MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Mini Review

Open Access

CrossMark

Application of GLM (logistic regression) on serological data of malaria infection

Volume 8 Issue 1 - 2019

Getachew Tekle
Department of Statistics, Wachemo University, Ethiopia

Correspondence: Getachew Tekle, MSc. in Biostatistics, Department of Statistics, Wachemo University, Ethiopia, Email getch.55ekle@gmail.com

Received: December 11, 2018 | **Published:** January 09, 2019

Introduction

As a nation reduces the burden of falciparum malaria, identifying areas of transmission becomes increasingly difficult. Over the past decade, the field of utilizing malaria serological assays to measure exposure has grown rapidly, and a variety of serological methods for data acquisition and analysis of human IgG against falciparum antigens are available.[1]

The main Objective of this case study is to model the probability of infection as a function of age (the prevalence of malaria infection).

Methods

Variables

The predictor variable (age) is continuous and the dependent variable serology/ disease status is binary (where, sero-positive or sero-negative).

Data: Serological data of malaria

Serology is the scientific study of plasma serum and other bodily fluids. In practice, the term usually refers to the diagnostic identification of antibodies in the serum. Serological tests may be performed for diagnostic purposes when an infection is suspected, in rheumatic illnesses, and in many other situations, such as checking an individual's blood type.[2]

Antibodies produced in response to an infectious disease like malaria remain in the body after the individual has recovered from the disease. A serological test detects the presence or absence of such antibodies. An individual with such antibodies is termed sero-positive.

A sample which has taken at a certain time point, the information for each individual:

A. Age at test.

B. Infected or not.

   a. Prevalence of sero-positivity in the sample: This is the probability to become infected before the age at test.

   b. In this example the information about each subject in the experiment is the disease status (infected or not by malaria) and the age group of the subject.

   c. The variables are: the sample size, the number of sero-positive at each sample size (=the number of infected subjects) and the age.

Binary data

Binary data may occur in two forms:

Ungrouped in which the variable can take one of two values, say success/failure. Grouped in which the variable is the number of successes in a given number of trials.

a. The natural distribution for such data is the Binomial (n, p) distribution; where in the first case n = 1.

b. The observation is a binary variable which takes the value of 1 with probability P.

$$P = \frac{e^{\alpha+\beta \quad age}}{1+e^{\alpha+\beta \quad age}} \tag{2.1}$$

c. The probability of infection.

d. If $\beta > 0$ then there is a positive association between the probability and age. This means that the probability of infection increase with age.

e. If $\beta > 0$ then there is a negative association between the probability and age. This means that the probability of infection decrease with age.

Generalized linear models (GLM)

Generalized linear models (GLM) are used to fit fixed effect models to certain types of data that are not normally distributed. Generalized– not limited to normally distributed data. Linear–models use a linear combination of variables to 'predict' the response. Exponential family of Binomial distribution, Dobson.[3]

$$Z_i = \begin{cases} 1 \\ 0 \end{cases} \Rightarrow Y_i = \sum_{i=1}^{n} Z_i \Rightarrow Y_i \sim B(n, \pi_i)$$

$$p(y_i \mid \theta) = \exp\left\{ y_i \log\left[\frac{\theta_i}{1-\theta_i}\right] + n_i \log(1-\theta_i) + \log\binom{n_i}{y_i} \right\}$$

The link function

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

$$a_i(\varphi) = 1, \quad b(\theta_i) = \log(1+\exp(\theta_i))$$

$$c(y) = \log\binom{n_i}{y_i}$$

$$\log\left(\frac{\mu}{1-\mu}\right)=\log\left(\frac{\frac{e^{\theta}}{1+e^{\theta}}}{\frac{1}{1+e^{\theta}}}\right)=\log\left(e^{\theta}\right) \quad \Leftarrow \quad \begin{matrix} E(y)=\mu=b'(\theta_i)=e^{\theta_i}(1+\exp(\theta_i))^{-1} \\ var(y)=\mu(1-\mu)/n \end{matrix} \qquad (2.2)$$

Components of GLM

a. Random component- the probability distribution of the response.

b. Systematic component (linear predictor)-the predictor variables are (e.g., X1, X2, etc). These variables enter to the model in a linear manner.

c. Link function-Specify the relationship between the mean random component (i.e., E(Y)) and the systematic component.

Random component

$Y_{ij}=\begin{cases}1\ seropositive\\0\ seronegative\end{cases}$ then $E(Y_{ij})=P(Y_{ij}=1)=\pi_{ij}$ which will also be $\sum\frac{Y_{ij}}{n_j}$, where $Z_i=\sum Y_{ij}$

To show the sum of Bernollis is binomially distributed, $Z_i=\begin{cases}1\ seropositive\\0\ seronegative\end{cases}$ and

$$Z_i=\sim Bin(1,\pi_{ij}) \qquad (2.3)$$

$Z_i=\sum Y_{ij}$ Vs $Z_i=Bin(n_i,\pi_{ij})$

Number of sero-positive at each age group n_i: sample size at each age group

P_i is the probability to be infected (the prevalence). We use logistic regression in order to model the prevalence as a function of age.

Systematic component: - dependency of the predictor – the linear predictor

The systematic component of the model consists of a set of explanatory variables and some linear function of them.[4]

$$\pi_j=f(seropositive_i)=f(S_i)\leftrightarrow\pi_j=f(S_i)=f(\beta_0+\beta_1 S_i) \qquad (2.4)$$

Binomial link functions

Logit link function: $n(p)=b(\frac{p}{1-p})$

$\upsilon=\frac{\exp(X\beta)}{1+\exp(X\beta)}=\frac{1}{1+\exp(-X\beta)}$, mean of the response with logit link

Probit link function: $\eta(p)=\varphi^{-1}(p)$

Complementary log log function: $\eta(p)=\ln(-\ln(1-p))$

Analysis of designed matrices

a. For logistic regression

Define a (design) matrix X so that for response variable $Y: E(Y)=X\beta$; Where β is a vector of parameters and X is a design matrix of predictors.

b. For binomial model

$E(Y)=X\beta$; Where β is a vector of parameters and X is a design matrix of predictors.

Model Selection Techniques

The most commonly known model selection criteria are Akaike Information Criterion (AIC) (Sakamoto, 1986), and Log-likelihood were used.

$AIC = -2\log L + 2p$

Where, -2 log L is twice the negative log-likelihood value for the model

P: - is the number of estimated parameters.

Smallest value of AIC, best is the model.

Results and Discussions

Exploratory analysis of data

The above plot indicates the prevalence of malaria infection will be increased with age, as age increases the probability of infection will increases. Thus, there is almost a linear relationship among the probability of malaria infection and age (Figure 1). The line indicates the fitted proportion of infection linearly as given below:

$$\log it(\hat{pi})=-2.71+0.044*age \qquad (3.1)$$

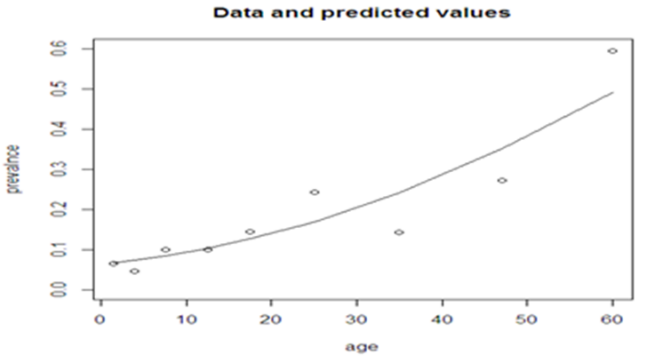

Figure 1 Plot of prevalence of malaria vs. age, posi/N.

Model Diagnosis

As the above plot describes, there is a pattern the residuals fit and the residuals are not constant through fitted values; the variation among the predicted probability of infection is not the same. Thus, it indicates some assumption/constant variance of the model has not been satisfied (Figure 2).

The above normal plot shows that the normality assumption has been satisfied (Figure 3).

Citation: Tekle G. Application of GLM (logistic regression) on serological data of malaria infection. *Biom Biostat Int J.* 2019;8(1):1–4.
DOI: 10.15406/bbij.2019.08.00261

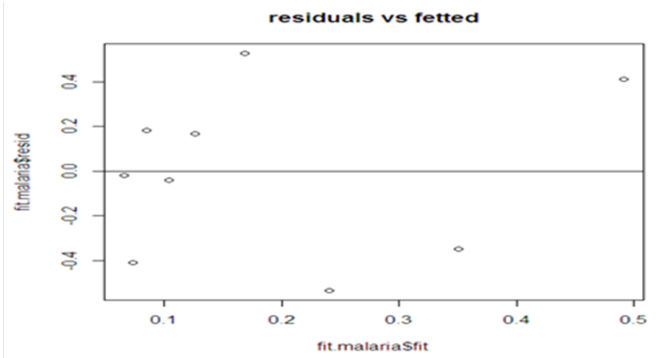

Figure 2 Plot of residuals vs. Fitted values.

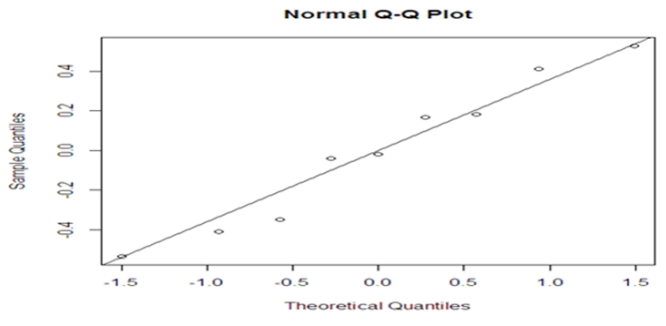

Figure 3 Normal plot.

Models with different link functions

Model with logit link

Deviance Residuals:

a) Min 1Q Median 3Q Max

b) $-2.78685 \ -1.31863 \ -0.05053 \ 0.66752 \ 2.38275$

c) Coefficients: Estimate Std. Error z value $\Pr(>|z|)$

d) (Intercept) $-2.714074 \ 0.151740 \ -17.886 \ < 2e-16$ ***

e) agei $0.044672 \ 0.004511 \ 9.904 \ < 2e-16$ ***

f) Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

g) (Dispersion parameter for binomial family taken to be 1)

h) Null deviance: 124.037 on 8 degrees of freedomResidual deviance: 21.865 on 7 degrees of freedom

i) AIC: 66.388

Complementary log log or (c-log-log) link:

Deviance Residuals:

a) Min 1Q Median 3Q Max

b) $-2.6301 \ -1.3864 \ -0.1393 \ 0.6994 \ 2.5276$

c) Coefficients: Estimate Std. Error z value Pr(>|z|)

d) (Intercept) $-2.709235 \ 0.139261 \ -19.45 \ < 2e-16$ ***

e) agei $0.039671 \ 0.003746 \ 10.59 \ < 2e-16$ ***

f) Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

g) (Dispersion parameter for binomial family taken to be 1)

h) Null deviance: 124.037 on 8 degrees of freedom

i) Residual deviance: 20.658 on 7 degrees of freedom

j) AIC: 65.181

Model with log link:

a) Deviance Residuals:

b) Min 1Q Median 3Q Max

c) $-2.428 \ -1.474 \ -0.146 \ 0.751 \ 2.682$

d) Coefficients: Estimate Std. Error z value $\Pr(>|z|)$

e) (Intercept) $-2.699659 \ 0.126483 \ -21.34 \ < 2e-16$ ***

f) agei $0.034705 \ 0.002997 \ 11.58 \ < 2e-16$ ***

g) Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

h) (Dispersion parameter for binomial family taken to be 1)

i) Null deviance: 124.037 on 8 degrees of freedom

j) Residual deviance: 19.312 on 7 degrees of freedom

k) AIC: 63.836

Model with Identity link:

a) glm(formula = dew ~ agei, family = binomial(link = "identity"))

b) Deviance Residuals:

c) Min 1Q Median 3Q Max

d) $-3.2921 \ -0.8959 \ -0.1462 \ 0.8583 \ 3.0276$

e) Coefficients:

f) Estimate Std. Error z value Pr(>|z|)

g) (Intercept) $0.0381457 \ 0.0123993 \ 3.076 \ 0.00209$ **

h) agei $0.0063542 \ 0.0006656 \ 9.547 \ < 2e-16$ ***

i) Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

j) (Dispersion parameter for binomial family taken to be 1)

k) Null deviance: 124.037 on 8 degrees of freedom

l) Residual deviance: 26.165 on 7 degrees of freedom

m) AIC: 70.689

Models Comparison

Selection of terms for deletion or inclusion is based on Akaike's information criterion (AIC). In R, the function "extractAIC(model) will give AIC (Table 1). According to the AIC criteria and Likelihood, the model with log link function will be chosen as a good model; though its mean estimate is the second smallest next to identity, its AIC and Likelihood are the smallest of all. Hence, the chosen model with the log link function should be given as follows:

$$\hat{P}_i = \frac{e^{2.699659+0.034705*age}}{1+e^{-2.699659+0.034705*age}}$$

$E(Y) = -2.699659 + 0.034705 * \text{age}$, which indicates that for a unit increase in age since at infection, the proportion of developing the antibiotics will increase by 0.0347(3.5%).

Table 1 Model comparison

Model	Estimate (β)	Likelihood	No. parameters	AIC
Logit	0.044672	-31.1941	2	66.388
Logit	0.034705	-29.9179	2	63.836
Identity	0.006354	-33.3445	2	70.689
C-log-log	0.039671	.30.59063	2	65.181

The odds ratio: point estimator

How to calculate the odds ratio? For continuous predictor the odds ratio is given by $\theta = \exp(\beta)$. The meaning of a logistic regression coefficient is not as straightforward as that of a linear regression coefficient. While B is convenient for testing the usefulness of predictors, exp (B) is easier to interpret. Exp (B) represents the ratio-change in the odds of the event of interest for a one-unit change in the predictor. Exp (0.0347) =1.0353, in this case the odds for malaria infection in sero-positive people is 0.035(3.5%) times the odds for malaria infection in sero-negative people.[5]

Conclusion

Serological data is explored and analyzed as is shown above. From the summary part it is indicated that in all models fitting, the p-value is very small and the predictor variable age is significant for the prediction of the prevalence of malaria. Comparison of the four models indicated that the model with log link function is chosen as the best model based on AIC criteria, in which case the predicted value of model coefficient is 0.0347, which indicates for a unit increase in mid age the proportion of malaria infection will increase by 0.0347.

Acknowledgments

None.

Conflicts of interest

Author declares that there is no conflict of interest.

References

1. Eric Rogier, Wiegand R, Moss D, et al. Multiple comparisons analysis of serological data from an area of low Plasmodium falciparum transmission. *Malaria Journal*. 2015;4(14):436.
2. Collet D. *Modeling Binary Data*. London: Chapman & Hall; 1991.
3. Dobson AJ. *An Introduction to Generalized Linear Models*. 2nd edn. London: Chapman &Hall; 2001.
4. McCullagh P, JA Nelder. *Generalized Linear Models*. London 2nd edn. London: Chapman & Hall. 1989.
5. Lindsey JK, G Mersch. Fitting and comparing probability distributions with log linear models. *Comput Statist Data Anal*. 1992;13:373–384.