International Journal of Statistics and Probability; Vol. 4, No. 1; 2015
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

# Estimating Explained Variation of a Latent Scale Dependent Variable Underlying a Binary Indicator of Event Occurrence

Dinesh Sharma[1], Amanda Miller[2], & Caroline Hollingsworth[1]

[1] Department of Mathematics and Statistics, James Madison University, Harrisonburg, VA 22807, USA

[2] Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Correspondence: Dinesh Shatma, Department of Mathematics and Statistics, James Madison University, Harrisonburg, VA 22807, USA. Tel: 1-540-568-6404. E-mail: sharmadr@jmu.edu

Received: August 25, 2014 Accepted: September 15, 2014 Online Published: January 26, 2015

doi:10.5539/ijsp.v4n1p148 URL: http://dx.doi.org/10.5539/ijsp.v4n1p148

## Abstract

The coefficient of determinant, also known as the $R^2$ statistic, is widely used as a measure of the proportion of explained variation in the context of a linear regression model. In many real life events, interests may lie on measuring the proportion of explained variation, $\rho^2$, of a latent scale dependent variable $U$ which follows a multiple regression model. But in practice, $U$ may not be observable and is represented by its binary proxy. In such situations, use of logistic regression analysis is a popular choice. Many analogues to $R^2$ type statistics have been proposed to measure explained variation in the context of logistic regression. McFadden's $R^2$ measure stands out from others because of its intuitive interpretation and its independence on the proportion of success in the sample. It, however, severely underestimates the proportion of explained variation of the underlying linear model. In this research we present a method for estimating the explained variation for the underlying linear model using the McFadden's $R^2$ statistics. When used in a real life dataset, our method estimated $\rho^2$ of the underlying model within an acceptable margin of error.

**Keywords:** logistic regression, measures of explained variation, latent scale dependent variable, multilevel nonlinear model, Chapman-Richards model

## 1. Introduction

Logistic regression modeling is a popular and powerful tool to describe the relationship between a binary outcome variable to several independent variables. Motivation to use the logistic formulation also follows if we consider the dependent variable $Y$ to be a binary proxy for a latent continuous variable $U$, that follows the multiple linear regression model. This formulation of logistic model is explained below.

Many diseases, including several mental and health disorders, are progressive in nature. Health practitioners use some predefined criteria to determine whether a person has a particular disease or some mental/health condition. In many instances, researchers may have information on whether a subject has a particular health condition or not but they may not have access to the actual measurements on the degree of progression of the condition. Under such circumstances it is reasonable to assume the existence of a latent scale dependent variable, which is not observable but is represented by its binary proxy. This situation can be modeled as follows.

Let $U$ be a continuous random variable, such that

$$Y = \begin{cases} 1 & \text{if } U > c, \quad \text{for some } c \in \mathbb{R} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Let $\mathbf{X}' = (X_1, X_2, \cdots, X_p)$ be a vector of $p$ predictors. We may assume that $U$ is related to $\mathbf{X}$ through an ordinary linear model

$$U = \beta_0 + \beta'\mathbf{X} + \varepsilon, \tag{2}$$

where $\varepsilon$ is a random error term such that $\varepsilon \sim N(0, \sigma_\varepsilon)$. The usual coefficient of determinant $\rho^2 = 1 - E\,[\text{Var}(U|\mathbf{x})]\,/\text{Var}(U)$ can be used to measure the extent to which the covariates of interest explain the underlying outcome variable $U$. As mentioned above, measurements on $U$ are not available, and consequently, the answer to the question "How well the predictors $(X_1, X_2, \cdots, X_p)$ explain $U$?" has to be based on the proportion of explained variation obtained from a logistic regression analysis of $Y$ on $\mathbf{X}$. In such situation it is desirable to compute an $R^2$ analog from the logistic model and use it as an estimate of $\rho^2$. There are, however, two main issues that need to be addressed first.

First, unlike the ordinary least square ($OLS$) regression analysis, where the $R^2$ statistic is almost unanimously used as measure of explained variation, there are many $R^2$ analogs suggested for logistic regression models (Mittlböck & Schemper, 1996; Menard, 2000; DeMaris, 2002; Liao & McGee, 2003; Sharma, 2006). Mittlböck & Schemper (1996) reviewed 12 $R^2$ analogs for logistic regression, Menard (2000) six, DeMaris (2002) seven, and Sharma (2006) 14, with some overlap. Other authors have proposed adjusted $R^2$ analogs (see Mittlböck & Schemper, 2002; Liao & McGee, 2003, for example). But there is no clear consensus on the "best" $R^2$ measure for use with logistic models. Second, almost all of the measures of explained variation for the logistic regression analysis severely underestimate the explained variation in the underlying latent scale variable (Hosmer & Lemeshow, 2000), if one exists.

A "good" $R^2$ measure should i) have intuitively reasonable interpretation (interpretability); ii) be numerically consistent with the $R^2$ of an underlying model; and iii) be least dependent of the proportion of successes in the sample (base rate sensitivity) (Sharma, 2006; Menard, 2000). The McFadden's $R^2$ (McFadden, 1974) has clear advantages over others, because of its intuitively reasonable interpretation as a proportional reduction in error measure, parallel to the $R^2$ in linear regression analysis (Menard, 2000) and lowest base rate sensitivity (Menard, 2000; Sharma et al., 2011). The McFadden's $R^2$ measure is defined as

$$R_L^2 = 1 - \frac{log\,(L_M)}{log\,(L_0)}, \tag{3}$$

where, $L_0$ and $L_M$ are the likelihood of the null and full logistic models, respectively. In spite of many of its advantages over other $R^2$ measures, $R_L^2$ can not be directly used as an estimator of $\rho^2$, as it severely underestimates the the parameter of interest (Hosmer & Lemeshow, 2000).

In this paper we propose a computational method for estimating the proportion of explained variation $\rho^2$ for the underlying linear model using $R_L^2$ obtained from the logistic regression analysis. In section 2 we explain the two-level nonlinear model used for estimating $\rho^2$. The simulation study, results of model fit and model validation are discussed in Section 2. An application to a real data is presented in Section 4 and some concluding remarks are given in Section 5.

## 2. Method

Consider $n$ observations on a binary response variable $Y$ as defined in Eq. (1) and a covariate vector $\mathbf{X}' = (X_1, \ldots, X_p)$. The relationship between $Y$ and $X$ is modeled by the logistic model

$$Pr(Y = 1|\mathbf{x}) \equiv \pi(\mathbf{x}) = \frac{e^{\beta_0 + \boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\beta_0 + \boldsymbol{\beta}'\mathbf{x}}}, \tag{4}$$

with the unconditional mean

$$Pr(Y = 1) \equiv \bar{y} = \frac{\sum\limits_{i=1}^{n} y_i}{n}, \tag{5}$$

where $\boldsymbol{\beta}'$ is a vector of $p$ regression parameters. For a logistic model with binary $y$, it can be shown that the mean of conditional probability of success over all possible combinations of the covariate values ($\bar{y}$) equals the probability of success in the population $\bar{\pi}$.

### *2.1 Two Level Nonlinear Model*

We propose a two-level nonlinear model to estimate the explained variation $\rho^2$ of the underlying linear model using $R_L^2$ obtained from logistic regression analysis. Results of a preliminary simulation study suggests a nonlinear relationship between $\rho^2$ and $R_L^2$. In addition, the dependent variable is a measure of explained variation and needs to be constrained in [0,1]. Therefore, we proposed the following Chapman-Richards function for level-I model.

Level-I Model:

$$\rho^2 = \theta_0 \left(1 - \theta_1 e^{(-\theta_2 R_L^2)}\right)^{(1/(1-\theta_3))}, \tag{6}$$

In the above model, $\theta_0$ is the maximum attainable value of $\rho^2$ and hence is set to 1. $\theta_1$ is related to the initial value of the response variable. $\theta_2$ is the parameter governing the rate at which the response variable approaches its potential maximum, and $\theta_3$ affects near which asymptote maximum growth occurs and determines curve shape and the location of the inflection point.

The level-II models assumes $\theta_i,\ i = 1, 2$ and 3 to be some linear functions of the probability of success $\bar{\pi}$ and the sample size $n$.

Level-II Model:

$$\theta_i = \beta_{i0} + \beta_{i1}\bar{\pi} + \beta_{i2}n + \epsilon_i, \quad i = 1, 2, 3 \tag{7}$$

where $\beta_{ij}$ are regression coefficients and $\varepsilon_i$ is the random error term of $i^th$ level-II model.

### *2.2 Parameter Estimation*

Parameter estimation for the proposed model involves the following steps:

Step 1 - Simulating Datasets:

A Monte Carlo study was designed to simulate datasets of various sample sizes from populations with different levels of $\bar{\pi}$. For a Binary dependent variable $Y$ (Eq. 1), representing an unobservable latent scale continuous random variable $U$ (Eq. 2), the probability of success $\bar{\pi}$ is given by

$$\bar{\pi} \equiv Pr(Y = 1) = Pr(U > c) = Pr\left(Z > \frac{c - \mu}{\sigma}\right), \tag{8}$$

where $\mu$ and $\sigma$ respectively are the mean and standard deviation of $U$ and $Z \sim N(0, 1)$. Therefore, $\bar{\pi}$ can be expressed as a function of three key parameters: the cutoff value $c$, and the mean and the standard deviation of $U$ as below.

$$\bar{\pi} = 1 - Pr\left(Z \leq \frac{c - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{c - \mu}{\sigma}\right), \tag{9}$$

where $\Phi$ is the standard normal cumulative distribution function (CDF). It is, therefor, possible to simulate two populations with different proportion of successes by varying any combination of these three parameters. However, in many practical situations the cutoff value is usually held fixed. It is also reasonable to assume that the underlying latent scale variable $U$ has the same mean but different spread in two subgroups of a population. For example, in a study of determinants of diabetes in male and female populations, the same cutoff value of fasting plasma glucose (FPG) level is used for classifying diabetes status for both populations. However, studies have shown that while the mean FPG level among men is usually higher than female, the standard deviation mostly remains the same (for example see Faerch et al., 2010). Therefore, in order to generate datasets with different proportion of successes, we first simulated $U$s with different means but same standard deviations and then generated binary $Y$s using Eq. 1 with fixed $c$.

We manipulated three variable in our simulation: $\rho^2$ of the underlying linear model (Eq. 2), proportion of success ($\bar{\pi}$), and sample size ($n$). We used 19 configurations of $\rho^2$ varying by 0.05 from 0.05 to 0.95, 10 configurations of $\bar{\pi}$ varying by 0.05 from 0.05 to 0.5 and five sample sizes: 50, 100, 250, 500, 1000. Simulation variables were completely crossed creating a total of 950 simulation conditions. Each simulation condition was replicated 10,000 times, resulting a total of 9,500,000 logistic models.

Step 2 - Estimating Level-I Model Parameters

We used PROC NLIN of SAS$^{®}$ to fit level-I model to the simulated data and estimate the parameters. The Marquardt (1963) iterative method was used as it represents a compromise between the linearization (Gauss-Newton) method and the steepest descent method and appears to combine the best features of both while avoiding their most serious limitations. The Marquardt iterative method, however, requires that an initial value for each model parameter be specified first. There are four parameters to be estimated in the level-I model. The methods used to determine the starting values of these parameters are described below.

$\theta_0$ is the maximum possible value of the dependent variable, which in our case is $\rho^2$, and therefore was set to 1. $\theta_2$ parameter is the rate constant at which the response variable approaches its maximum possible value of 1. On the

Table 1. Model fit summery and parameter estimates for Level-2 model. P-values are given in parenthesis with respective parameter estimates.

| Level-1 | Model Fit Summary | | Parameter Estimates | | |
|---|---|---|---|---|---|
| Parameter | $F$-Stat | $R^2$ | $\beta_{0j}$ | $\beta_{1j}$ | $\beta_{2j}$ |
| $\theta_1$ | 2595.8 | 0.8303 | 0.903457 | -0.1274 | 0.00001136 |
| | (0.000) | | (0.000) | (0.000) | (0.016) |
| $\theta_2$ | 1011.94 | 0.6561 | 3.84093 | 0.085962 | 0.000000702 |
| | (0.000) | | (0.000) | (0.000) | (0.901) |
| $\theta_3$ | 6631.79 | 0.9259 | 0.490623 | 0.087287 | -0.000009741 |
| | (0.000) | | (0.000) | (0.000) | (0.000) |

basis of this definition we used the expression $(u_2 - u_1)/(v_2 - v_1)$ to estimate the starting value of $\theta_2$. Here $u_1$ and $u_2$ are values of $\rho^2$ corresponding to some large $R^2_L$ values in the range $(v_1, v_2)$. For the classical Chapman-Richards model $\theta_3$ is between zero and one ($0 < \theta_3 < 1$). $\theta_1$ depends on the initial value of the response variable, $\rho^2$, and can be thought as the "intercept" on Y-axis for $R^2_L$=0. Its starting value can be specified by evaluating $\rho^2$ when $R^2_L = 0$. From equation (6) we get $\rho^2(0) = (1 - \theta_1)^{(1/1-\theta_3)}$, where $\rho^2(0)$ is ideally zero, but one should choose a relatively small positive number close to zero.

Step 3 - Estimating Level-II Model Parameters

Estimates of $\theta_i$'s obtained in step 2) are regressed on corresponding sample sizes (n) and probability of successes ($\bar{\pi}$) to obtain estimates of $\beta_{ij}$, $i = 1, 2, 3,\ j = 0, 1, 2$ for the level-II models in Eq. 7.

## 3. Results and Model Validation

Scatter plots for level-II model suggest a nonlinear effect of $\bar{\pi}$ on $\theta_i$. An inverse square-root transformation of $\bar{\pi}$ appeared to address the problem of nonlinearity. Accordingly, we fit the following system of linear equations to obtain the least square estimates of the level-II model parameters.

$$\theta_i = \beta_{i0} + \beta_{i1}(\bar{\pi})^{-1/2} + \beta_{i2}n + \epsilon_i, \quad i = 1, 2, 3 \tag{10}$$

Model fit summary statistics and estimated level-II model parameters along with the respective p-values are presented in Table 1. The relationship between level-I model parameters and the proportion of success, $\bar{\pi}$, is statistically significant ($p < 0.000$ for all $\beta_i$, $i = 1, 2, 3$). Though the coefficients for $n$ are very small, they are statistically significant for estimating $\theta_1$ and $\theta_3$. The results presented in Table 1 clearly indicate that the proportion of success in a data set and the sample size are good predictors of the level-I model parameters which are used to estimate $\rho^2$ of the underlying linear model.

In order to validate our model, we simulated a validation dataset using the same 19 levels of $\rho^2$ ranging from .05 to .95 and three sample sizes, n: 50, 100 and 500 and four levels of $\bar{\pi}$: 0.05, 0.2, 0.35 and 0.5. Use of a sample size smaller than 50 (e.g. n=30) caused numerical problems including no variation in the dependent variable, complete severation and quasi complete severation. These numerical problems were more frequent for low values of $\bar{\pi}$, especially when $\bar{\pi} = 0.05$. Simulation is implemented using the statistical software $SAS^{\copyright}$ 8.1. Proc logit is used to fit the logistic models. The simulation algorithm is outlined below:

For each level combination of $\rho^2$, and $n$

1. Simulate the underlying linear model $U = \beta_0 + \beta'\mathbf{X} + \varepsilon$ by generating $X \sim N(\mu_x, \sigma^2_x)$ and $\varepsilon \sim N(0, \sigma^2_\varepsilon)$.
2. Generate $U$ such that the coefficient of determination for the linear model is $\rho^2$.
3. Generate the binary dependent variable $Y$ such the proportion of success in the dataset is $\bar{\pi}$ ($\bar{\pi}$ = 0.05, 0.2, 0.35 and 0.5).
4. For each dataset thus generated fit a logistic model and then compute $R^2_L$
5. For each combination of $\bar{\pi}$ and $n$, estimates of $\theta_i$'s using equation (10) and estimates of $\beta_{ij}$'s using the regression coefficients from Table 1.

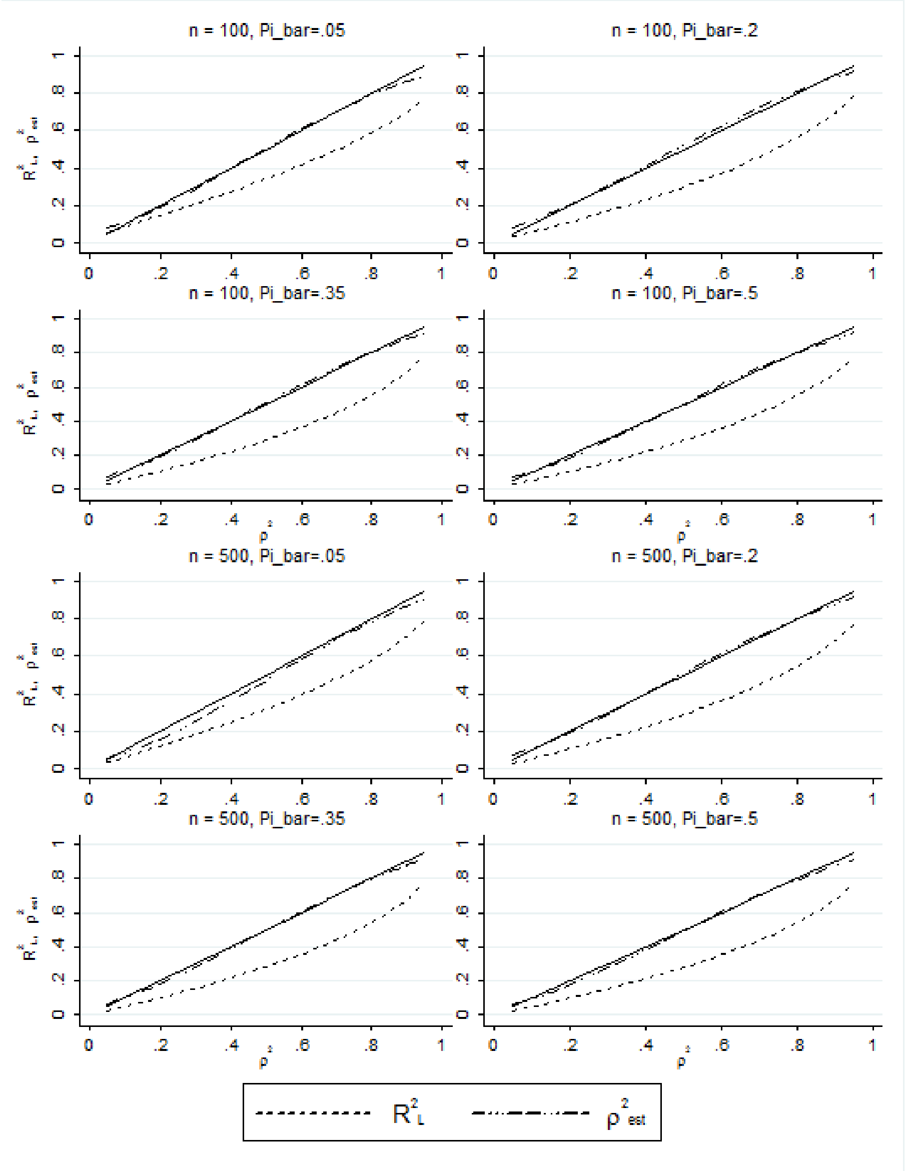

Figure 1. $R_L^2$ and $\hat{\rho}^2$ plotted against $\rho^2$. Departure from the 45 ° line indicates bias.

6. Obtain $\hat{\rho}^2$ using Eq. 6.

7. Repeat steps 1) to 3) 1000 times and calculate mean and SE of $\hat{\rho}^2$.

The graphs in Fig. 1 compare $\hat{\rho}^2$ and $R_L^2$ as estimators of $\rho^2$, proportion explained variation of the underlying linear model, for selected simulation conditions. Graphs for $n = 50$ so the similar patterns and are not presented here. The 45 ° angle solid line was obtained by plotting $\rho^2$ against itself. The distance of a point from this ideal 45 ° angle line indicates how well or how poorly the prediction performed. As can be seen in Fig. 1, our model clearly out performs $R_L^2$ in estimating $\rho^2$ for all eight simulation conditions.

In order to evaluate the quality of our estimate we computed relative root mean square error (RRMSE) of $\hat{\rho}^2$. RRMSE is a relative measure of prediction accuracy and is calculate as

$$\text{RRMSE} = \sqrt{\frac{\sum_{i=1}^{R}(\hat{\theta}_i - \theta)^2}{\theta^2}}, \tag{11}$$

Table 2. Relative Root Mean Square Errors (RRMSE) of $\hat{\rho}^2$ for selected sample sizes and levels of $\bar{\pi}$. RRMSE values are presented as percentage points.

| | $n$ = 50 | | | | $n$ = 100 | | | | $n$ = 500 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho^2$ | $\bar{\pi}$=0.05 | $\bar{\pi}$=0.2 | $\bar{\pi}$=0.35 | $\bar{\pi}$=0.5 | $\bar{\pi}$=0.05 | $\bar{\pi}$=0.2 | $\bar{\pi}$=0.35 | $\bar{\pi}$=0.5 | $\bar{\pi}$=0.05 | $\bar{\pi}$=0.2 | $\bar{\pi}$=0.35 | $\bar{\pi}$=0.5 |
| 0.05 | 13.422 | 13.059 | 13.533 | 13.057 | 12.102 | 11.945 | 10.533 | 9.385 | 10.140 | 10.140 | 10.055 | 7.995 |
| 0.10 | 4.571 | 4.530 | 2.501 | 3.596 | 3.745 | 3.745 | 0.828 | 4.129 | 4.113 | 3.053 | 1.247 | 1.248 |
| 0.15 | 2.103 | 2.100 | 1.500 | 1.490 | 1.126 | 1.133 | 1.149 | 1.126 | 1.126 | 1.126 | 1.149 | 0.580 |
| 0.20 | 1.571 | 1.339 | 1.325 | 1.127 | 1.127 | 1.127 | 1.150 | 1.150 | 1.127 | 1.107 | 1.127 | 0.750 |
| 0.25 | 1.601 | 1.081 | 1.105 | 1.151 | 1.128 | 1.111 | 1.105 | 1.151 | 1.128 | 1.128 | 1.105 | 0.800 |
| 0.30 | 1.400 | 1.049 | 1.284 | 1.526 | 1.500 | 1.511 | 1.470 | 1.470 | 1.500 | 1.518 | 1.500 | 0.870 |
| 0.35 | 0.891 | 0.828 | 1.154 | 1.144 | 1.126 | 1.308 | 1.720 | 1.694 | 1.126 | 1.449 | 1.804 | 1.804 |
| 0.40 | 0.836 | 0.842 | 0.468 | 1.127 | 1.127 | 0.778 | 1.071 | 1.127 | 1.127 | 0.903 | 1.232 | 1.250 |
| 0.45 | 1.265 | 1.265 | 0.242 | 1.128 | 1.128 | 0.567 | 0.665 | 1.115 | 1.128 | 0.470 | 0.738 | 0.738 |
| 0.50 | 1.570 | 1.570 | 0.489 | 1.496 | 1.500 | 0.780 | 0.159 | 1.503 | 1.500 | 0.594 | 0.367 | 0.364 |
| 0.55 | 1.476 | 1.476 | 0.735 | 1.065 | 1.065 | 0.935 | 0.254 | 1.161 | 1.160 | 0.796 | 0.148 | 0.053 |
| 0.60 | 1.336 | 1.336 | 0.697 | 0.608 | 0.940 | 0.948 | 0.431 | 0.789 | 0.828 | 0.812 | 0.313 | 0.138 |
| 0.65 | 1.239 | 1.239 | 0.763 | 0.569 | 0.825 | 0.833 | 0.403 | 0.484 | 0.760 | 0.760 | 0.333 | 0.166 |
| 0.70 | 1.143 | 1.137 | 0.580 | 0.541 | 0.732 | 0.736 | 0.366 | 0.417 | 0.583 | 0.591 | 0.291 | 0.156 |
| 0.75 | 0.822 | 0.822 | 0.484 | 0.320 | 0.477 | 0.477 | 0.174 | 0.395 | 0.395 | 0.370 | 0.094 | 0.032 |
| 0.80 | 0.445 | 0.445 | 0.145 | 0.416 | 0.417 | 0.149 | 0.181 | 0.549 | 0.553 | 0.120 | 0.200 | 0.198 |
| 0.85 | 0.770 | 0.251 | 0.317 | 0.710 | 0.704 | 0.397 | 0.448 | 0.783 | 0.774 | 0.444 | 0.473 | 0.482 |
| 0.90 | 1.271 | 0.617 | 0.699 | 0.992 | 0.977 | 0.763 | 0.832 | 1.019 | 1.036 | 0.775 | 0.811 | 0.826 |
| 0.95 | 1.770 | 1.083 | 1.116 | 1.319 | 1.341 | 1.143 | 1.155 | 1.327 | 1.327 | 1.164 | 1.167 | 1.167 |

where $\theta$ and $\hat{\theta}$ are respectively the desired and the estimated value of the parameter of interest and $R$ represents the number of simulation. The RRMSE has a minimum value of 0.0 for a perfect prediction. Values closer to 0.0 indicate better prediction. RRMSEs of the estimates for all twelve simulation conditions are presented in Table 2. Except for very small value of $\rho^2$, the RRMSEs are acceptably small (less than 5%). When $\rho^2 = .05$ the RRMSEs range from 8% to 13.5%. However, it should be noted that a model with $\rho^2 = .05$ is not very useful and thus may not be used in practice.

## 4. Application to a Real Life Dataset

The dataset used in our example comes from Exam 3 of Framingham Offspring Study (Feinleib et al., 1975). We used Exam 3 data mainly because information about fasting blood glucose (FBG), which was used as the unobserved latent scale variable in our model, was collected starting at this point of the offspring study. The dataset consists of 3371 men and women who were not taking any diabetes medicine at the time of the exam and were not previously identified as diabetic. For the purpose of this example we selected five potential predictors of FBG. They were gender (SEX), age at the time of Exam 3 (AGE3), hypertension (HYP: 1 if hypertensive, 0 otherwise), smoking (SMOKE: 1 if currently smoking, 0 otherwise) and body mass index (BMI). In the standard model formulae syntax, our model is

$$FBG = CONST + AGE3 + BMI + SEX + HYP + SMOKE \tag{12}$$

A multiple linear regression analysis of the above model resulted an $R^2 = 0.1861$ with all of the predictors being statistically significant. A 95% bootstrap confidence interval, based on 1000 bootstrap samples, for $\rho^2$ was (0.1629, 0.2092).

According to the American Diabetes Association criteria, a person is classified as having impaired fasting glucose (IFG), a type of prediabetes, if the FBG level is between 100 mg/dL and 125 mg/dL, inclusive. We used this criteria to create a binary variable IFG, a proxy of the continuous dependent variable FBG to be explained by the above mentioned predictors, such that IFG=1 if $100 \leq FBG \leq 125$ and IFG=0 if $FBG < 100$. In our dataset 17.83% of the subjects were identified with IFG (i.e. $\bar{\pi} = 0.1783$). A logistic regression analysis between IFG and the five predictors resulted a model with $R_L^2 = 0.1080$, which, as expected, is considerably smaller than $R^2$, the proportion of variation in FBG explained by the underlying linear model (12). The predicted value of $\rho^2$, using our proposed method is 0.16403, which is well within the 95% bootstrap confidence interval of $\rho^2$.

## 5. Conclusion

Researchers often are interested in estimating how well a set of predictors explains the outcome of a dependent continuous variable. If the relationship is modeled using a linear regression model then the coefficient of determinant can be used to estimate the proportion of variation in the dependent variable explained by the predictors. But in practice, the dependent variable of interest may not be observable and is represented by its binary proxy. In such situation, interests may lies on estimating the proportion of explained variation by use of a logistic regression analysis. In this paper, we have proposed a computational method for this purpose. We used McFadden's $R^2$ measure mainly because of its intuitive interpretation and base rate invariant property. In addition, it is easy to compute using standard logistic regression output of most of the statistical analysis softwares. When applied to a real life dataset, our method estimated the proportion of explained variation of the underlying model within an acceptable margin of error.

**Acknowledgements**

Part of this research is funded through the *Jeffrey E. Tickle '90 Family Endowment in Science & Mathematics*, College of Sciences and Mathematics, James Madison University. The authors would like to extend their sincere thanks to the Biologic Specimen and Data Repository Information Coordinating Center of National Institute of Health for providing access to the Framingham Offspring Study dataset for use.

## References

DeMaris, A. (2002). Explained variance in logistic regression. A Monte Carlo study of proposed measures. *Sociological Methods and Research, 31*, 27-74.

Faerch, K., Borch-Johnsen, K., Vaag, A., Jorgensen, T., & Witte, D. R. (2010). Sex differences in glucose levels: a consequence of physiology or methodological convenience? The Inter99 study. *Diabetologia, 53*, 858-865.

Feinleib, M., Kannel, W., Garrison, R., McNamara, P., & Castelli, W. (1975). The Framingham Offspring Study. Design and preliminary data. *Preventive Medicine, 4*, 518-525.

Hosmer, D. & Lemeshow, S. (2000). *Applied Regression Analysis (2nd ed)*. New York, NY: Wiley.

Liao, J. & McGee, D. (2003). Adjusted coefficient of determination for logistic regression. *The American Statistician, 73*, 161-165.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P. (Ed.), *Frontiers in Econometrics*, 105-142. Academic Press.

Menard, S. (2000). Coefficient of determination for multiple logistic regression analysis. *The American Statistician, 54*, 17-24.

Mittlbck, M. & Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine, 15*, 1987-1997.

Mittlbck, M. & Schemper, M. (2002). Explained variation for logistic regression - small sample adjustments, confidence intervals and predictive precision. *Biometrical Journal, 44*, 263-272.

Sharma, D. (2006). *Logistic Regression, Measures of Explained Variation and the Base-rate Problem*. PhD thesis, Florida State University.

Sharma, D., McGee, D., & Kibria, B. (2011). Measures of explained variation and the base-rate problem for logistic regression. *American Journal of Biostatistics, 2*, 11-19.

## Copyrights