Mikailalsys Journal of Mathematics and Statistics

e-ISSN : 3030-816X
p-ISSN : 3030-8399

Index : Harvard, Boston, Sydney University, Dimensions, Lens, Scilit, Semantic, Google, etc

https://doi.org/10.58578/MJMS.v3i2.5145

Fitting Linear Probability Model and Logit to Prevalence of Hepatitis B and C Data in Donga Local Government Area of Taraba State

Adeniyi Oyewole Ogunmola & Garsama Sambo
Federal University Wukari, Taraba State, Nigeria
ogunmola@fuwukari.edu.ng

Article Info:

Submitted:	Revised:	Accepted:	Published:
Feb 12, 2025	Feb 24, 2025	Mar 9, 2025	Mar 14, 2025

Abstract

Hepatitis B and C are significant global public health concerns, responsible for a substantial burden of liver disease. The prevalence of hepatitis B and C varies widely across different populations and regions, influenced by factors such as age, sex, geographic location, and risk behaviors. This study focuses on examining the effects of age and sex on the prevalence of hepatitis B and C. By analyzing data on the presence or absence of these infections across different age categories and between sexes, we aim to identify patterns that could inform targeted public health efforts. Linear probability model and logit through generalised linear model were fitted on the data collected at the first referral hospital laboratory Donga. Results showed that both models fit the data, and significant factors are age category and the interaction of age category and sex. But it is discovered that the logit model fitted the data more with a lower value of AIC and BIC.

Kewords: Hepatitis B, Hepatitis C, Linear Probability Model, Logit and Generalised Linear Model

INTRODUCTION

Hepatitis, a significant global health concern, represents a group of viral infections that impact the liver and vary in severity from acute to chronic conditions (WHO 2024, CDC 2021). The prevalence of hepatitis B and C varies widely across different populations and regions, influenced by factors such as age, sex, geographic location, and risk behaviors. Anaedobe et. al, (2015) also studied prevalence, demographic features and risk factors of HBV among pregnant women. They employed descriptive statistics, used Chi square odd ratios to determine significance multiple sexual partners and early age at sex debut. Mohammed et al, (2022), studied prevalence and risk factors associated with HBV. They used Chi-square test for proportion to determine significance between genders, and among age categories.

Prevalence data, typically binary in nature—indicating the presence or absence of hepatitis in individuals—necessitate the use of specialized regression techniques for accurate modeling. Two commonly employed approaches for analyzing such data are the Linear Probability Model (LPM) and Logistic Regression (logit model). In this study, we explore the application of both the Linear Probability Model and Logistic Regression to prevalence data on hepatitis. By comparing these two approaches, we aim to highlight the strengths and limitations of each in the context of modeling binary health outcomes, and to provide insights into their practical use in epidemiological research.

METHODOLOGY

Data for this study is secondary data collected from the laboratory of First Referral Hospital Donga in Taraba state. Patients are tested for infectious diseases including Hepatitis. Hepatitis test outcome for patients are either positive or negative. Two predictors are considered based on the available data; age and sex of patients. Patients' age are categorized. To predict positive outcome of the test for when the patient's age category and sex is known, linear probability model (LPM) is first fitted to the data.

LPM Specification:

The Linear Probability Model for this study is specified as

$$Hepatitis\ present = \beta_0 + \beta_1 Age1 + \beta_2 Age2 + \beta_3 Age3 + \beta_4 Male + \varepsilon$$

where Age1 is age category 0-14 years, Age2 is age category 15-24 years, Age3 is age category 25=64 years, while Age category 65 years and above is considered as the baseline for the age category variable and so, it is not explicitly included in the model. For sex category, female sex is the baseline. The parameters $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 would be estimated using the maximum likelihood estimation method.

Logit model Specification:

Fitting logit model to the data, the model for the study is specified as

$$ln\left(\frac{P(y=1|\boldsymbol{X})}{1-P(y=1|\boldsymbol{X})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Where X_1 is the age categorical variable, and X_2 is the sex categorical variable. $P(y=1|\boldsymbol{X})$ is the probability of positive test outcome given a specified age category and sex category. $\ln\left(\frac{P(y=1|\boldsymbol{X})}{1-P(y=1|\boldsymbol{X})}\right)$ is the log odds ratio which is a linear function of the independent variables.

The Generalised Linear Model perception of the LPM and the Logit model:

For the random component; $P(hepatitis\ presence) = \mu$ and $P(hepatitis\ absence) = 1-\mu$

For the systematic component; the linear predictor is XB, the linear combination of the independent categorical variables.

The Link function is the identity function, which is

$$g(\mu) = \mu = X\beta$$

That is, the mean of the outcome of hepatitis test which is the probability of the disease being present in a patient is directly modeled as a linear function of the age categorical and sex categorical independent variables.

For the logit model, the GLM uses logit link function, given as

$$g(\mu) = ln\left(\frac{\mu}{1-\mu}\right) = \boldsymbol{X\beta}$$

AIC and BIC are used to determine the model that performs better between LPM and logit.

RESULTS

For each model, the Wald Chi-square test result for the effect of the model is shown in Table 1 below. The estimated parameters values with their Wald test statistic values and corresponding p-values are given. In both the LPM and logit model, all the independent variables are significant at 5 percent level of significance.

Table 1: Test of model effect

LPM					Logit Model				
Source	$\boldsymbol{\beta}$	**Wald X^2**	**df**	**P-value**	**Source**	$\boldsymbol{\beta}$	**Wald X^2**	**df**	**P-value**
Intercept	1.836	87364.7	1	0.000	Intercept	1.645	2829.453	1	0.000
0-14 Years	0.108	74.128	1	0.000	0-14 Years	1.080	87.513	1	0.000
15-24 Years	0.029	13.958	1	0.000	15-24 Years	0.211	20.930	1	0.000
25-64 Years	0.049	43.987	1	0.000	25-64 Years	0.378	54.376	1	0.000
Male	-0.027	21.014	1	0.000	Male	-0.219	34.352	1	0.000
Scale	0.123								

For LPM, the estimated equation is:

Probability that hepatitis is present given Age and sex categories of the patient = 1.836 + 0.180 presence of 0-14 years old of patient + 0.029 presence of 15-24 years old of patient + 0.049 presence of 25-64 years old of patient – 0.027 presence of male patient.

This means, compared to the baseline patient who is at 65 years and above, when a patient of age 0-14 years is tested for hepatitis, the probability of the patient will be positive increases by 0.180; for a patient within the age 15-24 years, the chances of being positive will increases by 0.029; and for a patient within the age 25-64 years, the chances of being positive to hepatitis increases by 0.049. Compared to a female patient as a baseline, a male patient tested for hepatitis, the chance that the male patient will be positive decreases by 0.027.

For Logit model, the estimated equation is:

The log odds of a patient tested positive to hepatitis is 1.645 + 1.080 presence of 0-14 years old of patient + 0.211presence of 15-24 years old of patient + 0.378 presence of 25-64 years old of patient – 0.219 presence of male patient.

This means, compared to the baseline patient who is at 65 years and above, the log odds when a patient of age 0-14 years is tested for hepatitis increases the log odds by 1.080; for a patient within the age 15-24 years, the log odds of being positive will increases by 0.211; and for a patient within the age 25-64 years, the log odds of being positive to hepatitis increases by 0.378. Compared to a female patient as a baseline, a male patient tested for hepatitis, the chance that the male patient will be positive decreases by 0.219

To determine the model that performs better between the LPM and the logit model, result of Akaike Information Criterion (AIC) and Bayesian Information Criteria are given in Table 2 below. Both AIC and BIC values for the logit model are much smaller than those of the LPM. The logit model performs better than the LPM in modeling the outcome of Hepatitis test screening in the hospital.

Table 2: AIC and BIC values for LPM and Logit Model

Information Criterion	**LPM**	**Logit Model**
AIC	10575.161	85.538
BIC	10620.549	123.362

CONCLUSION

Both LPM and logit model fit the data. But logit model far performs better than the LPM. A patient within the age category 0-14 years is most likely to test positive to hepatitis in that hospital and locality. A female patient is more likely to test positive to hepatitis compared to a male patient in that hospital and locality.

Recommendation

Between LPM and logit model for binary outcomes, it is strongly recommended that logit model be fitted to the data than LPM. It is clear that LPM modeling power is less than that of logit model.

Age category 0-14 years who suffered the highest rate of being tested positive requires that more investigation be done to ascertained that the it is just the frequency due to the age rather than the a high level factor of contracting hepatitis.

REFERENCES

Anaedobe C.G., Fowotade A., Omoruyi C.E., Bakare R.A. (2015). Prevalence, socio-demographic features and risk factors of Hepatitis B virus infection among pregnant women in Southwestern Nigeria. Pan Afr Med J. 2015; 20: 406.

CDC (2024). Global Hepatitis: Millions of people are affected. https://www.cdc.gov/hepatitis/globa/index.html September 15th, 2024

Mohammed H., Eshetie A., Melese D. (2022). Prevalence of hepatitis B virus and associated risk factors among adults patients at Dessie referral and Kemise general hospitals in northeastern Ethiopia. Health Sci. Rep. 2022 May 22;5(3):e659. doi:10.1002/hsr2.659. eCollection 2022 May.

World Health Organisation (WHO) (2024). Hepatitis B. https://www.who.int/news-room/fact-sheets/detail/hepatitis-b April 9th, 2024.