On Statistical Analysis of Blood Pressure with respect to Some Demographic Factors

Babalola B.T[1]

Olubiyi A.O[2]

Abstract

This research examined the relationship that exists between the between the blood pressure and some demographic fators viz: age , sex and status(alive or dead) of some patients. Ordinal logistic regression was employed for the analysis of the data. The data of 1,966 patients(men and women) were obtained from Ekiti State Teaching Hospital, Ado Ekiti. The result showed that only age is statistically significant of the three factors. Pseudo R-square (Cox and Snell, Nagelkerke and McFadden) values were very low (0.005, 0.006, 0.002 respectively), suggesting that these predictors can only account for a little change in blood pressure patients. Logit models were obtained to calculate probabilities of the various possible outcomes.

Keyword: Blood pressure, Patients, Regression and Age, Hypertension

Introduction

Blood pressure, sometimes referred to as arterial blood pressure exerted by circulating blood upon the walls of the blood vessels. Hypertension which is also called high blood pressure is elevated pressure of the blood in the arteries. It is one of the most common worldwide diseases afflicting humans. Because of the associated morbidity and mortality and the cost to society, hypertension is an important public health challenge. Over the past several decades, extensive research, widespread patient education and a concerted effort on the part of the health care professionals have led to decreased mortality and morbidity rate from the multiple organ damage arising from years of untreated hypertension.

Hypertension is the most important modifiable risk factor for coronary heart disease, stroke, congestive heart failure, end-stage renal disease and peripheral vascular disease. Hypertension results from two major factors which can be presently independent or together. When the heart pumps blood with excessive force and if the the body's smaller blood vessels (known as the arterioles) narrow, so that blood flow exerts more pressure against the vessels' walls. Stress and other behavioral factors have linked to a board range of cardiovascular disease outcome. Two numbers are used to describe blood pressure: the systolic pressure (the higher and first number) and the diastolic pressure (the lower and second number). Health dangers from blood pressure may vary among different age groups and depending on whether systolic or diastolic pressure (or both) is elevated.

A third measurement, pulse measure may also be important as an indicator of severity.Based on the recommendations of the Joint National Committee (JNC 7) on prevention, detection, evaluation and treatment of high blood pressure, classification of blood pressure (BP) (expressed in mmHg) for adults aged 18years or older is as follows; Normal:

[1] Department of Mathematical and Physical Sciences, Afe Babalola University,Ado Ekiti, Nigeria.
[2] Department of Mathematical Sciences, Ekiti State University, Ado Ekiti, Nigeria.

Systolic lower than 120mmHg,Diastolic lower than 80mmHg.Prehypertension: Systolic 120mmHg-139mmHg,Diastolic 80mmHg - 89mmHg.Stage 1: Systolic 140mmHg - 159mmHg,Diastolic 80mmHg - 89mmHg.Stage 2: Systolic 160mmHg or greater, Diastolic 100mmHg or greater. In this work, we made use of logistic regression analysis to test if being hypertensive is dependent on sex, age and if it determines whether a person will live or die. We combined patients in Stages 1 and 2 together as hypertensive. Consequently, we have patients who are hypotensive, normal, prehypertensive and hypertensive.

The data used in this project work are secondary data collected from the health records department of Ekiti State University Teaching Hospital, Ado Ekiti, Ekiti State from the year 2002 to 2011. Logistic regression was used to obtain regression equations and Wald's statistic was used to test whether the predictors were statistically significant.

Material and Methods

Logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable (a dependent variable that can take on a limited number of categories) based on one or more predictor variables. The probabilities describing the possible outcome of a single trial are modeled as a function of explanatory variables using a logistic function. Logistic regression measures the relationship between categorical dependent variable and usually a continuous independent variable, by converting the dependent variable to probability scores. Logistic regression can be binomial or multinomial.

Binomial or binary logistic regression refers to the instance in which the observed outcome can have only two possible types (example: "dead" or "alive"; "yes" or "no"). Multinomial logistic regression refers to cases where the outcome can have three of more possible types (example: "better" vs. "no change" vs. "worse"). Like other forms of regression analysis, logistic regression makes use of one or more predictor, variables that may either be continuous or categorical data. Also, like other linear regression models, the expected value of a Bernoulli distribution is simply the probability of a case. In other words, in logistic regression, the base rate of a case for the null model is fit to the model including one or more predictors.

Unlike ordinary linear regression, however, logistic regression is used in predicting binary outcomes rather than continuous outcomes. Given this difference, it is necessary that logistic regression take the natural logarithm of the odds (referred to as a logit) to create a continuous criterion. The logit of success is then fit to the predictors using regression analysis. The results of the logit are not intuitive, so the logit is converted back to the odds via exponential function or the inverse of the natural logarithm. The logit is referred to as the link function in logistic regression, although the output in logistic regression is binomial and displayed in a contingency table, the logit is an underlying continuous criterion upon which linear regression is conducted.

Logistic Function

An explanation of logistic regression begins with an explanation of the logistic function, which always takes on values between 0 and 1.

$$\pi(x) = \frac{exp^{(\beta_0+\beta_1 x)}}{1 + exp^{-(\beta_0+\beta_1 x)}}$$

$$= \frac{1}{1 + exp^{-(\beta_0+\beta_1 x)}}$$

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

$$= \beta_0 + \beta_1 x$$

$$\frac{\pi(x)}{1-\pi(x)} = exp^{(\beta_0+\beta_1 x)}$$

The input is $\beta_0 + \beta_1 x$ and the output is $\pi(x)$. The logistic function is useful because it can take as an input any value from negative infinity to positive infinity whereas the output is confined to values between 0 and 1. In the above equation,

- $g(x)$ refers to the logit function of some given predictor X
- ln denotes the natural logarithm
- $\pi(x)$ is the probability of being a case
- $\beta_0$is the intercept from the linear regression equation
- $\beta_1 x$ is the regression coefficient multiplied by some value of predictor
- Base $(\exp)$ denotes the exponential function.

The first formula illustrates that the probability of being a case is equal to the odds of the exponential function of the linear regression equation.

Model Fitting

Maximum Likelihood Estimation:

The regression coefficients are usually estimated using maximum likelihood estimation. Unlike linear regression with normal distribution residuals, it is not possible to find a closed-form expression for the coefficient values that maximizes the likelihood function, so an iterative process must be used instead. In some instances, the model may not reach convergence. When a model does not converge, this indicates that the coefficients are not meaningful because the iterative process was not able to find appropriate solutions. A failure to converge may occur for number of reasons: multicollinearity, sparseness or complete separation.

Deviance and Likelihood Ratio Test:

In linear regression analysis, one is concerned with partitioning variance via the sum of squares calculation, variance in the criterion is essentially divided into variance accounted for by predictors and residual variance.

In logistic regression analysis, deviance is used in lieu of sum of squares calculations. Deviance is analogous to the sum of squares calculations in linear regression and is a measure of the lack of fit to the data in a logistic regression model. Deviance is calculated by comparing a given model with the saturated model (a model with a theoretically perfect fit). This computation is called the likelihood ratio test.

$$D = -2ln\left[\frac{(likelihood\ of\ the\ fitted\ model)}{(likelihood\ of\ the\ saturated\ model)}\right]$$

In the above equation,

- D represents deviance
- ln represents natural logarithm

The results of the likelihood ratio will produce a negative value, so the product is multiplied by negative two times its natural logarithm to produce a value with an approximate "chi-square distribution."

Regression Coefficient

After fitting the model, it is likely that researchers will want to examine the contribution of individual predictors. To do so, they will want to examine the regression coefficients. In linear regression, the regression coefficient presents the change in the criterion for each unit change in the logit for each unit change in the predictor. In logistic regression, there are a couple of different tests designed to assess the significance of an individual predictor, most notably, the likelihood ratio test and the Wald statistic.

Logistic Regression Model

The linear logistic model assumes a dichotomous dependent variable, that is the dependent variable can take the value of 1 with a probability of success π, or the value of zero with a probability of failure $1 - \pi$.

$$Y = \frac{\pi}{1 - \pi}$$

Where:

π Is hypertensive

$1 - \pi$ Is not hypertensive

The general multiple regression model for p predictors is:

$$Y(\theta) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

For the sake of this project work, we shall use:

$$Y(\theta) = \alpha_i + \beta_1 (age)_i + \beta_2 (sex)_i + \beta_3 (outcome)_i$$

Where:

$Y(\theta)$= the independent variable. Which is the natural logarithm of odds ratio of $y_i = 1$ versus $y_i = 0$. That is, the log of the odds of patients having hypertension relative to not having hypertension.

α_i= intercept or treshhold. That is, the log of the odds of a patient having hypertension relative to not having hypertension when all the predictors are zero.

β_1= the partial regression coefficient for x_1, holding the remaining predictors constant. That is, the change in the log of the odds of a patient having hypertension relative to not having hypertension for a change in age, holding sex and outcome (or status) of the patient constant.

β_2= the partial regression coefficient for x_2, holding the remaining predictors constant. That is, the change in the log of the odds of a patient having hypertension relative to not having hypertension for a single unit increase in sex, holding age and outcome (or status) of the patient constant.

β_3= the partial regression coefficient for x_3, holding the remaining predictors constant. That is, the change in the log of the odds of a patient having hypertension relative to not having hypertension for a single unit change in the outcome (or status: i.e. dead/alive), holding age and sex constant.

The variables are coded as follows: Y(Response):0- hypertensive,1-prehypertensive,2-normal and 3-hypotensive.X_1(Age): 0-≤ 40,1- >40. X_2(Sex):0- Male, 1-Female. X_2(Status): 0- Dead, 1-Alive. The whole regression can be viewed as, Hypertensive vs Prehypertensive vs Normal Blood Pressure vs Hypotensive.

Results and Discussion

In this section, we discussed the data analysized based on some statistical tools.

Overall Model Test

The table 1 gives the result for our imputed data

H_0: Model 1= Model 2

H_1: Model 1≠ Model 2

Where Model 1 is a model without any predictor and Model 2 is a model with predictors. Since the chi- square value which is 10.054> 5.991($X^2_{(0.05,3)}$) have significant level which is P= 0.018<0.05, we then reject the null hypothesis that the model without predictor is as good as model with predictors (i.e. we reject that all independent variables are equal to zero).

Goodness of Fit Test

We can test the goodness of fit using the result obtained from our analysis provided in table 2 compared to the tabulated value of chi- square. Good models have large observed significance levels. Since the goodness of fit measures have large observed significant level that is 0.479 (Pearson) and 0.344 (Deviance) are relatively high when compared with the normal significant level (0.05), coupled with the fact that X^2 =17.646< $X^2_{(0.05,18)}$ =28.869 and P=0.05<0.479 for Pearson and X^2 =19.808< $X^2_{(0.05,18)}$ =28.869 and P=0.05<0.344=P-value for Deviance measure. So the model fits.

Test of Parallelism

When we fit a logistic regression, we assume that the relationship between the independent variables and the logit are the same for all logits. This assumption is tested below. Table 3 shows the result of the test for the model. The row labeled null hypothesis contains -2log likelihood for the constrained model, the model that assumes the line or plane are parallel. The row labeled general is for the model with separate planes or lines. The entry labeled chi- square is the difference between the two -2log likelihood value.

The X^2 value =1.784 compared with $X^2_{(0.05,6)}$ =12.592 and P=0.05<P-value=0.938 which shows that we do not have significant evidence to reject null hypothesis (which states that there is no significant difference between the regression coefficient across the response categories), suggesting that the model assumption of equality is satisfied. This shows that we have chosen the right link function.

Test of regression parameters

Here we are going to use the Wald's Statistic to find out which of the predictors are significant to the outcome blood pressure. The Wald's Statistic is the square of the ratio of the coefficient to its standard error based on the observed significant level.

Tests for Parameter Significance

We are going to use the Wald's Statistic to find out which of the predictors are significant to the outcome blood pressure. The Wald's Statistic is the square of the ratio of the coefficient to its standard error based on the observed significant level.

Age

H_0: Age=0 (hypertension is independent of age ≤40)

H_1: Age≠0 (hypertension is dependent on age > 40)

Let's compare Z^2 =8.297, which is the Wald's value with the chi-square value of $X^2_{(0.05,1)}$=3.841 and P-value of 0.004. The values obtained showed that we need to reject the null hypothesis. This means that the statement that "age has no influence on blood pressure" is false and we conclude that age has a huge influence on blood pressure.

Sex

We are to test whether sex has an influence on blood pressure or not. The test hypothesis is:

H_0: S=0 (sex has no influence on blood pressure)

H_1: S≠0 (sex has an influence on blood pressure)

It is evident that Z^2=0.137<$X^2_{(0.05,1)}$=3.841 and P-value=0.711>0.05=α, so we conclude that H_0 needs to be rejected and we say that sex does not have any influence on blood pressure. Consequently, the sex of a hypertensive patient does not contribute to his or her predicament.

Outcome

We are testing whether,

H_0: O=0 (hypertension outcome is dependent on blood pressure)

H_1: O≠0 (hypertension outcome is dependent on blood pressure)

Since Z^2= 1.617< $X^2_{(0.05,1)}$=3.844 and P-value=0.203> 0.05=α, therefore we conclude that the blood pressure of patients does not determine their outcome (status: dead/alive).This means that one can actually be hypotensive, normal, prehypertensive or hypertensive and still be alive or dead.

The Logistic Regression Model

Here, we are interested in extracting the logistic regression models from the result in table 4. The estimates of the regression parameters are in the column labeled "estimate". We will generate three logit models:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_i + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Where i= 0, 1, 2

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_i + 0.373X_1 + 0.032X_2 - 0.355X_3$$

Logit 1 is:

$$\ln\left(\frac{\pi_0}{1-\pi_0}\right) = 0.154 + 0.373X_1 + 0.032X_2 - 0.355X_3$$

$$\pi_0 = \frac{1}{1+\exp^{-(0.154+0.373\,X_1+0.032\,X_2-0.355\,X_3)}}$$

Where π_0 is the Probability that a patient is Hypertensive. That is

π_0= P(Hypertensive)

Logit 2 is:

$$ln\frac{\pi_1}{1-\pi_1} = 0.884 + 0.373X_1 + 0.032X_2 - 0.355X_3$$

$$\pi_1 = \frac{1}{1+\exp^{-(0.884+0.373\,X_1+0.032\,X_2-0.355\,X_3)}}$$

Where π_1 is the Probability that a patient is Prehypertensive or Hypertensive. That is

π_1=P (Prehypertensive or Hypertensive)

Logit 3 is:

$$\ln\left(\frac{\pi_2}{1-\pi_2}\right) = 1.275 + 0.373X_1 + 0.032X_2 - 0.355X_3$$

$$\pi_2 = \frac{1}{1+\exp^{-(1.275+0.373\,X_1+0.032\,X_2-0.355\,X_3)}}$$

Where π_2 is the probability that a patient has a Normal blood pressure or Hypertensive or Prehypertensive. That is

π_2= P(Normal Blood Pressure or Prehypertensive or Hypertensive)

Also,

P (Hypotensive)+ P(Normal Blood Pressure)+ P (Prehypertensive)+ P(Hypertensive) =1

P (Hypotensive)= 1- [P(Normal Blood Pressure)+ P (Prehypertensive)+ P(Hypertensive)]

P (Hypotensive)= 1-π_2

Interpretation of Regression Parameters

The logit model

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_i + 0.373X_1 + 0.032X_2 - 0.355X_3$$

has similar interpretation as with the ordinary least square regression equation but the difference being that the threshold (α_i) does not have direct interpretation like the intercept in ordinary least square.

From the logits above, the estimated coefficients represent the change in the log odds for one unit increase in the corresponding independent variable. By implication, for age, we expect a 0.373 unit increase in the ordered log odd given that all other variables are kept constant. For sex, we expect a 0.032 unit increase in the log odd and for outcome, we expect a 0.355 unit decrease in the log odd.

The predictor outcome (dead) gives negative value (-0.355). This indicates that it is associated with the lower response category (that is they tend to favour the lower ranked category which is hypertensive)

Recommendation

As discussed during the course of this study, only age finally affect the final outcome of blood pressure. The causes of hypertension are to be strictly looked into which should help in reducing the number of people with hypertension. I recommend that people should go to the hospital for regular medical check-up in order to be intimate with their blood pressure and they should try as much as possible to live a healthy life to avoid high blood pressure.

Table 1
Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	111.309			
Final	101.255	10.054	3	.018

Link function: Logit.

Table 2
Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	17.646	18	.479
Deviance	19.808	18	.344

Link function: Logit.

Table 3
Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	101.255			
General	99.471	1.784	6	.938

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

Table 4

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[BLOODPRESSURE = 0]	.154	.059	6.835	1	.009	.039	.270
	[BLOODPRESSURE = 1]	.884	.062	200.547	1	.000	.762	1.007
	[BLOODPRESSURE = 2]	1.275	.066	368.813	1	.000	1.145	1.405
Location	[AGE=0]	.373	.130	8.297	1	.004	.119	.627
	[AGE=1]	0[a]	.	.	0	.	.	.
	[SEX=0]	.032	.088	.137	1	.711	-.139	.204
	[SEX=1]	0[a]	.	.	0	.	.	.
	[OUTCOME=0]	-.355	.279	1.617	1	.203	-.903	.192
	[OUTCOME=1]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

References

Agresti A.(1996), "An Introduction to Categorical Data Analysis, Wiley

Babalola B.T (2012), 'Application of Ordinal Logistic Regression to Pregnanncy Outcome'', Lambert Academic Publishing, Germany.

Cifkova R. et al (2004), "Prevalence, awareness, treatment and control of hypertension in the Czech Republic", Journal of human hypertension.

Fisher R.A. (1992), "The goodness of fit of regression formulae, and the distribution of regression coefficient".

Gropelli et al (1992), "Persistent blood pressure increase induced by heavy smoking", Journal of hypertension.

Mitchell C. and Dayton (1992), "Logistic Regression Analysis", University of Maryland

Honser B.W. and S. Lemeshow, "Applied Logistic Regression", New York (2000)

John P. Hoffmann (2010), "Linear Regression Analysis: Application and Assumptions (second edition)".

Kutner M.H ; Nachtsheim C.J and Neter J. (2004), "Applied linear regression models (fourth edition)".

World Health Organization, International Society of Hypertension writing group (2003), "Journal of Hypertension".