IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 12, Issue 4 Ver. V (Jul. - Aug.2016), PP 125-130
www.iosrjournals.org

Comparison between Logistic and Calibration Linear Regression

Asma Ali Mohammedkhair[1]*,
Ahamed Mohamed Abdalla Hamdi[2],Mohammedelameen Eissa Qurashi [3]
[1, 2 ,3]*Sudan University of Science & Technology, Faculty of science, Department of Statistics Researcher*

***Abstract**: This paper hosts a comparison between logistic regression model and calibration linear model . Both models were applied on random sample of 120 people , 100 are infected with blood cancer and 20 are fit . And we have 3 independent variables , age , pcv , mch . When applying both models we discovered that the values of standard errors in calibration regression model are less than the value of standard errors in logistic regression model , meaning that calibration regression method was better . Some other results were reached , like when applying logistic all variables mentioned above have significant influence on cancer infection , we also found that pcv variable is the most influential in cancer infection , followed by the rest age and msh.*
***Keyword:** Linear model, classical estimator, calibration model,vce(robust) ,efficiency, Blood cancers.*

I. Introduction

Statistical calibration analysis provide away to predict a quantity from the observation of another one by using adose-response type relationship. The problem occurs in biological sciences when the quantity to be calibration is hard or expensive to measure or is not observable. It is important in any topic of calibration to distinguish between absolute and comparative calibration these two activities are both called calibration, they are conceptually different and lead to different issues in statistical modeling.

In absolute calibration a quick or non –standard measurement is either known or made with negligible error. With comparative calibration one instrument or measurement technique is calibrated against another with neither one being inherently a standard so that there is no standard measurement X.we discuss here absolute calibration.In the another hand logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more metric (interval or ratio scale) independent variables.

Logistic regression assumes that the dependent variable is a stochastic event. That is that for instance if we analyze a pesticides kill rate the outcome event is either killed or alive. Since even the most resistant bug can only be either of these two states, logistic regression thinks in likelihoods of the bug getting killed. If the likelihood of killing the bug is > 0.5 it is assumed dead, if it is < 0.5 it is assumed alive.

The outcome variable – which must be coded as 0 and 1 – is placed in the first box labeled Dependent, while all predictors are entered into the Covariates box (categorical variables should be appropriately dummy coded). SPSS predicts the value labeled 1 by default, so careful attention should be paid to the coding of the outcome (usually it makes more sense to examine the presence of a characteristic or "success."

Mathematically logistic regression estimates a multiple linear regression function defined as:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_1 - \hat{\beta}_2 \mathrm{X}_2 - \hat{\beta}_3 \mathrm{X}_3 + \ldots \hat{\beta}_p \mathrm{X}_\mathrm{p} \ldots\ldots\ldots(1)$$

$$\mathrm{i} = 1,2,\ldots,\mathrm{n}$$

When selecting the model for the logistic regression analysis another important consideration is the model fit. Adding independent variables to a logistic regression model will always increase its statistical validity, because it will always explain a bit more variance of the log odds (typically expressed as R^2). However, adding more and more variables to the model makes it inefficient and over fitting occurs.

Nevertheless, many people want an equivalent way of describing how good a particular model is, and numerous pseudo- R^2 values have been developed. These should be interpreted with extreme caution as they have many computational issues which cause them to be artificially high or low. A better approach is to present any of the goodness of fit tests available; Hosmer-Lemeshow is a commonly used measure of goodness of fit based on the Chi-square test (which makes sense given that logistic regression is related to cross tabulation).(www.Statisticsolutions.com)

II. Blood Cancers

Blood cancers affect the production and function of your blood cells. Most of these cancers start in your bone marrow where blood is produced. Stem cells in your bone marrow mature and develop into three types of blood cells: red blood cells, white blood cells, or platelets. In most blood cancers, the normal blood cell development process is interrupted by uncontrolled growth of an abnormal type of blood cell. These abnormal blood cells, or cancerous cells, prevent your blood from performing many of its functions, like fighting off infections or preventing serious bleeding.

There are three main types of blood cancers:

Leukemia, a type of cancer found in your blood and bone marrow, is caused by the rapid production of abnormal white blood cells. The high number of abnormal white blood cells are not able to fight infection, and they impair the ability of the bone marrow to produce red blood cells and platelets.

Lymphoma is a type of blood cancer that affects the lymphatic system, which removes excess fluids from your body and produces immune cells. Lymphocytes are a type of white blood cell that fight infection. Abnormal lymphocytes become lymphoma cells, which multiply and collect in your lymph nodes and other tissues. Over time, these cancerous cells impair your immune system.

Myeloma is a cancer of the plasma cells. Plasma cells are white blood cells that produce disease- and infection-fighting antibodies in your body. Myeloma cells prevent the normal production of antibodies, leaving your body's immune system weakened and susceptible to infection. (www.hematology.org)

III. Variables Research

3.1 Age:

Cancer is primarily a disease of older people, with incidence rates increasing with age for most cancers Children aged 0-14, and teenagers and young adults aged 15-24, each account for less than one per cent of all new cancer cases in the UK (2011-2013) for example Adults aged 25-49 contribute a tenth (10%) of all new cancer cases, with twice as many cases in females as males in this age group. Adults aged 50-74 account for over half (53%) of all new cancer cases, and elderly people aged 75+ account for over a third (36%), with slightly more cases in males than females in both age groups. There are more people aged 50-74 than aged 75+ in the population overall, hence the number of cancer cases is higher in 50-74s, but incidence rates are higher in 75+.www.cancerresearchuk.org

.

3.2 Pcv:

The Pcv test is used to measure the amount of cells in the blood; blood is made up of cells and plasma. The amount of cells in the blood is expressed as a percentage of the total volume of blood; for example, a Pcv measurement of 50% means that there are 50 millilitres of cells per 100 millilitres of blood. The Pcv measurement may increase or decrease depending on the individual's health; if they are dehydrated, the measurement will rise and the measurement will decrease if the individual has a condition, such as anaemia.

The Pcv test is usually ordered as part of the series of tests that make up the full blood count. The test is used to diagnose and monitor conditions including anaemia, polyctyhaemia and dehydration. The test may also be used to determine whether an individual is fit to have a blood transfusion; the test may also be repeated regularly to check whether the transfusion has worked effectively.The test is usually ordered to monitor the condition of people who have been diagnosed with anaemia; it may also be used to monitor those with dehydration and persistent bleeding.

The test is done by taking a sample of blood from the patient; in most cases, the sample is taken from a vein in the patient's arm. A needle is inserted into the vein and the blood is drawn out and collected in a syringe. Once the sample has been collected, it will be bottled, labelled with the patient's name and sent off to a laboratory for testing.

In children, a sample may be collected from the fingertip; in infants it may be collected from the heel. The samples are obtained by pricking the finger or the heel with a needle. A decreased Pcv result usually indicates anaemia. A low Pcv count may also indicate vitamin or mineral deficiencies, liver cirrhosis and bleeding episodes.Increased Pcv results are usually associated with dehydration; in most cases, the Pcv result will return to normal once the individual has increased their fluid intake.

High Pcv results may also be caused by polycythaemia vera, a condition which occurs when an individual has too many red blood cells; this is caused by a problem with the function of the bone marrow. Living at high altitude usually increases Pcv. Pregnancy usually causes Pcv results to be slightly lower than normal. (www.medic8.com)

3.3 Mch:

Mch is the initialism for Mean Corpuscular Hemoglobin. Taken from Latin, the term refers to the average amount of hemoglobin found in red blood cells. A CBC (complete blood count) blood test can be used to monitor Mch levels in blood. Lab Tests Online explains that the Mch aspect of a CBC test "is a measurement of the average amount of oxygen-carrying hemoglobin inside a red blood cell. Macrocytic RBCs are large so tend to have a higher Mch, while microcytic red cells would have a lower value." Mch levels in blood tests are considered high if they are 35 or higher. A normal hemoglobin level is considered to be in the range between 26 and 33 picograms per red blood cell.

High Mch levels can indicate macrocytic anemia, which can be caused by insufficient vitamin B12. Insufficient folic acid can be another cause of macrocytic anemia. Alcohol abuse can be a contributing factor, and should be disclosed in the diagnostic process to better enable accuracy in diagnosis and in treatment determination. A simple calculation is used to determine the mean corpuscular hemoglobin level in blood. According to Med Friendly, the total amount of hemoglobin in the sample is multiplied by ten and then divided by the number of red blood cells present.

The method recommended to treat a high Mch level will depend upon what is causing it in the patient. The treatment will also depend upon other medical conditions and any medications the patient may be taking. Allergies will also be taken into account. Any person with a high Mch level should carefully discuss treatment with his physician and follow the directions carefully. Any dietary supplements and over-the-counter medications should be disclosed in order to prevent any negative results. If the cause is macrocytic anemia, the treatment could involve adding liver to the diet or adding more vitamin B12.(www.brighthub.com)

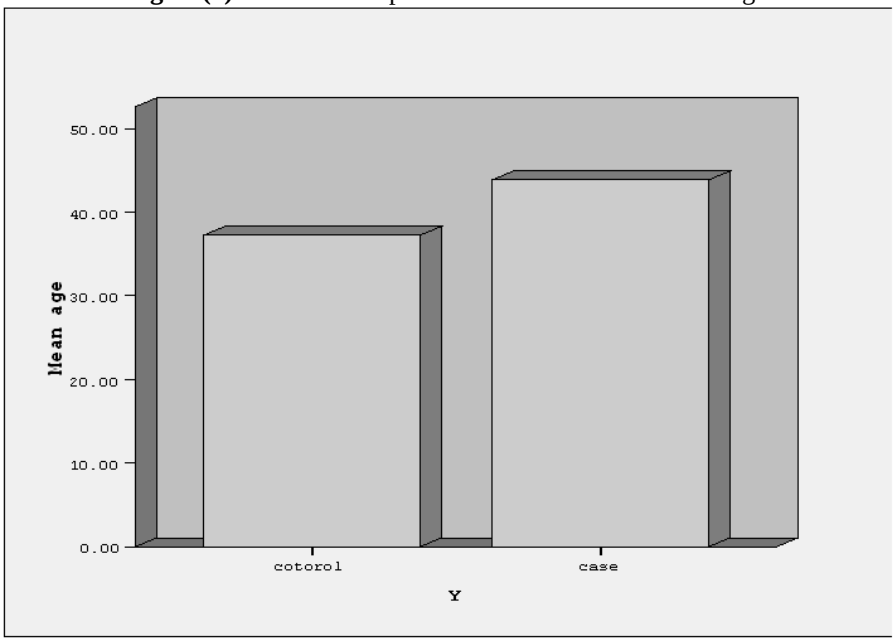
IV. Application Aspect

The application aspect includes to what explained in the theoretical aspect and depending on Statistical software "SPSS & STATA", we would describe the data, estimate parameters models and comparison between calibration and logistic models.

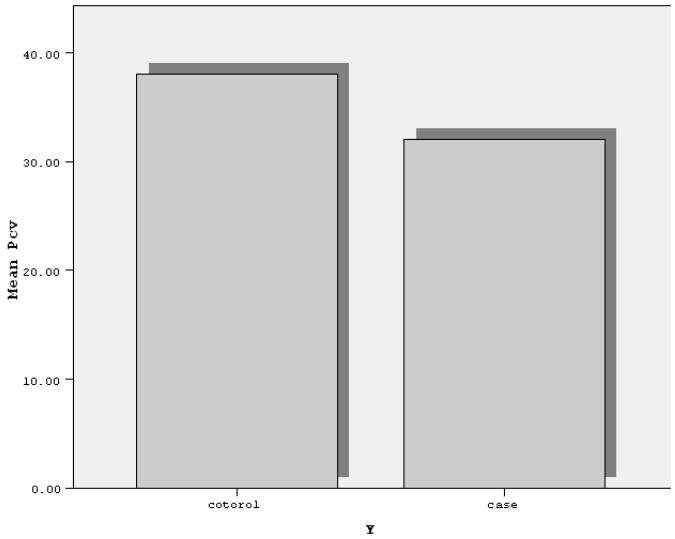
Table (1): Describtives

	N	*Mean*	*Std.Devision*	*Std.Erorr*
Age control case	20	37.250	6.307	1.410
	100	43.870	10.556	1.056
Pcv control case	20	38.050	1.791	0.400
	100	31.985	5.454	0.545
Mch control caserr	20	28.550	1.234	0.276
	100	27.030	2.401	0.240

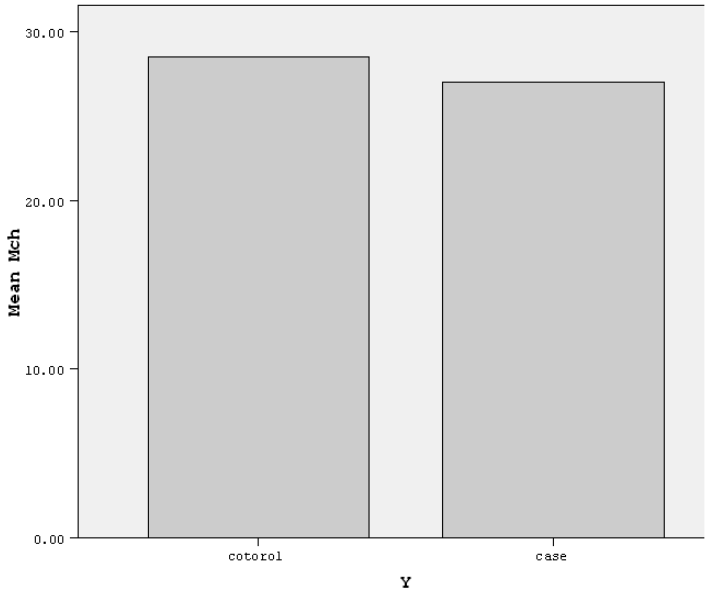
Figure(1): Bar Chart Explain Variable Y And Mean of Age



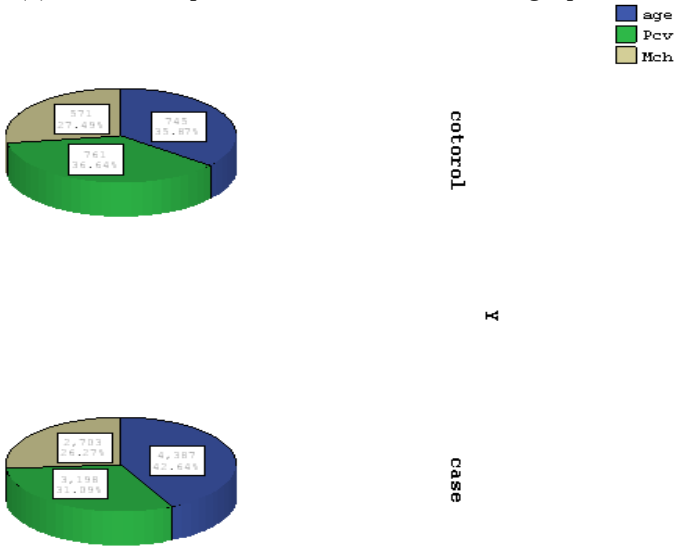
Figure(2): bar chart explain variable Y and mean of pcv



Figure(3): bar chart explain variable Y and mean of mch



Figure(4): bar chart explain variable Y and variables age ,pcv and mch



From the table(1) and the graph 1, 2, 3 and 4 in the variable Age we find the mean of the infected 37,250 while the fit 43,870 year with standard error 6.307and 10,556 respectively .for the variable Pcv the average of the infected is 38,050 and the fit 31,985 years with standard error 1,791 and 5,454 .

The average of Mch for the fit is 27,030 year and the infected 28,550 with standerd error of 2,401 and 1,234 .

Table(2): Variables in Equation

	β	*S.E*	*Wald*	*d.f*	*sig*	$Exp(\beta)$	0.95% C.I for Exp(β)	
							Lower	*Upper*
Age	0.121	0.044	7.572	1	0.006	1.128	1.035	1.229
Pcv	-0.288	0.084	11.617	1	0.001	0.750	0.636	0.885
Mch	-0.344	0.160	4.611	1	0.032	0.709	0.518	0.970
Constant	16.591	5.690	8.503	1	0.004	16046912		

The table 2 includes all the models parameters in addition to some statistics like standard errors of the model , Wald statistics , degrees of freedom , and the final two columns are the exponential function added to both confidence interval columns. The null hypotheses assumption that is to be tested to know if the model parameters are influencing the dependent variable or not $(H_0 = 0)$. And we notice that all parameters are significant which means the non existence assumption is rejected meaning that the variables age pcv and mch affect cancer resilience.

Logistics inclination model can be written as follows:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 16.591 + 0.121\text{Age} - 0.288\text{Pcv} - 0.344\text{Mch}$$

Where $\hat{p}$ refers to blood cancer infection and these estimations clarify the relation between the dependent variable (infection / non infection) and the independent variable using ((log it)) units yet we noticed that the independent variable is most affected by the Pcv variable with a parameter -0.288 and the value of $sig = 0.001$ and on the lower degree of affecting the independent variable comes the Age variable with a parameter of 0.121 and $sig = 0.006$. On the final level comes the variable Mch with a parameter of -0.344 and a probable value of 0.032.

Table (3): Omnibus Tests of Model Coefficients

	$Chi-square$	$d.f$	*sig*
Model	38.450	3	0

The table 3 tests the quality of the logistic model being used , we found the value of χ^2 test equals 38,450 and is significant at $\alpha = 0.05$ and the probable value equals zero which asserts the previous result , meaning that the null hypothese assumption has been rejected which means the model is significant.

Table(4): Hosmer and Lemeshow test

$Chi-square$	$d.f$	*sig*
7.287	8	0.506

The table 4 clarifies that χ^2 test equals 7,285 with a degree of freedom equal 8 and probable value 0,506 which asserts the goodness of the method as a whole.

Table(5): Classification Table

Observed		predicted		
		Y		Percentage Correct
		Control	Case	45.0
Y	Control	9	11	94.0
	Case	6	94	85.8
Over all precentage				

The table(5) clarifies the percentage of the right classification which is 85,8% and the percentage of wrong classification 14,2% which means the model presents the data very well.

Table(6): Standard Error by Using Vce(Robust)

Y	*Coef.*	*Robust* Std.Error	*t*	$P>\|t\|$	0.95% C.I	
Age	0.009	0.003	3.39	0.001	0.004	0.015
Pcv	-0.024	0.006	-4.13	0	-0.356	-0.012
Mch	-0.028	0.012	-2.29	0.024	-0.051	-0.004
Constant	1.987	0.282	7.04	0	1.428	2.546

The table (6) clarifies regress of the dependent variable (y) (infected / not infected) on the dependent variables mentioned before , the aim behind establishing this model is to find standard errors , known here by (Robust Std. Error) then comparing it with standard errors column in table (2).

We notice here that the value of trusted calibration errors is less from standard errors in logistic regression model table (2). And it is known that the less errors in a model , the more accurate and more capable of forecast it becomes . Meaning that by using calibration linear regression model , error were made less with more accuracy. However logistic was well tested in table (5) and (6) yet calibration regression was noticeably better.

V. Conclusions

In this paper calibration regression and logistic regression models were applied on data that was taken from Khartoum hospital , 100 are infected with blood cancer and 20 fit. Reaching the fact that calibration regression model was better in illustrating the data. Also discovered t he variable Pcv is the most influential in blood cancer infection followed by Age and Mch

Acknowledgement

I would take this opportunity to thank my research supervisor Dr. Ahamed Mohamed Abdalla Hamdi, and special thanks to our great teacher and my idle role Prof. Zainelabdian A.El Beshir, professor in alneelain university department of statistic for their support and guidance without which this research would not have been possible**.**

References

[1]. MARIE-ANNE GRUET,(1996) , A non Parametric Calibration Analysis
[2]. Christine Osborne . (1991). Statistical Calibration: A Review. School of Mathematical Sciences, University of Bath, Bath BA2 7A Y, England
[3]. AITCHISON, J., and DUNSMORE, I. R. (1975), Statistical Prediction Analysis, London: Cambridge University Press.
[4]. DUNSMORE, I. R. (1968), "A Bayesian Approach to Calibration," Journal of the Royal Statistical Society.
[5]. Cox, D. R. (1981). Theory and general principle in statistics: the Address of the President (with Proceedings). J. R. Statist.
[6]. Simultaneous pairwise linear structural relationships. Biometrics 25, 129-142. Berkson, J. (1969).
[7]. Estimation of a linear function for a calibration line: consideration of a recent proposal. Technometrics 11, 649-660. Breiman, L. & Friedman, J.H. (1985).
[8]. Estimating optimal transformations for multiple regression and correlation. J. Am. Statist. Assoc. 80, 580-597. Brown, G.H. (1979).
[9]. An optimisation criterion for linear inverse estimation. Technometrics 21, 575-579. Brown, P.J. (1982).
[10]. Multivariate calibration (with discussion). J. R. Statist. Soc. B 44, 287-321. Brown, P.J. & Sundberg, R. (1987)
[11]. Confidence and conflict in multivariate calibration. J. R. Statist. Soc. B 49, 46-57. Brown, P.J. & Sundberg