

Comparison between Robust and Classical Analysis in Bivariate Logistic for Medical Data

Fadhil Abdul AA*

Department of Statistics, Al-Furat Al-Awast Technical University, Iraq

Abstract

Representing medical data and biological important part in experiments are concerned with Human life, the primary objective of this research is to use the statistical optimization method analysis for the data and knowledge of the important factors affecting the variables of the study (liver fat, liver size), where the variables are interconnected there is a need for statistical method to examines the degree of their relationship, we used bivariate logistic.

To achieve the of the research on the field study will be done in Al-Sadr medical city in the province of Najaf by taking a sample of 150 people auditors diabetes and liver disease center, from the statistical analysis results we observed the degree of diagnosis model in both method are good, and also we monitored that impact factors in responses (liver fat, liver size) and some comment as multivariate logistic in the Future.

Keywords: Bivariate logistic; Robust analysis; Logistic regression; Binary regression

Aim of Study

This research(study) aims to review the method of bivariate logistic distribution in order to study and analysis effecting factors on the response variables (the degree of liver fatty and increase of liver size in Human beings) using the data of medical tests to compare classical and robust analysis when some values are outlier in the sample.

Introduction

In the medical and biological field studies, the experiments are often related to the nature of the response adopted for non- continuous variable data (variables), but is the occurrence/non-occurrence of the score after taking a certain treatment (having effect or non-effect), regardless of the nature of the variables of the study at the left side of the General model equation $y = x\beta + \epsilon$, whether continuous, discrete or categorical.

Where the method of analysis depending on the type of data at the right side (y). If the binary response (0,1 depends on binary logit the method while ordinal response if rank (1st, 2nd, 3rd, ...), for example, the degree of healing of an illness or the degree of incidence of a particular disease is formulated into (generalized) ordinal logit/probit regressions, but if the response takes symbolic letters (A, B,) can rely on logit (probit) multinomial

From a historical of view Barry W [1] studied binary data analysis with bivariate response under the influence of some independent variables and assuming a correlation between paired observations a disturbance depending on the Logistic regression model to estimate parameters the results were obtained are efficient compared with MLE to other researchers.

Kimberlee G, et al. [2] studied the analysis of correlated binary outcomes using Multivariate Logistic Regression for the case of two outcomes, a form of the cumulative bivariate logistic distribution proposed by Gumbel is used to characterize their joint probabilities in terms of logistic marginal probabilities and the correlation coefficient of the responses. They applied this technique in two different situations. When the correlation among responses is not significantly/and is significantly from zero.

In 1997 Sean M., David B [3] studied Bayesian analyses of multivariate binary (categorical) outcomes for $\text{logit Pr}(y_i = 1/x_i, \beta) = x_i' \beta$; where β is a vector of unknown regression coefficients with prior Normal distribution $\pi(\beta) = N(\beta_0, \Sigma_\beta)$.

Thomas Y [4] studied Bivariate Binomial Responses by vgam family Functions B= breathlessness, W=wheeze; (B=i, W=j; i, j=0 ,1), but Hun, M. in 2009 studied the Regression Models for binary dependent variables and the analysis data by Using Stata, SAS etc., He used data from clinicians and practitioners simulation study two responses (trust 1 respondent,0 otherwise, www internet used 1 respondent, 0 otherwise) and five independents variables.

In 2014 Tabatabai, MA [5] and others had studied methods for robust Logistic and probit compared with MLE when there are outlier values, they have been rely on real data and simulation experience for ($x_i=1,2$) as independent variables, they proved robust method is efficient. but they didn't apply bivariate logistic of response.

Theory

Logistic distribution

In this part of paper, we display some basic concepts of application distribution Logistic experiments in which the data to the variable appears to stop responding (adopted) is continuous or binary data such as nominal or countable (classified), which does not require a well-known hypotheses linear regression model and there should be no outliers in the data, logistic regression assumes linearity of independent variables and log odds. Only it requires quite large sample sizes, because of maximum likelihood estimate require classification according to the

*Corresponding author: Fadhil Abdul Abbas Al-Abidy, Department of Statistics, Al-Furat Al-Awast Technical University, General Street-Player city, Najaf, Iraq; Tel: +9647801376429; E-mail: Abidy_fadhil@yahoo.com

Received October 06, 2015; Accepted October 26, 2015; Published November 02, 2015

Citation: Fadhil Abdul AA (2015) Comparison between Robust and Classical Analysis in Bivariate Logistic for Medical Data. J Biom Biostat 6: 262. doi:10.4172/2155-6180.1000262

Copyright: © 2015 Fadhil Abdul AA. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

number of response variables (the number of dependent variables) classified to [6,7]:

Binary regression model

This model depends on the following equation:

$$y_i = \pi(x_i; \beta) + \varepsilon_i ; i = 1, 2, \dots, n$$

Where ε_i are independent $\forall i = 1, 2, \dots, n$

$$y_i \sim \text{IBer with men} = \pi(x_i; \beta) \text{ and var} = \pi(x_i; \beta)(1 - \pi(x_i; \beta)) ; i = 1, 2, \dots, n$$

a binary response variable (Y=1 or Y=0) is associated a set of explanatory variables, as from the following functions:

$$\pi = \text{prob}(y = \text{outcomes of itersted} / X = x)$$

$$\text{logit} \left(\frac{p(y=1)}{1-p(y=1)} \right) = \delta(x_i; \hat{a})$$

is call simple logistic regression

Multiple logistic model

$$\text{logit}(\pi(x)) = \ln \left(\frac{p(y=1)}{1-p(y=1)} \right) = \frac{\text{Exp} \sum_{j=1}^p \beta_j X_j}{1 + \text{Exp} \sum_{j=1}^p \beta_j X_j} ; x_0 = 1$$

After a series of mathematical operations we get the following formula of binary multiple regression:

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

We see that the logit of the probability of an event given X is a linear function and then MLE or OLS can be applied to complete inference of parameters β .

Bivariate logistic distribution

This model shows of analysis Logistic in medical studies and biological, for example, eye tests when; one where the response may vary from other response (uncorrelated) or possible interrelated as well as responding in agricultural experiments when two plants in the piece tested one of our show bilateral response, that is, $Y = (Y_1, Y_2)$ and the values of this vectors are as:

$$Y_i = \begin{cases} 1 \text{ success} \\ 0 \text{ failure} \end{cases} \quad i = 1, 2$$

$$\pi = \text{prob}(y = \text{outcomes of itersted} / X = x)$$

We can use bivariate probit regression models and these models have two equations for two binary dependent variables as the following equations:

$$y_1 = a_1 + b_1 y_2 + x' \beta + \varepsilon_1$$

$$y_2 = a_2 + x' C + \varepsilon_2$$

Where:

(y_1 and y_2): are a dummy responses as binary number (0,1), and y_2 as endogenous dummy in first equation.

β, C : vectors of parameter's model be estimated.

X: matrix with size (p.n) of independent variables.

\hat{a}_1 & \hat{a}_2 errors terms in equations assumed to be independent and

identically bivariate standard normal as pdf

$$N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right] = \varnothing(\varepsilon_1, \varepsilon_2, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left[\frac{-1}{2(1-\rho^2)} (\varepsilon_1^2 + \varepsilon_2^2 - 2\rho\varepsilon_1\varepsilon_2) \right]$$

and the research relies on statistical estimation methods such as

bivariate OLS or Robust multivariate as M_estimator, R_estimator [8] but as a result of developments in the software can be relied upon one of these software (Stat) including several functions explains us in the application part [9,10].

Statistical Techniques

It was used set of metrics (standards) to determine the statistical model appropriate to the data as well as to test the effect of the independent variables dependent on factors such as: [1,9]

Likelihood-ratio chi-square test statistics or Wald chi-square test and P-values;

Akaike's Information Criteria (AIC) and Bayesian information criteria (BIC);

Parameter estimates and standard errors of the study/exposure variable.

Some tables about spread phenomena (fatty liver, increased liver size) according to Gender and Age classes.

Application

Been relying on a random sample data of medical experience in the field of pathological analyzes conducted by researcher [11] in Al-Sadr Teaching Hospital, the province of Najaf on the study the relationship between blood variables and liver disease (fat & size) on the human where selected a random sample of 150 patients, the study variables:

Dependent variables

Fatty_type - the emergence of the fatty degree on the liver (0 lack of fat, 1 the presence of fat).

Liver_type - increase in the Liver size (0 normal, 1 an increase in size).

Independent variables

Age, Gender, Body mass index (BMI), liver enzymes (Gpt, Alk, Got), cholesterol Cho, triglycerides TG, sugar controlled Hb1Ac.

The program was adopted by the statistical software (Stata ver.2014) through analysis Bivariate Binary Logistic Regression, because the response variables in the experiment are binary numbers (1,0), taking into account the lack of independence of the response variables and can be continued to reach the method of implementation through the following illustrative screen [12] (Figure 1),

The data were analyzed according to the following cases:

Classical method: Do not use Robust analysis we got the following results

Through the results shown in the Table 1 above it is clear to display the following:

The test value (Likelihood Ratio LR) show an appropriate model used for analysis ($\chi^2=58.34$ with $P<0.0001$) as well as the existence of a positive relationship between the variables of the study accredited

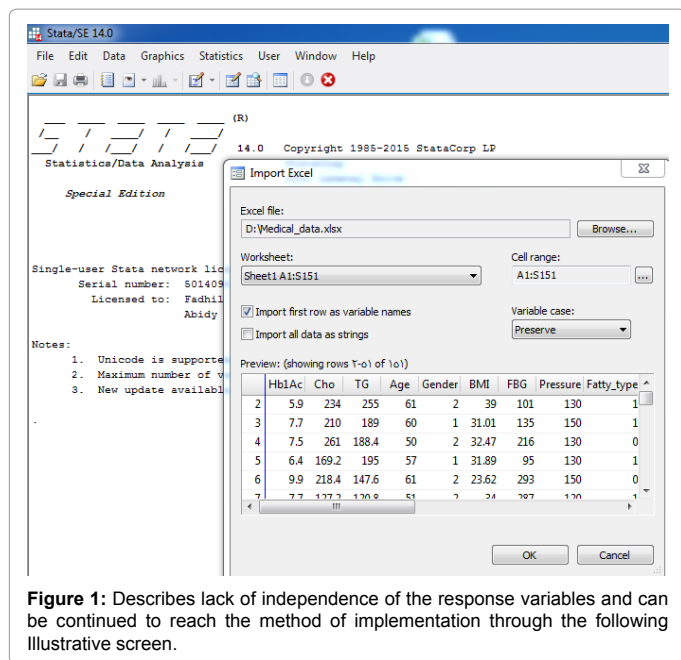


Figure 1: Describes lack of independence of the response variables and can be continued to reach the method of implementation through the following illustrative screen.

	Coef.	Std. Err	z	P> z	(95% Conf. Interval]
Fatty_type					
Age	-0.0138739	0.0126286	-1.10	0.272	-0.038626 0.0108778
Gender	0.0034449	0.2610207	0.01	0.989	-0.508146 0.515036
BMI	0.0701801	0.0269094	2.61	0.009	0.0174386 0.1229216
Hb1Ac	-0.1545435	0.0853633	-1.81	0.070	-0.321853 0.0127656
Cho	-0.0001859	0.0025515	-0.07	0.942	-0.005187 0.004815
TG	0.0008252	0.0014075	0.59	0.558	-0.001933 0.0035838
Pressure	0.0074593	0.0076659	0.97	0.331	-0.0075657 0.0224843
Goe	0.1296733	0.0602018	2.15	0.031	-0.01168 0.094098
Gpe	0.0184007	0.0386218	0.48	0.634	-0.057297 0.0108221
ALI<	0.0019094	0.0045473	0.42	0.675	-0.007003 0.0108221
cons	-2.474073	1.517789	-1.63	0.103	-5.448886 0.5007396
Liver_type					
Age	0.0249884	0.0125168	2.00	0.046	0.0004559 0.0495209
Gender	0.1890935	0.2448802	0.77	0.440	-0.290863 0.6690499
BMI	-0.0045514	0.0225854	-0.20	0.840	-0.048818 0.0397153
Hb1Ac	-0.0882778	0.0824425	-1.07	0.284	-0.249862 0.0733065
Cho	0.0070667	0.0028511	2.48	0.013	0.001479 0.0126545
TG	-0.0012979	0.0013507	-0.96	0.337	-0.003945 0.0013493
Pressure	0.0059725	0.00675	0.88	0.376	-0.007257 0.0192021
Goe	0.1645504	0.0582545	2.82	0.005	0.0503737 0.2787271
Gpe	-0.0513203	0.0367339	-1.40	0.162	-0.123318 0.0206769
ALI<	-0.0023338	0.0046188	-0.51	0.613	-0.011386 0.0067188
cons	-3.352374	1.487627	-2.25	0.024	-6.268068 -0.4366789
/athrho	0.12393	0.1677924	0.74	0.460	-0.204937 0.4527971
rho	0.1232994	0.1652415			-0.202115 0.4241955

Table 1: Result analysis of bivariate logistic model by classical method.

biprobit Fatty_type Liver_type Age Gender BMI Hb1Ac Cho 'J'C Pressure Got Gpt ALK,

Bivariaee probie regression Number of obs = 150

LR chi2(20) = 58.34

Lcq likelihcod = -158.66871 Prob>chi2 = 0.0000

LR<ces<cofrho=O:chi2(1)=.550495 Prob>chi2=0.4581

	Coef.	Std. Err	z	P> z	(95% Conf. Interval]
Fatty_type					
Age	-0.0138739	0.0135203	-1.03	0.305	-0.040373 0.0126253
Gender	0.0034449	0.241705	0.01	0.989	-0.470288 0.477178
BMI	0.0701801	0.0267631	2.62	0.009	0.0177254 0.1226349
Hb1Ac	-0.1545435	0.0847402	-1.82	0.068	-0.320631 0.0115443
Cho	-0.0001859	0.0024457	-0.08	0.939	-0.004979 0.0046076
TG	0.0008252	0.001472	0.56	0.575	-0.00206 0.00371
Pressure	0.0074593	0.0073579	1.01	0.311	-0.006962 0.0218805
Goe	0.1296733	.057284	2.26	0.024	0.0173988 0.2419479
Gpe	0.0184008	0.0386689	0.48	0.634	-0.057389 .0941904
ALI<	0.0019094	0.0044956	0.42	0.671	-0.006902 0.0107207
cons	-2.474072	1.703948	-1.45	0.147	-5.81375 0.8656052
Liver_type					
Age	0.0249884	.0116016	2.15	0.031	0.0022496 .0477272
Gender	0.1890935	0.2414576	0.78	0.434	-0.284155 0.6623416
BMI	-0.00455	0.023124	-0.20	0.844	-0.04987 0.04077
Hb1Ac	-0.0882778	0.0835978	-1.06	0.291	-0.252127 0.075571
Cho	0.0070667	0.0029229	2.42	0.016	0.001338 0.0127955
TG	-0.0012979	0.0013006	-1.00	0.318	-0.003847 0.0012511
Pressure	0.0059725	0.007624	0.78	0.433	-0.00897 0.0209152
Goe	0.1645504	0.0522388	3.15	0.002	0.0621643 .2669365
Gpe	-0.0513203	0.031757	-1.62	0.106	-0.113563 0.0109223
ALI<	-0.0023338	0.0041128	-0.570	0.570	-0.010395 .0057272
cons	-3.352374	1.492694	-2.25	0.025	-6.278 -0.4267488
/athrho	0.1239301	0.1580029	0.78	0.433	-0.18575 0.43361
rho	0.1232995	0.1556008			-0.183643 0.4083338

Table 2: Result analysis of bivariate logistic model by Robust method.

biprobit Fatty_type Liver_type Age Gender BMI Hb1Ac Cho 'J'C Pressure Got Gpt AW<, vce (robust)

Bivariaee probie regression Number of obs=150
Waldchi 2 (20)=71.71

Lcqpsedu like lihcod=-158.66871 Prob>chi2=0.0000

(Fatty_type and Liver_type) (r=0.123), but not significant (p_value=0.458).

There is effect of some of the independent variables on the dependent variable (Fatty_type) where he showed variable (BMI) very high significant with p_value 0.009, enzyme liver (Got) with probability (p = 0.031) and the degree of a simple effect of sugar control (Hb1Ac) with p_value 0.07, while significant effect of other variables did not show.

The existence of the impact of some independent variables on the dependent variable (Liver_type) where it showed enzyme liver (Got) high impact on increasing the liver size with p_value=0.005, fat cholesterol valued likely would (p=0.013) while the age factor is effect with p_value= 0.046, while effect of other variables did not significant.

Robust method: We chosen robust method of analysis to determine the effect of outlier values on the independent variables which assumed some of the values (10%) of them, because the dependent variables are binary data (0,1), the results showed in the following Table 2:

The test value (Likelihood Ratio LR) show an appropriate model used for for analysis ($\chi^2 = 71.71$ with $p < 0.0001$).

The results in Robust analysis are not different without Robust, for (Fatty_type), where the variable (BMI) has very high effect with

p_value=0.009, enzyme liver (Got) with p_value=0.024, but simple effect of sugar control(Hb1Ac) with p_value=0.068, while no significant effect of other independent variables.

The existence of the impact of some of the independent variables on the dependent variable (Liver_type) as variable enzyme liver (Got) high impact on increasing the liver size with p_value=0.002 and fat cholesterol is significant with p_value=0.016, while the age factor influential degree p_value=0.031, while significant effect of other variables did not show.

Additional statistical tables are needed to show the numbers of the spread of fatty liver and liver size of the study sample according to social standards (Gender, Age group).

Through the above Table 3 it is clear to us that:

Gender Fatty Type.

Gender	Fattytype		Total
	0	1	
1	18	44	62
2	28	60	88
Total	46	104	150

Age class Liver Type

Ageclass	Livertype		Total
	0	1	
40-55 years	23	42	65
55-70 years	19	43	62
<40 years	8	7	15
>70 years	1	7	8
Total	51	99	150

Age class Fatty Type

Ageclass	Fattytype		Total
	0	1	
40-55 years	17	48	65
55-70 years	19	43	62
<40 years	5	10	15
>70 years	5	3	8
Total	46	104	150

Table 3: Study sample according to social standards (Gender, Age group).

The spread of fatty liver in the study sample is 69% (104/150) and the proportion of its spread in females 59% (88/150) compared with males, while increasing the liver size in the research sample that 66% have a problem with increasing the size from the normal, and the highest proportions in the age groups (55-70 years; 40-55 years) 43% and 42% respectively.

While the fat spread problems by 46% in the age group (40-55 years) and close to this ratio in the category (55-70 years), and is identical with the medical condition after the injury fat in the previous age group starts the problem of increasing the liver size in next age.

Conclusions

There is a positive correlation between liver fatty and increase of the size liver responses, but this correlation not significant.

Impact of the explanatory variables (BMI, Age, Hb1Ac, Got, Cho) is approved on liver fatty changes and increased liver size, while do not receive the influence of other explanatory variables on these responses.

Spread of fatty liver disease in females than in males according to the research sample was more prevalent in the age group (40-55 years) while the prevalence of liver size ratio in the age group (55-70 years), a medically acceptable because Category Previous showed the spread of fatty liver disease where after years leads to an increase (inflation) size of the liver.

Recommendations

Use simulations to test the success of the methods used under several influences (errors distribution type, sample size and assuming a high correlation between the two variables of response values).

More studies are required to study the analysis of logistics multi-variable models.

References

- Mc Donald BW (1993) Estimating Logistic Regression Parameters for Bivariate Binary Data. *Journal of the Royal Statistical Society* 55: 391-397.
- Kimberlee G, Marcello P (1997) The Analysis of Correlated Binary Outcomes Using Multivariate Logistic Regression. *Biometrical Journal* 39: 309-325.
- O'Brien SM, Dunson DB (2004), Bayesian Multivariate Logistic Regression. *Biometrics* 60: 739-746.
- Yee TW (2008) VGAM Family Functions for Bivariate Binomial Responses.
- Tabatabai MA, Li H, Eby WM, Kengwoung-Keumo JJ, Manne U, et al. (2014) Robust Logistic and Probit Methods for Binary and Multinomial Regression. *J Biomet Biostat* 5: 202.
- Cessie S, Van Houwelingen JC (1994) Logistic regression for correlated binary data. *Applied Statistics* 43: 95-108.
- Hosmer D, Lemeshow S (2000) Applied Logistic Regression. John Wiley & Sons Ltd, UK.
- Robert P (2011) Robust Multivariate linear regression using the student t-distribution. Tampere University of Technology.
- Hun MP (2009) Regression Models for Binary Dependent Variables Using Stata, SAS, R, LIMDEP, and SPSS.
- Stata Press (2007) Stata Base Reference Manual.
- Rusul FA (2014) Relationship between Type II Diabetes Mellitus and Fatty Liver in Human.
- Palmgren J (1989) Regression models for bivariate binary responses.