International Journal of Sciences:
Basic and Applied Research
(IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Binary Logistic Regression to Identify the Risk Factors of Eye Glaucoma

Nidal Mohamed Mustafa Abd Elsalam*

Department of Statistics & Computation
Faculty of Technology of Mathematical Sciences & Statistics, Al Neelain University
E-mail: nidalmm2@gmail.com

Abstract

A binary logistic regression is performed to predict the presence or absence of Eye Glaucoma and to identify the risk factors of the disease. Five predictor variables are included in the model, which are: age, gender, inheritance, diabetes and hypertension. The results of the logistic regression analysis show that the full model, that considered all the five independent variables together, is statistically significant. . The most significant predictor in the model is the age factor and when it is raised by one year, the person is six times more likely to get sick with eye glaucoma while the other variables in the model are controlled but are regarded as clinically important. Moreover, the logistic model explains about 75.5% of the cases for absence of glaucoma, 83.9% for its presence and it correctly classifies 80% of the included cases.

***Keywords*: Binary; Logistic regression; Maximum Likelihood; Discriminant Function.**

1. Introduction

Usually in a medical setting, an outcome might be presence or absence of a certain disease. Logistic regression analysis is widely known to be a valuable tool in extending the techniques of multiple regression analysis to research situations in which the outcome variable is categorical , through taking on two or more possible values. This is very clearly stated by [1,2,3,4].

----

* Corresponding author.

E-mail address: nidalmm2@gmail.com.

In this paper, the risk factors for the disease of Eye Glaucoma are identified using logistic regression analysis. Finding the risk factors and the potential risk factors can help prevent the development of the disease. The logistic regression model has become the standard method of analysis in this situation, as stated by [5,6,7]. And like any other model building technique, the goal of the logistic regression analysis is to find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables.

Glaucoma is a condition that causes damage to your eye's optic nerve and gets worse over time. It's often associated with a buildup of pressure inside the eye. Glaucoma tends to be inherited and may not show up until later in life. The increased pressure, called intraocular pressure, can damage the optic nerve, which transmits images to the brain. This can happen when eye fluid isn't circulating normally in the front part of the eye. If damage to the optic nerve from high eye pressure continues, glaucoma will cause permanent loss of vision. Without treatment, glaucoma can cause total permanent blindness within a few years. Normally, that fluid, called aqueous humor, flows out of the eye through a mesh-like channel. If this channel becomes blocked, fluid builds up, causing glaucoma. The direct cause of this blockage is unknown, but doctors do know that it can be inherited, meaning it is passed from parents to children. Less common causes of glaucoma include a blunt or chemical injury to the eye, severe eye infection, blockage of blood vessels in the eye, inflammatory conditions of the eye, and occasionally eye surgery to correct another condition. Glaucoma usually occurs in both eyes, but it may involve each eye to a different extent.

There are two main types of eye glaucoma: The first type is Open-angle glaucoma which is also called wide-angle glaucoma and it is the most common type of glaucoma. The structures of the eye appear normal, but fluid in the eye does not flow properly through the drain of the eye, (trabecular meshwork).The second type is Angle-closure glaucoma, and sometimes is also called acute or chronic angle-closure or narrow-angle glaucoma. This type of glaucoma is less common but can cause a sudden buildup of pressure in the eye. Drainage may be poor because the angle between the iris and the cornea (where a drainage channel for the eye is located) is too narrow.

For most people, there are usually few or no symptoms of eye glaucoma. The first sign of eye glaucoma is often the loss of peripheral or side vision, which can go unnoticed until late in the disease. Detecting glaucoma early is one reason you should have a complete exam with an eye specialist every one to two years. Occasionally, intraocular pressure can rise to severe levels. In these cases, sudden eye pain, headache, blurred vision, or the appearance of halos around lights may occur. Thus the need for medical care emerges if any of the following symptoms occur: Seeing halos around lights, Vision loss, Redness in the eye, Eye that looks hazy (particularly in infants), Nausea or vomiting, Pain in the eye and Narrowing of vision (tunnel vision). To diagnose glaucoma, an eye doctor will test your vision and examine your eyes through dilated pupils. The eye exam typically focuses on the optic nerve which has a particular appearance in glaucoma. In fact, photographs of the optic nerve can also be helpful to follow over time as the optic nerve appearance changes as glaucoma progresses. The doctor will also perform a procedure called tonometry to check for eye pressure and a visual field test, if necessary, to determine if there is loss of side vision. Glaucoma tests are painless and take very little time. Glaucoma treatment may include prescription eye drops, laser surgery, or microsurgery. Moreover, Infant or congenital glaucoma -- meaning you are born with it -- is primarily treated with surgery since the cause of the problem is a much distorted drainage system. Glaucoma cannot be prevented, but if it is diagnosed and treated early, the disease can be controlled.

The main objectives of the study are to : identify the most important risk factors of eye glaucoma through the use of logistic regression, to find an appropriate discriminant function that helps in the diagnosis of the disease, and to know which risk factors are the most statistically significant.

The study resulted in the fact that a test of the full model against a constant only model was statistically significant, indicating that the predictors as a set reliably distinguished between presence and absence of Glaucoma. In addition to this, there exists a medium relationship between prediction and grouping. Prediction success overall was 80% (75.5% for absence of glaucoma and 83.9% for presence of glaucoma). It was revealed that only age factor made a significant contribution to prediction, while the rest of the predictors (Gender, inheritance, diabetes, and hypertension) were not significant ones. Moreover it was found that when age is raised by one year, the person is six times more likely to get sick with eye glaucoma.

2. Methods and Materials

The logistic regression model indirectly models the response variable based on probabilities associated with the values of the dependent variable Y. We will use P(x) to represent the probability that Y =1, which is the presence of Glaucoma. Similarly, we will define 1-P(x) to be the probability that Y =0, which is absence of Glaucoma. These probabilities are written in the following form:

$$P(x) = P(Y = 1|X_1, X_2, X_3 \dots, X_n)$$

$$1 - P(x) = P(Y = 0|X_1, X_2, X_3 \dots, X_n)$$

The log distribution (or logistic transformation of p) is also called the logit of p or logit (p) which is the log (to base e) of the odds ratio or likelihood ratio that the dependent variable is 1. In symbols it is defined as:

$$\text{Logit}(P) = \log [P / (1 - P)] = \ln [P / (1 - P)] \qquad \rightarrow (1)$$

Whereas P can only range from 0 to 1, logit (p) scale ranges from negative infinity to positive infinity and is symmetrical around the logit of 0.5 (which is zero). Formula (2) below shows the relationship between the usual regression equation ($Y = \alpha + \beta X$), which is a straight line formula, and the logistic regression equation. The form of the logistic regression equation is thus rewritten as:

$$\text{Logit } P(x) = \log [P(x) / (1 - P(x))] = \ln [P(x) / (1 - P(x))] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \dots \qquad \rightarrow (2)$$

This looks just like a linear regression and although logistic regression always finds a ‘best fitting’ equation, just as linear regression does, the principles on which it does so are rather different, Lei (8) and Peng (9). Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, Hosmer (10), which maximizes the probability of getting the observed results given the fitted regression coefficients. A consequence of this is that the goodness of fit and overall significance statistics used in logistic regression is different from those used in linear regression. P can be calculated with the following formula:

$$P = \frac{Exp^{\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3 \dots}}{1+ Exp^{\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3 \dots}} \qquad \rightarrow (3)$$

Where:

P = the probability that a case is in a particular category,

Exp = the base of natural logarithms (approx. 2.72),

α = the constant of the equation and,

β= the coefficient of the predictor variables.

A model including the intercept (constant) only is carried out firstly. Logistic regression compares this model with a model including all the predictors (Age, Gender, Inheritance, Diabetes and Hypertension) to determine whether the latter model is more appropriate in representing the data in a better way. Then information about the variables that are not included in the equation is provided to tell whether each independent variable (IV) improves the model. Moreover, we often want to look at the proportion of cases we have managed to classify correctly, for this reason a classification process is adopted.

Consequently, results after adding predictors are presented and model goodness of fit tests are carried out to show if each predictor contributes significantly to the model. The overall significance is tested and an approximation to the coefficient of determination R_2 is done.

A classification plot or histogram of predicted probabilities may be presented to provide a visual demonstration of the correct and incorrect predictions. Finally, the case-wise list produces a list of cases that didn't fit the model well which are known as outliers.

2.1 Statistical Tests

For more elaboration, the statistical tests of goodness of fit are revised and described briefly as follows:

Starting by Wald test, together with the associated probabilities, it provides an index of the significance of each predictor in the equation, as noted by[11]. The Wald statistic has a chi-square distribution. The simplest way to assess Wald is to take the significance values and if they are less than a value of 0.05, we reject the null hypothesis as the variable does make a significant contribution.

The Logits (log odds) are the *b* coefficients (the slope values) of the regression equation. The slope can be interpreted as the change in the average value of Y, from one unit of change in X. Logistic regression calculates changes in the log odds of the dependent, not changes in the dependent value as OLS regression does. For a dichotomous variable the odds of membership of the target group are equal to the probability of membership in the target group divided by the probability of membership in the other group, [12]. Odds value can range from zero to infinity and tell you how much more likely it is that an observation is a member of the target group rather than a member of the other group. If the probability is 0.80, the odds are 4 to 1 or 0.80/0.20; if the probability is 0.25, the odds are0.33 (0.25/0.75). If the probability of membership in the target group is 0.50, the odds are 1 to 1 (0.50/0.50), as in coin tossing when both outcomes are equally likely.

Another important concept is the odds ratio (OR), which estimates the change in the odds of membership in the target group for a one unit increase in the predictor. It is calculated by using the regression coefficient of the predictor as the exponent or Exp. Generally, in statistical packages, the odds ratio is called EXP (*b*). It eases our calculations of changes in the dependent variables DV due to changes in the independent variables IV. So the standard way of interpreting a ‘b’ in logistic regression is using the conversion of it to an odds ratio using the corresponding EXP (*b*) value. As an example, if the logit $b = 1.5$, then the corresponding odds ratio will be 4.48. We can then say that when the independent variable increases one unit, the odds that the case can be predicted increase by a factor of around 4.5 times, when other variables are controlled.

Maximum Likelihood (or ML) is used instead of the least squares approach to find the function that will maximize our ability to predict the probability of Y based on what we know about X. This is because the values of Y can only range between 0 and 1. In logistic regression, two hypotheses are of interest: the null hypothesis, which is when all the coefficients in the regression equation take the value zero, and the alternative hypothesis that the model with predictors currently under consideration is accurate and differs significantly from the null of zero, i.e. gives significantly better than the chance or random prediction level of the null hypothesis. We then work out the likelihood of observing the data we actually did observe under each of these hypotheses. The result is usually a very small number, and to make it easier to handle, the natural logarithm is used, producing log likelihood (LL). Probabilities are always less than one, so LL’s are always negative. Log likelihood is the basis for tests of a logistic model.

The likelihood ratio test is based on –2LL ratio. It is a test of the significance of the difference between the likelihood ratios for the researcher’s model with predictors (called model chi square) minus the likelihood ratio for baseline model with only a constant in it. Significance at the 0.05 level or lower means the researcher’s model with the predictors is significantly different from the one with the constant only (all ‘b’ coefficients being zero). It measures the improvement in fit that the explanatory variables make compared to the null model. Chi square is used to assess significance of this ratio. When probability fails to reach the 5% significance level, we retain the null hypothesis that knowing the independent variables (predictors) has no increased effects (i.e. make no difference) in predicting the dependent variable.

Cox and Snell’s R-Square attempts to imitate multiple R-Square based on ‘likelihood’, but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. Here it is indicating the how much is the percentage of the variation in the dependent variable DV is explained by the logistic model.

The Nagelkerke modification that does range from 0 to 1 is a more reliable measure of the relationship. Nagelkerke's R_2 will normally be higher than the Cox and Snell measure and it is the most-reported of the R-squared estimates, as stated by [13].

Hosmer and Lemeshow test constitutes an alternative to model chi-square in which subjects are divided into 10 ordered groups of subjects and then compares the number actually in the each group (observed) to the number predicted by the logistic regression model (predicted). The 10 ordered groups are created based on their estimated probability; those with estimated probability below 0.1 form one group, and so on, up to those with probability 0.9 to 1.0. Each of these categories is further divided into two groups based on the actual observed outcome variable (success, failure). The expected frequencies for each of the cells are obtained from the model. A probability (p) value is computed from the chi-square distribution with 8 degrees of freedom to test the fit of the logistic model. If the H-L goodness-of-fit test statistic is greater than 0.05, as we want for well-fitting models, we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level. That is, well-fitting models show non-significance on the H-L goodness-of-fit test. This desirable outcome of non-significance indicates that the model prediction does not significantly differ from the observed. The H-L statistic assumes sampling adequacy, with a rule of thumb being enough cases so that 95% of cells (typically, 10 decile groups times 2 outcome categories = 20 cells) have an expected frequency > 5.

2.2 Data

Considering the data used in this study, a Quata sample of size n = 55 patients is chosen from "Makah Hospital for Eye Diseases" in Al Riyadh – Khartoum – Sudan , and is composed of a dependent and categorical variable represented in the presence or absence of glaucoma, in addition to five risk factors of : Age, Gender, inheritance, diabetes and hypertension. The categories of the risk factors together with their description are demonstrated in table (1).

Table 1: Risk Factors for Eye Glaucoma

Risk Factors (Code) (Categories)	Description
Age (X1) (1,2,3,4)	Whether the age is less than 20 years, or between 20 & 35, or between 35 & 50 or above 50 years old
Gender (X2) (1,2)	Either male or female
Inheritance (X3) (1,2)	Either presence or absence of the Inheritance factor
Diabetes (X4) 1,(2)	Either presence or absence of Diabetes
Hypertension (X5) (1,2)	Either presence or absence of Hypertension

3. Results

The objectives of this study are to identify the most important risk factors of Eye Glaucoma, to find an appropriate discriminant function that helps in the diagnosis of the disease and to know which risk factors are the most statistically significant. Accordingly, a binary logistic regression analysis is carried out, according to [14]. The results of the study are summarized as follows:

Table (2) presents the results of the logistic regression with the constant only included before any coefficients (i.e. those relating to Age, Gender, inheritance factor, diabetes and hypertension) are entered into the equation. Logistic regression compares this model with a model including all the predictors to determine whether the latter model is more appropriate. The table suggests that if we knew nothing about our variables and guessed that a person would be sick with eye glaucoma, we would be correct 56.4% of the time.

Table 2: Classification Table

Observed Y	Predicted Y [0 (Absence)]	Predicted Y [1 (presence)]	Percentage correct
0	0	24	0.0
1	0	31	100.0
Overall Percentage			56.4

The cut value is 0.500

Table (3) illustrates the variables in the equation, which is the constant term at the moment. It can be realized that the intercept-only model has ln (odds) = 0.256. If we exponentiate both sides of this expression we find that our predicted odds [Exp (B)] = 1.292. That is, the predicted odds of having eye glaucoma is 1.292. Since 31 of the sampled persons have eye glaucoma and 24 are not sick, our observed odds are 31/24 = 1.292. Wald statistic is computed and since it is 0.886, the null hypothesis is accepted, indicating that the constant does not make a significant contribution to the model.

Table 3: Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Constant	.256	.272	.886	1	.347	1.292

The variables not in the equation table tell us whether each IV improves the model, (Table 4). The answer is yes for variables X1 (Age) and X3 (Inheritance) as both are significant and if included would add to the predictive power of the model. The rest of the risk factors seem to be not important at this step, but the overall significance P-value of 0.001 assures that the logistic model will represent the data very much.

Table (5) presents the model chi square value of 27.719, with 5 degrees of freedom, and a probability value of p = 0.000. Thus, the indication is that the model has a poor fit, with the model containing only the constant, while the predictors do have a significant effect and create essentially a different model. So we need to look closely at the predictors and from later tables determine if one or all are significant ones.

Table 4: Variables Not in the Equation

Variables	Score	df	Sig.
X1	16.857	1	0.000
X2	0.245	1	0.620
X3	9.462	1	0.002
X4	0.296	1	0.587
X5	0.111	1	0.739
Overall Statistics	21.436	5	0.001

Table 5: Omnibus Tests of Model coefficients

Step 1	Chi-square	df	Sig.
Step	27.719	5	0.000
Block	27.719	5	0.000
Model	27.719	5	0.000

Although there is no close analogous statistic in logistic regression to the coefficient of determination R_2, the Model Summary in table (6) provides some approximations. In our case its value (Nagelkerke's R_2) is 0.531, indicating a medium relationship between the predictors and the prediction. Under Model Summary the value of the -2 Log Likelihood statistic is 47.634.

Table 6: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	47.634a	.396	.531

Estimation terminated at iteration number 10 because maximum iterations have been reached. Final solution cannot be found.

Table (7) illustrates Hosmer-Lemeshow (H-L) test. The statistic under consideration has a significance of 0.780 which means that it is not statistically significant and therefore leading to the fact that the model is quite a good fit.

Table 7: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.789	8	.780

Considering table (8), it can be realized that it is a classification table, including the constant term and the rest of the predictors. It reveals that 75.0% were correctly classified for the absence of glaucoma while 83.9% for its presence. And overall 80.0% of the cases were correctly classified. This is a considerable improvement on the 56.4% correct classification with the constant model, indicating that the model with predictors is a significantly better one.

Table 8: Classification Table

Observed Y	Predicted Y (no sick)	Predicted Y (sick)	Percentage correct
0	18	6	75.0
1	5	26	83.9
Overall Percentage			80.0

The cut value is 0.500

Table (9) is about the variables that are included in the logistic regression equation. This is illustrated in the following equation:

$$\text{Ln(odds)} = 14.305 + 1.732\text{Age} + 0.192\text{Gender} - 11.012\text{Inheritance} + 0.263\text{Diabities} + 0.565\text{Hypertension} \quad \rightarrow (3)$$

Wald statistic states that the risk factor Age is the only statistically significant factor, while the others are not. Exp (B) of Age factor is approximately 6, indicating that if Age is raised by one year, the person is six times more likely to get sick with eye glaucoma as noted by [15].

Table 9: Variables in the Equation

Step	B	S.E.	Wald	df	Sig.	Exp(B)	L. 95% CI for EXP (B)	U. 95% CI for EXP (B)
X1	1.732	0.593	8.542	1	0.003	5.654	1.769	18.065
X2	0.192	0.843	0.052	1	0.820	1.211	0.232	6.323
X3	-11.012	81.348	0.018	1	0.892	0.000	0.000	2.892E64
X4	0.263	0.952	0.076	1	0.783	1.300	0.201	8.395
X5	0.565	0.850	0.442	1	0.506	1.759	0.333	9.301
Constant	14.305	162.723	0.008	1	0.930	1.632E6		

Finally, the case-wise list of table (10) produces a list of cases that didn’t fit the model well (outliers). If there are a number of cases, this may reveal the need for further explanatory variables to be added to the model. But fortunately, only one case (No. 12) falls into this category and therefore the model is reasonably sound (this is the only person who did not fit the general pattern). We do not expect to obtain a perfect match between observation and prediction across a large number of cases. But no excessive outliers should be retained as they can affect results significantly. The standardized residuals for outliers (ZResid) should be inspected and removed if they exceed > 2.58 (outliers at the .01 level). And in table (10) it is 3.028.

Table:10 Casewise Listb

Case	Selected Status a	Observed Y	Predicted	Predicted Group	Resid	ZResid
12	S	1**	.098	0	.902	3.028

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

4. Conclusion

A binary logistic regression analysis was conducted to predict the persons sick with eye glaucoma. This was for 55 cases selected using Quata sample from the patients in ‘Makah Hospital for Eye Diseases”, using Age, Gender, inheritance, diabetes, and hypertension as predictors. A test of the full model against a constant only model was statistically significant, indicating that the predictors as a set reliably distinguished between sick and non-sick of Glaucoma (chi-square 27.719, p-value 0.000 with df 5). Nagelkerke’s R_2 of 0.531 indicated a medium relationship between prediction and grouping. Prediction success overall was 80% , classified into 75.5% for glaucoma absence and 83.9% for its presence. The Wald criterion demonstrated that only age factor made a significant contribution to prediction (p-value 0.003). Gender, inheritance, diabetes, and hypertension were not significant predictors. EXP (*b*) value indicates that when age is raised by one year, the person is six times more likely to get sick with eye glaucoma. Thus the only statistically significant factor is the age, as expected , while the other factors are considered as clinically important.

One constraint of this study is revealed on the use of the non- probability Quato Sampling Technique, which is limited in size, and which was used to facilitate the process of data collection through time saving.

The recommendations of the study are summarized in two points; Fisrstly, Equation (3) is preferred to be used in the needed situations to help in deciding for the absence or the presence of the disease. Secondly, for future work , selection of data on basis of a more complex probabilty sample design is recommended so as to be able to enlarge the sample size , and to be able to get more presice and reliable results.

References

[1] Agresti, A. A.. *An introduction to categorical data analysis*. New York : John Wiley & Sons, 1996, pp. 156-178.

[2] Agresti, A.. *Categorical data analysis, 2nd edition*. New York : John Wiley & Sons, 2000, pp. 52-143.

[3] Long, J. S. *Regression models of categorical and limited dependent variables*. Sage, Thousand Oaks, CA, 1997, pp. 96-132.

[4] Peng, C. Y., Manz, B. D., & Keck, J. "Modeling categorical variables by logistic regression". *American Journal of Health Behavior*, vol. 25, pp. 278–284, 2001.

[5] Hosmer, David W and Stanley Lemeshow. *Applied Logistic Regression*. New York : Wiley Series, 1989, pp. 145-196.

[6] Hosmer, David W, Scott Taber, and Stanley Lemeshow. "The importance of Assessing the Fit of Logistic Regression Models: A Case Study". *American Journal of Public Health* , vol. 81, pp. 30-35, 1 991.

[7] Hosmer, D. W. and Lemeshow, S. *Applied logistic regression, 2nd edition*. New York : John Wiley & Sons, 2000, pp. 79-125.

[8] Lei, P.-W., & Koehly, L. M. "Linear discriminant analysis versus logistic regression: A comparison of classification errors". Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA., 2000.

[9] Peng, C. Y., & So, T. S. H. "Logistic regression analysis and reporting: A primer". *Journal of Understanding Statistics*, vol. 1, pp. 31–70, 2002

[10] Hosmer, T., D.W. Hosmer and L.L. Fisher. "A comparison of the maximum likelihood and discriminant function estimators of the coefficients of the logistic regression model for mixed continuous and discrete variables". *Journal of Communications in Statistics*, vol. 12, pp. 577-593, 1983.

[11] Hauck, W.W. and A. Donner. "Wald’s Test as applied to hypotheses in logit analysis". *Journal of the American Statistical Association*, vol. 72, pp. 851-853, 1977.

[12] Cleary, P. D., & Angel, R. "The analysis of relationships involving dichotomous dependent variables". *Journal of Health and Social Behavior*, vol. 25, pp. 334–348, 1984.

[13] Nagelkerke, N.J.D. "A note on the general definition of the coefficient of determination". *Journal of Biometrika*, vol. 78, pp. 691-692, 1991.

[14] Cox, D.R. and E.J. Snell,. *The Analysis of Binary Data, 2nd edition*. Chapman and Hall, London, 1989, pp. 144-163.

[15] http://www.webmd.com/eye-health/glaucoma-eyes. (Retrieved on 6/11/2013)