag{4}$$

This statistic cannot have a value of one even the model fits the data perfectly. However the adjusted measure obtained dividing R_M^2 by its maximum possible value permits a value of one. The last R^2 statistic considered in this study is the one proposed by Nagelkerke (1991), denoted by R_N^2

$$R_N^2 = \frac{1 - \exp\left\{-\frac{2}{n}\left[\ln(L_M) - \ln(L_0)\right]\right\}}{1 - \exp\left[2\ln(L_0)/n\right]} = \frac{1 - \left(\frac{L_0}{L_M}\right)^{2/n}}{1 - (L_0)^{2/n}} \tag{5}$$

For more information about R^2 statistics in logistic regression analysis, see Menard (2000, 2002), Mittlböck and Schemper (1996) and Hu, Palta and Shao (2006).

We will compare these statistics in terms of their asymptotic relative efficiency under misspecification. The asymptotic relative efficiency of a statistic against another equals the ratio of their mean square errors. For example, the ARE of R_L^2 versus R_M^2 is

$$ARE\left(R_L^2, R_M^2\right) = \frac{MSE\left(R_M^2\right)}{MSE\left(R_L^2\right)} \tag{6}$$

A value of ARE (R_L^2, R_M^2) greater than 1 indicates that R_L^2 statistic is more efficient than R_M^2, whereas ARE (R_L^2 , R_M^2) less than 1 indicates that R_M^2 statistic is more efficient than R_L^2 .

4. SIMULATION STUDY

Using this simulation study, we demonstrate how misspecification and the distribution of the explanatory variable affect the efficiency of coefficient of determination. We consider misspecified explanatory variable, omission of the other explanatory variable. A logistic regression model, with binary response variable (y), one continuous (x) and one discrete (z) explanatory variable was built. Our aim is to see the effects of distributions and parameters of *x* on ARE. The distribution of x is chosen to be normal(0,1), normal(0,3), normal(3,1) in order to see the effect of the changes of mean and standard deviation, exponential($\lambda = 1$) and exponential($\lambda = 3$) where $\lambda = 1/mean$ in order to reveal the effect of skewness. $\beta_0 = -3$, $\beta_1 = 2$ and $\beta_2 = 2$ are considered as coefficient values. To be consistent with real life, we set the approximate correlations between *x* and *y*, *x* and *z*, *z* and *y* are 0.65, 0.15 and 0.35, respectively. The simulation was conducted using R-programming version of 2.15.

The targeted population consists of *N* = 100,000 units. From this targeted population we randomly draw 10,000 samples with size of *n* = 100, with B = 500 bootstrap replications. The binary response variable *Y* was generated from the Bernoulli distribution with a success probability $p = (1 + \exp(-\theta))^{-1}$. A functional logistic regression model was fitted to the generated data. The models built with the population values without any misspecification are called original model, while the models built with the sample values without any misspecification are denoted by *x*-cont. Misspecifications chosen for this study include categorizing the continuous explanatory variable x into $k = 2$ and $k = 3$ categories and using wrong functional form of x, such as; x^3 and e^x which are denoted by x-cb and exp(x) respectively and omitting the discrete covariate z from the model. For example, if we use *x*-cube instead of *x*, equation 1 will take the form of

$$E(y)=\frac{\exp\left(\beta_0^* + \beta_1^* x^3 + \beta_2^* z\right)}{1+\exp\left(\beta_0^* + \beta_1^* x^3 + \beta_2^* z\right)} \tag{7}$$

For discretizing *x*, the cut points given with Table 1 are selected as mentioned in section 2.1.1

Table 1. Number of categories and location of cut points

Distribution	***k=2***	***k=3***
Norm(0,1)	mean	-0.612 and 0.612
Norm(0,3)	mean	-1.836 and 1.836
Norm(3,1)	mean	2.388 and 3.612
Exp(1)	1.594	1.018 and 2.611
Exp(3)	0.531	0.339 and 0.870

Table 2 shows that the sample of size *n* = 100 is sufficient for this data set. Studies with small sample size employing logistic regression overestimate the effect measure (Nemes and friends, 2009). Therefore, we determined *n* as 100 and reduced the bias effect.

Table 2. Medians of estimations of the logistic regression coefficients for x is continuous

Distribution	**Estimate**
Normal (0,1)	2.082
Normal (0,3)	2.217
Normal (3,1)	2.126
Exp (1)	2.077
Exp (3)	2.019

We report the results of R^2 for original model, and the medians and the variances of R^2 for other models in Table 3. For example, in the third row and the second column, the value 0.377 is the median of R_L^2 obtained from 10,000 simulations when *x*-cb is used instead of *x* for standard normally distributed *x*. The value in parenthesis is the variance of its sampling distribution. The most interesting result presented in the corresponding table is the reduction occurring when *k* = 2 and *z* is omitted. It means that the medians of R^2s are reduced when we categorize *x* into 2 different categories and omit *z* from the model. Besides, if *x* follows an exponential distribution, not only *k* = 2 and *z* is omitted but also *k* = 3 models have reduced values. Especially, we see that with 10% the greatest reduction is in *z* omitted when $\lambda = 3$.

Table 3. The real values of R^2 for original model and the medians and the variances of R^2 for other models

	N(0,1)			N(0,3)			N(3,1)			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	R_L^2	R_M^2	R_N^2
Original model	0.412	0.367	0.547	0.727	0.620	0.843	0.357	0.228	0.442	0.361	0.391	0.523	0.210	0.213	0.313
x-cont	0.436 (0.0090)	0.378 (0.0050)	0.566 (0.0096)	0.751 (0.0057)	0.626 (0.0016)	0.855 (0.0026)	0.391 (0.0122)	0.244 (0.0051)	0.475 (0.0139)	0.380 (0.0065)	0.400 (0.0043)	0.538 (0.0078)	0.228 (0.0063)	0.224 (0.0049)	0.331 (0.0102)
x-cb	0.377 (0.0109)	0.335 (0.0064)	0.503 (0.0130)	0.703 (0.0097)	0.600 (0.0029)	0.820 (0.0051)	0.376 (0.0126)	0.236 (0.0051)	0.459 (0.0145)	0.356 (0.0074)	0.380 (0.0052)	0.511 (0.0094)	0.218 (0.0063)	0.216 (0.0050)	0.319 (0.0105)
exp(x)	0.401 (0.0106)	0.353 (0.0060)	0.528 (0.0121)	0.688 (0.0128)	0.591 (0.0042)	0.807 (0.0074)	0.376 (0.0130)	0.236 (0.0052)	0.458 (0.0149)	0.369 (0.0073)	0.391 (0.0050)	0.526 (0.0090)	0.227 (0.0063)	0.224 (0.0049)	0.330 (0.0103)
k=2	0.329 (0.0077)	0.301 (0.0053)	0.451 (0.0105)	0.556 (0.0080)	0.516 (0.0040)	0.705 (0.0063)	0.234 (0.0035)	0.156 (0.0025)	0.302 (0.0057)	0.304 (0.0060)	0.335 (0.0049)	0.450 (0.0088)	0.207 (0.0058)	0.206 (0.0048)	0.305 (0.0099)
k=3	0.391 (0.0083)	0.347 (0.0051)	0.519 (0.0099)	0.606 (0.0058)	0.548 (0.0020)	0.748 (0.0037)	0.315 (0.0082)	0.203 (0.0042)	0.393 (0.0107)	0.338 (0.0061)	0.366 (0.0044)	0.492 (0.0080)	0.202 (0.0057)	0.202 (0.0047)	0.298 (0.0098)
z omitted	0.327 (0.0078)	0.299 (0.0051)	0.448 (0.0104)	0.702 (0.0058)	0.601 (0.0018)	0.821 (0.0030)	0.355 (0.0125)	0.224 (0.0052)	0.435 (0.0149)	0.261 (0.0052)	0.295 (0.0046)	0.397 (0.0084)	0.092 (0.0031)	0.097 (0.0031)	0.144 (0.0067)

Table 4. ARE's of each R^2 statistics under both correct and misspecified models

		ARE		
		$R_L^{2^*}, R_L^2$	$R_M^{2^*}, R_M^2$	$R_N^{2^*}, R_N^2$
N(0,1)	x-cb	0.83	0.78	0.74
	exp(x)	0.85	0.83	0.79
	k=2	1.17	0.94	0.91
	k=3	1.08	0.98	0.97
	z omitted	1.15	0.98	0.92
N(0,3)	x-cb	0.59	0.55	0.51
	exp(x)	0.45	0.39	0.35
	k=2	0.71	0.41	0.41
	k=3	0.99	0.80	0.70
	z omitted	0.99	0.87	0.87
N(3,1)	x-cb	0.97	0.99	0.96
	exp(x)	0.94	0.97	0.93
	k=2	2.53	1.78	1.97
	k=3	1.48	1.23	1.29
	z omitted	0.98	0.98	0.92
Exp(1)	x-cb	0.89	0.83	0.83
	exp(x)	0.90	0.87	0.87
	k=2	1.09	0.88	0.88
	k=3	1.08	0.97	0.97
	z omitted	1.26	0.93	0.93
Exp(3)	x-cb	0.99	0.98	0.98
	exp(x)	0.99	0.99	0.99
	k=2	1.07	1.03	1.03
	k=3	1.11	1.05	1.05
	z omitted	1.99	1.57	1.51

Table 4 evaluates the ARE's of each R^2 statistics under correct model versus misspecified model. For example the ARE of R_L^{2*} versus R_L^2 is 0.83 meaning R_L^2 is more efficient than its misspecified version when x-cb is used and the loss in efficiency is 17%. For standard normal distributed data, there are about 15% and 20% loss if we use wrong functional form, but if we categorize x or omit z there is no significant difference in efficiency. However, when we increase the variance, remarkable differences occur. Especially, consequences of using wrong functional form of x are more serious. The most effected statistic by misspecification is R_N^2, since there is 65% loss. Categorization causes some loss too but not as much as in case of z omitted. The increase of mean does not cause any undesired result except for wrongly categorizing explanatory variable. Using exponential distribution causes at most 17% efficiency loss and increasing the rate, i.e., decreasing the mean, reduces the efficiency loss except for z omitted scenario. One more remarkable result is that R_M^2 and R_N^2 have generally the same ARE for exponentially distributed x.

Table 5. ARE's of three R^2 statistics with each other

		ARE		
		R_L^2, R_M^2	R_N^2, R_L^2	R_N^2, R_M^2
N(0,1)	Cont.	0.55	0.94	0.52
	x-cb	0.58	0.84	0.49
	exp(x)	0.57	0.88	0.50
	k=2	0.69	0.73	0.51
	k=3	0.61	0.84	0.52
	z omitted	0.65	0.76	0.50
N(0,3)	Cont.	0.28	2.19	0.61
	x-cb	0.31	1.91	0.58
	exp(x)	0.32	1.73	0.56
	k=2	0.50	1.25	0.62
	k=3	0.35	1.56	0.55
	z omitted	0.32	1.90	0.60
N(3,1)	Cont.	0.42	0.88	0.37
	x-cb	0.41	0.87	0.36
	exp(x)	0.41	0.87	0.35
	k=2	0.72	0.62	0.44
	k=3	0.50	0.77	0.39
	z omitted	0.42	0.83	0.35
Exp(1)	Cont.	0.66	0.85	0.56
	x-cb	0.71	0.78	0.56
	exp(x)	0.69	0.81	0.56
	k=2	0.81	0.69	0.56
	k=3	0.73	0.76	0.56
	z omitted	0.89	0.62	0.56
Exp(3)	Cont.	0.78	0.61	0.48
	x-cb	0.79	0.60	0.48
	exp(x)	0.78	0.61	0.48
	k=2	0.81	0.59	0.48
	k=3	0.82	0.58	0.48
	z omitted	0.99	0.47	0.46

Table 5 presents the ARE values of three R^2 statistics with each other. For example, suppose that *x* follows N(0,1), and is discretized into $k = 3$ intervals, then preferring R_M^2 statistic instead of R_L^2 is reasonable since ARE (R_L^2, R_M^2) is 0.61. This will prevent 39% loss in efficiency. If the variance gets larger for normally distributed *x*, the efficiency of R_L^2 reduces substantially relative to both R_M^2 and R_N^2 . R_M^2 is the more efficient one, explicitly. If the mean gets larger, R_M^2 is more efficient than R_L^2

again, but this time with ARE < 1, R_L^2 is more efficient than R_N^2. Finally, if *x* follows exponential distribution, the efficiency of R_N^2 reduces and its asymptotic relative efficiency against R_M^2 is the same regardless of misspecification type except for *z*-omitted.

5. CONCLUSION

In this study, effects of the model misspecification on the various pseudo R^2 statistics have been studied for binary logistic regression model. Asymptotic relative efficiency has been used to compare the performances. Simulations show that, if we have normally distributed explanatory variable, increased variance leads to substantial losses in efficiency, whereas the increased mean does not. If we use a skewed distribution, on the other hand, misspecification does not cause any problem. Under the considered scenarios, we may order three R^2 statistics from the most efficient to the less one as R_M^2, R_L^2, R_N^2. Our final recommendation for researchers would be to select the coefficient of determination associated with the logistic regression analysis, carefully.

REFERENCES

Begg, M.D. and Lagakos, S. (1990). On The Consequences of Model Misspecification in Logistic Regression. *Environmental Health Perspectives,* 87: 69-75.

Begg, M.D. and Lagakos, S. (1992). Effects of Mismodelling on Tests of Association Based on Logistic Regression Models. *The Annals of Statistics,* 20(4): 1929-1952.

Begg, M.D. and 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.

Cox, D.R. (1957). Note on Grouping. *Journal of the American Statistical Association,* 53(280): 543-547.

Cox, D.R. and Snell, E.J. (1989). *The Analysis of Binary Data*. London: Chapman and Hall.

Erees, S. and Demirel, N. (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.

Hu, B., Palta, M., Shao, J. (2006). Properties of R^2 Statistics for Logistic Regression. *Statistics in Medicine,* 25: 1383-1395.

Keele, L.J. (2008). *Semiparametric Regression for The Social Sciences*. Wiley.

Kvalseth, T.O. (1985). Cautionary Note About R^2. *The American Statistician,* 39: 279-285.

Lagakos, S.W. (1988). 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., and Inoue, T. (2007). Tackling The Omitted Variables Problem Without The Strong Assumptions of Proxies. *European Journal of Operational Research*, 178: 819–840.

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, Second Edition*. SAGE.

McFadden, D. (1974). The Measurement of Urban Travel Demand. *Journal of Public Economies,* 3:303-328.

Mittlböck, M. and Schemper, M. (1996). Explained Variation for Logistic Regression. *Statistics in Medicine,* 15: 1987-1997.

Nagelkerke, N.J.D. (1991). A Note on A General Definition of The Coefficient of Determination. *Biometrica*, 78: 691-692.

Nemes, S., Jonasson, J.M., Genell, A. and Steineck, G. (2009). Bias in Odds Ratios By Logistic Regression Modeling and Sample Size. *BMC Medical Research Methodology*, 9(56).

Tosteson, T.D. and Tsiatis, A.A. (1988). The Asymptotic Relative Efficiency of Score Tests in A Generalized Linear Model with Surrogate Covariates. *Biometrica,* 75(3): 507-514.