أسلوب البوستراب لقاييس شبه معامل التحديد لنموذج متغير الاستجابة الثنائي

المدرس زكريا يحيى الجمال

قسم الإحصاء والمعلوماتية

كلية علوم الحاسبات والرياضيات/ جامعة الموصل

المستخلص

الاستدلال الإحصائي عادة ما يكون مبني على بعض المقدرات والتي بدورها تكون دوال للبيانات لذا فان أسلوب البوستراب يوفر تقدير أو تقريب للتقدير لتوزيع المعاينة للمقدر الإحصائي. يعتبر نموذج الانحدار اللوجستي من النماذج الكثيرة الاستخدام في مجالات العلوم التطبيقية. في بحثنا هذا تم التركيز على مقاييس شبه معامل التحديد في نموذج الانحدار اللوجستي من خلال دراسة أسلوب البوستراب عن طريق المحاكاة وبيانات حقيقية حيث استنتجنا باستخدام شبه معامل التحديد R^2_M أو شبه معامل التحديد R^2_D بسبب تقارب قيمهما إلى حد كبير.

تاريخ استلام البحث ٢٠١٢/٣/٥ تاريخ قبول النشر ٢٠١٢/٦/٢٤

1- Introduction

One of the important methods in statistics is that the regressing a binary response variable on a set of explanatory variables. Binary response variable has two values, typically coded 0 for the event did not occur and 1 for the event did occur [12]. For the standard linear regression model the familiar coefficient of determination, $\mathbf{R^2}$, is a widely used goodness of fit measure. The term bootstrap which is due to the Efron [5] is an illusion to the expression "pulling on self up by one's bootstraps" meaning doing the impossible [6]. The bootstrap is a method to derive properties like standard error, confidence intervals, of the sampling distribution of estimators. The bootstrap resampling consists of $\boldsymbol{n}$ elements that are drawn randomly from the $\boldsymbol{n}$ original data points with replacement [7]. This paper focus on the behavior of bootstrapping pseudo-$\mathbf{R^2}$measures. Simulation and real data results also presented. The contents of this paper may be divided into eight sections. In section 2 and 3 we review the binary response variable model and pseudo measures, respectively. In section 4 we introduce bootstrapping pseudo $\mathbf{R^2}$. The simulation results, real data results, conclusions and references are given in sections 5, 6, 7, and 8, respectively.

2- Binary Response Variable Model

Binary response is commonly studies in medical and epidemiologic research, for instance, the presence or absence of a particular disease, death during surgery. Models for mutually exclusive binary outcomes focus on the determinates of the probability $\mathbf{p}$ of the occurrence of one outcome rather than an alternative outcome that occurs with a probability of $\mathbf{1-p}$. In regression analysis we want to measure how the probability

$\mathbf{p}$ varies across individuals as a function of explanatory variables. Binary response variable has two values, typically coded 0 for the event did not occur and 1 for the event did occur [12]. The expected value of a binary variable is the probability that it takes the value 1.

Let $\mathbf{p(y_i=1)=\pi_i}$ and $\mathbf{p(y_i=0)=1-\pi_i}$, then

$$\mathbf{E(y_i)=0*(1-\pi_i)+1*\pi_i=p(y_i=1)=\pi_i} \qquad \ldots(1)$$

With explanatory variables, x_i's,

$$\mathbf{E(y_i|x_i)=\pi_i=p(y_i=1|x_1,x_2,.....,x_k)=\beta_0+\beta_1x_1+.....+\beta_kx_k} \qquad \ldots(2)$$

Since π_i is a probability, it must be between 0 and 1. The linear function given in (2) is not an adequate statistical model since the equation (2) can lie outside $[0, 1]$ and dose not represent a probability [3]. Binary response models directly describe the response probability that the response variable takes value 1 is modeled as

$$\mathbf{E(y_i|x_i) = p(y_i = 1|x_1, x_2, ..., x_k) = F(x_i'\beta)} \quad \ldots(3)$$

A binary response model is referred to as a probit model if $\mathbf{F(\cdot)}$ is the cumulative normal distribution function. It is called logit model if $\mathbf{F(\cdot)}$ is the cumulative logistic distribution function. The estimation problem is to estimate the unknown parameters $\boldsymbol{\beta}$. In practice, the ordinary least squares predictions of the conditional probability can be greater than one or less than zero [9]. The probit model is:

$$\mathbf{E(y_i|x_i) = \phi(x_i'\beta) = \int_{-\infty}^{x_i'\beta} f(t)dt = \int_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}t^2} dt} \quad \ldots(4)$$

and the logit model is

$$\mathbf{E(y_i|x_i) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}} \quad \ldots(5)$$

The probit and logit models are typically estimated by maximum likelihood (ML) method. Assuming independence across observations, the likelihood function is:

$$\mathbf{L = \prod_{i|y_i=0}^{n} p(y_i = 0|x_i) \prod_{i|y_i=1}^{n} p(y_i = 1|x_i)} \quad \ldots(6)$$

$$\mathbf{= \prod_{i=1}^{n} [1 - F(w_i)]^{1-y_i} F(w_i)^{y_i}} \quad \ldots(7)$$

where $\mathbf{p(y_i = 1|x_i) = F(w_i) = \phi(w_i)}$ in the probit model and $\mathbf{p(y_i = 1|x_i) = F(w_i) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}}$ in the logit model. The corresponding Log likelihood function is:

$$\log L = \sum_{i=1}^{n}\left[y_i \log F(w_i) + (1-y_i)\log(1-F(w_i))\right] \quad \ldots(8)$$

The first derivative of (8) is

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n}\left[y_i \frac{f(w_i)}{F(w_i)} + (1-y_i)\frac{-f(w_i)}{F(w_i)} \right] x_i' = 0 \quad \ldots(9)$$

Solving (9) using an iterative method one can get the maximum likelihood estimation of β [8].

3- Pseudo-R^2 Measures

For the standard linear regression model the familiar coefficient of determination, R^2, is a widely used goodness of fit measure. Application of this measure to binary response variable model such as logit and probit has no universal definition. A number of measures can be proposed. Pseudo-R^2 usually have the property that, on specialization to the linear model, the coincide with an interpretation of the linear model R^2 [1]. Many different Pseudo-R^2 measures have been proposed in the past four decades [11].

3-1 McFadden's Pseudo-R^2

McFadden's [13] defines the pseudo-R^2 based on the maximum log likelihood, it is:

$$R^2_M = 1 - \left[\frac{\log L(\text{Full model})}{\log L(\text{Null model})} \right] \quad \ldots(10)$$

where the Full model is the model with all variables in the model, whereas the null model is the model with intercept only. Theoretically the range of this coefficient is between 0 and 1.

3-2 Cragg and Uhler Pseudo-R^2

Cragg and Uhler [4] introduced a normal version of the transformation of the likelihood ratio, it is defined as:

$$R^2_{CU} = \left[\frac{\exp(\log L(\text{Full model}))^{2/n} - \exp(\log L(\text{Null model}))^{2/n}}{1 - \exp(\log L(\text{Null model}))^{2/n}} \right] \quad ..(11)$$

3-3 Deviance Pseudo-R^2

Mittlbick and Heinzl [14] proposed pseudo R^2measure for generalized linear models based on the concept of deviance. This measure, R^2_D, is defined as:

$$R^2_D = 1 - \frac{D(\text{Full model})}{D(\text{Null model})} \quad \ldots(12)$$

Where D(Full model) is the deviance of the full model, and D(Null model) is the deviance of the null model. It cannot become negative and increases monotonically with increasing number of explanatory variables.

4- Bootstrap Pseudo-R^2

The term bootstrap which is due to the Efron [5] is an illusion to the expression "pulling on self up by one's bootstraps" meaning doing the impossible [6]. The bootstrap is a method to derive properties like standard error, confidence intervals, of the sampling distribution of estimators. The bootstrap resampling consists of $\mathbf{n}$ elements that are drawn randomly from the n original data points with replacement [7]. In the term of regression analysis, we have two kind of bootstrapping, residual bootstrapping and paired bootstrapping. Consider a sample with $\mathbf{n}$ independent observations of the response variable $\mathbf{y}$ and $\mathbf{k+1}$ explanatory variables $\mathbf{x}$. A paired bootstrap sample is obtained by independently drawing rows with replacement from the pairs $(\mathbf{y_i}, \mathbf{x_i})$. The bootstrap sample has the same number of observations, however some observations appear several time and others never. The bootstrap involves drawing a large number $\mathbf{B}$ of bootstrap samples. An individual bootstrap sample is denoted $(\mathbf{y_b^*}, \mathbf{x_b^*})$ [2]

$$R^{2(boot)}{}_{pseudo} = \frac{\sum_{b=1}^{B} R^{2(b)}{}_{pseudo}}{B} \quad \ldots(13)$$

And

$$bias = R^{2(boot)}{}_{pseudo} - R^{2}{}_{pseudo} \quad \ldots(14)$$

5- Simulation Results

In this section, we examine by simulation the performance of the bootstrap procedure based on varies values of B and sample sizes of n of the three pseudo measures in section 3. Three cases of simulation were done; the first set the value of $\beta_0, \beta_1,$ and β_2 as 2, 0.5, and 0.5, respectively. The second case the parameters' values were 2, 1, and 1 respectively. While the third case parameter values were 2, 2, and 2 respectively. The number of bootstrap samples, B, set to be 1000, 10000, and 100000 respectively. We simulate different sample sizes (10, 25, 50, and 100) from uniform distribution with (-1,1) as explanatory variables and simulate the response variable according to logit model. Tables 1, 2, 3, and 4 show the results.

Table (1): Pseudo-R^2 when n = 10

B=1000 β_0 β_1 β_2	R^2_M (R^2_M (boot))	R^2_D (R^2_D (boot)	R^2_{CU} (R^2_{CU} (boot))
2 0.5 0.5	0.1417(0.521)	0.1054(0.505)	0.1417(0.571)
1 1	0.045 (0.464)	0.0451(0.417)	0.026 (0.4029)
2 2	0.037 (0.181)	0.038 (0.523)	0.0272(0.1761)
B=10000 β_0 β_1 β_2			
2 0.5 0.5	0.1417 (0.209)	0.1054(0.185)	0.1417 (0.21)
1 1	0.045 (0.426)	0.0451(0.43)	0.0268 (0.39)
2 2	0.037 (0.203)	0.038 (0.193)	0.0272 (0.155)

B=100000			
β_0 β_1 β_2			
2 0.5 0.5	0.1417 (0.21)	0.1054(0.181)	0.1417 (0.209)
1 1	0.045 (0.425)	0.0451(0.415)	0.0268 (0.394)
2 2	0.037 (0.197)	0.038 (0.228)	0.0272 (0.167)

Table (2): Pseudo-R^2 when $n = 25$

B=1000	R^2_M (R^2_M (boot))	R^2_D (R^2_D (boot)	R^2_{CU} (R^2_{CU} (boot))
β_0 β_1 β_2			
2 0.5 0.5	0.664 (0.843)	0.665(0.85)	0.548 (0.794)
1 1	0.822 (0.922)	0.821(0.901)	0.742 (0.903)
2 2	0.253 (0.393)	0.254 (0.464)	0.168 (0.297)
B=10000			
β_0 β_1 β_2			
2 0.5 0.5	0.664 (0.836)	0.665(0.791)	0.548 (0.803)
1 1	0.822 (0.931)	0.821(0.81)	0.742 (0.85)
2 2	0.253 (0.373)	0.254 (0.45)	0.168 (0.301)
B=100000			
β_0 β_1 β_2			
2 0.5 0.5	0.664 (0.838)	0.665(0.82)	0.548 (0.771)
1 1	0.822 (0.925)	0.821(0.85)	0.742 (0.82)
2 2	0.253 (0.351)	0.254 (0.395)	0.168 (0.293)

Table (3): Pseudo-R^2 when $n = 50$

B=1000 β_0 β_1 β_2	R^2_M (R^2_M (boot))	R^2_D (R^2_D (boot)	R^2_{CU} (R^2_{CU} (boot))
2 0.5 0.5	0.0732 (0.14)	0.073(0.141)	0.0489 (0.157)
1 1	0. 22 (0.258)	0.221(0.263)	0.1316 (0.1677)
2 2	0.592 (0.642)	0.593 (0.641)	0.458 (0.524)
B=10000 β_0 β_1 β_2			
2 0.5 0.5	0.0732 (0.143)	0.073(0.132)	0.0489 (0.155)
1 1	0. 22 (0.256)	0.221(0.26)	0.1316 (0.162)
2 2	0.592 (0.632)	0.593 (0.634)	0.458 (0.519)
B=100000 β_0 β_1 β_2			
2 0.5 0.5	0.0732 (0.135)	0.073(0.13)	0.0489 (0.143)
1 1	0. 22 (0.24)	0.221(0.252)	0.1316 (0.159)
2 2	0.592 (0.613)	0.593 (0.62)	0.458 (0.48)

Table (4): Pseudo-R^2 when $n = 100$

B=1000 β_0 β_1 β_2	R^2_M (R^2_M (boot))	R^2_D (R^2_D (boot)	R^2_{CU} (R^2_{CU} (boot))
2 0.5 0.5	0.0129 (0.0425)	0.013(0.045)	0.009 (0.0307)
1 1	0. 15 (0.173)	0.151(0.177)	0.1038 (0.1258)
2 2	0.375 (0.397)	0.3751 (0.39)	0.248 (0.27)
B=10000			

β_0 β_1 β_2			
2 0.5 0.5	0.0129 (0.0422)	0.013(0.042)	0.009 (0.0302)
1 1	0. 15 (0.171)	0.151(0.175)	0.1038 (0.1248)
2 2	0.375 (0.389)	0.3751 (0.39)	0.248 (0.255)
B=100000			
β_0 β_1 β_2			
2 0.5 0.5	0.0129 (0.041)	0.013(0.041)	0.009 (0.0299)
1 1	0. 15 (0.17)	0.151(0.171)	0.1038 (0.122)
2 2	0.375 (0.385)	0.3751 (0.38)	0.248 (0.25)

From tables 1,2,3, and 4 we observe that the value of the pseudo R^2_M **,** R^2_D**, and** R^2_{CU} **be larger than the original pseudo values and we see the convergency of the pseudo** R^2_M **and** R^2_D **values.**

6- Real Data Results

We will use a sample of 30 person from Ibn-Alatheer hospital. The response variable be binary with (1 if the person diagnostic to has Thalassemia and 0 if not). Nine variables have been studied as an explanatory variables, they are, sex (1 for male and 0 for female), age, body mass index, HB, PCV, ferritin, IL-6, TNF, and uric acid. The results shown in table 5.

Table (5): Pseudo-R^2 for the real data.

	R^2_M (R^2_M (boot))	R^2_D (R^2_D (boot)	R^2_{CU} (R^2_{CU} (boot))
B=1000	0.584 (0.578)	0.582 (0.614)	0.448 (0. 581)
B=10000	0.584 (0.611)	0.582 (0.5725)	0.448 (0.484)
B=100000	0.584 (0.602)	0.582 (0.5713)	0.448 (0.450)

7- Conclusions

1- In this paper, we have shown from table (1) when the sample size is 10 and for all values of B, the values of bootstrapped pseudo measures greatly difference from the original values of the pseudo measures, that is the bias is so large. This bias gradually be small when the sample size change from 25 to 50 and to 100.

2- From tables (1-5) we conclude and suggest to use either R^2_M or R^2_D, since they have convergence in there values.

3- We recommended that the bootstrap procedures may be not good to verify the asymptotic normal theory since we will get constant bootstrap samples contain only 1 or 0. But asymptotically of normal theory be met when the sample size is more than 50, the results for the real data in table (5) support our conclusion.

8-Refrences

1- Cameron, A., C., and Trivedi, P., K., (1998) " Regression Analysis of Count Data", Cambridge University Press, New York.

2- Carroll, R., J., Ruppert, D., Stefanski, L., A., and Crainiceanu, C., M.,(2006),"Measurement Error in Nonlinear Models, A Modern Prespective", 2nd ed., Chapman & Hall/CRC, Florida

3- Chatterjee, S., and Hadi, A., S. (2006), "Regression Analysis by Example", 4th ed., John Wiley and Sons, Inc., USA.

4- Cragg, J., G. and Uhler, R., (1970), "The Demand for Automobiles", Canadian Journal of Economics, Vol.3, pp.386-406.

5- Efron, B. (1979) "Bootstrap Methods: Another look at Jackknife", Annals of Statistics,Vol.7, pp.1-26.

6- Efron, B. and Tibshirani, R., (1993), "An introduction to the bootstrap", Chapman and Hall, New York.

7- Friedl, H. and Stampfer, E.,(2002), "Jackknife Resampling", Encyclopedia of Environmetrics, 2, pp.1089-1098.

8- Harrell, F., E., (2001) "Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression and Survival Analysis", Springer-Velag, Inc., USA

9- Horowitz, J., L., and Savin, N., E. (2001), "Binary Response Models: Logits, Probits and Semiparametrics", Journal of Economic Perspectives, Vol.15, No.4, pp.43-56.

10- Hosmer,D., W. and Lemeshow, S., (2000), "Applied Logistic Regression" 2nd ed., John Wiley & Sons, Inc., New York.

11- Hu, B., Shao, J. and Palta, M. (2006), "Pseudo-R^2 in Logistic Regression Model", Statistica Sinica, Vol.16, pp.847-860.

12- Long, S., J. (1997) "Regression Models for Categorical and Limited Dependent Variables", SAGE Publication, USA.

13- McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka(ed.), Frontiers in Econometrics, New York: Academic Press.

14- Mittlebick, M., and Heinzl, H., (2003), " Pseudo-R^2 Measures for Generalized Linear Models", 1st European Workshop on the Assessment of Diagnostic Performance, pp.71-80.