6-24-2021

A Review of Logistic Regression and its Application

Sultana Mubarika Rahman Chowdhury
Florida International University, schow034@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

A REVIEW OF LOGISTIC REGRESSION AND ITS APPLICATION

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

by

Sultana Mubarika Rahman Chowdhury

2021

To: Dean Michael Heithaus
College of Arts and Sciences

This thesis, written by Sultana Mubarika Rahman Chowdhury, and entitled A Review of Logistic Regression and Its Application, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

B M Golam Kibria

Sneh Gulati, Co-Major Professor

Florence George, Co-Major Professor

Date of Defense: June 24, 2021

The thesis of Sultana Mubarika Rahman Chowdhury is approved.

Dean Michael Heithaus
College of Arts and Sciences

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2021

DEDICATION

To my father Mr. Matiur Rahman Chowdhury , my mother Mrs. Shabnam Rahman Chowdhury.

ACKNOWLEDGMENTS

Firstly, I would like to express my gratitude to the Mathematics and Statistics Department for giving me the opportunity to pursue my master's degree here at Florida International University. I acknowledge my gratefulness to my respectable co - major professor Dr. Sneh Gulati for her guidance regarding this thesis and continuous support and encouragement. I also extend my heartfelt thanks to my committee member Dr. B M Golam Kibria for being supportive throughout this process. Most of all, I am fully indebted to my major - professor Dr. Florence George, for her understanding, wisdom, enthusiasm, and encouragement and for pushing me further than I thought I could go. I would also like to thank my professors and colleagues from Mathematics and Statistics Department for sharing their wisdom which would lead me to a better future. Furthermore, I would like to thank my friends, all office staffs that directly or indirectly helped me in completing my research work.

I would like to wish my profound gratitude to my beloved parents, and sisters for their immeasurable sacrifice, continuous inspiration and support, which lead me to success in this work. Lastly, I want to thank my partner, Asif Iqbal for the support he has given me in order for me to complete this Master's degree.

ABSTRACT OF THE THESIS

A REVIEW OF LOGISTIC REGRESSION AND ITS APPLICATION

by

Sultana Mubarika Rahman Chowdhury

Florida International University, 2021

Miami, Florida

Professor Florence George, Co-Major Professor

Professor Sneh Gulati, Co-Major Professor

The purpose of this thesis is to do an in-depth review of logistic regression and its application. Additionally, comparison of four different methods of coefficient standardization was done using Heart Disease Dataset. These methods were compared based on testing accuracy, training accuracy, area under the curve, sensitivity, and specificity. Furthermore, logistic regression analysis was applied to National Longitudinal Study of Adolescence Health Survey (Add health) dataset to examine the relationship between anxiety or panic disorder and history of childhood maltreatment, medical conditions such as ADHD, PTSD, some socio-economic conditions and addiction. Results indicated; history of abuse has a significant effect on anxiety disorder. The number of abuses experienced also has a significant association with developing anxiety disorder later in life. Women had higher odds of having such a disorder if they faced maltreatment in their childhood. Moreover, having PTSD and Depression also increased the odds of anxiety substantially.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

CHAPTER 1

INTRODUCTION

1.1 Logistic Regression

Logistic regression analysis is a specialized case of regression analysis, where the variable to be predicted is classified into two or more categories. In such cases, the traditional regression technique fails to explain the association between the independent variables and the response variable.

Binary logistic regression model or logit model is the most common form of this method of analysis in which the response variable takes only two values [Menard, 2000]. However, the method can also be extended when there are three or more categories of dependent variable present such as polytomous or multinomial [Wright, 1995].

Similar to any other regression model, the application of logistic regression can be divided mainly into two categories depending on the research question, either to find the association between the dependent and independent variables or to construct a predictive model that can facilitate forecasting. One advantage of logistic regression over linear regression is that in linear regression the parameters are estimated using the ordinary least square method which has a set of assumption that must be satisfied, otherwise the inferences made about the coefficients will not be correct [Berry et al., 1985, Berry, 1993]. Whereas, in logistic regression some of the assumptions such as linear relationship between the predictors and response variable, the error terms (residuals) being normally distributed or homogeneity of the variance of the variables are not required.

The specific form of a binary logistic regression model generally used is

$$P\left[Y=1\right]=\frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p}}=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p)}}, \qquad (1.1)$$

where Y is the dependent variable and $X_1, X_2, ..., X_p$ are the independent variables. The dependent variable Y takes a value either 0 or 1; 1 indicates the occurrence of a specific event and 0 indicates the absence. Therefore, Pr(Y=1) represents the probability of that event happening and Pr(Y=0) depicts the probability of the event being absent. As the total probability will always be one, $P[Y = 0] = 1 - P[Y = 1]$. Here, β_0 is the intercept and the rest of the $\beta's$ are the regression coefficients. The logit transformation of $P[Y = 1]$ is defined as

$$logit\,(Y) = \ln[\frac{p(Y=1)}{p(Y=0)}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p. \tag{1.2}$$

By taking exponential of the results in Equation 1.2, logit of Y can be converted back to the odds. Hence, results in the equation

$$Odds[Y = 1] = e^{ln[odds(Y=1)]} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}. \tag{1.3}$$

In addition, this odds can be used to calculate the probabilities using Equation 1.2. It must be understood that the odds and the logit explain the same thing in three different ways [Menard, 2002]. As the overall model is tested against the null hypothesis that all coefficients are equal to zero, where possible rejection of the null hypothesis indicates that the effect of the coefficient are not entirely zero in the population, leading to the conclusions implying a better prediction of the probability of the outcome of the dependent variable [Peng et al., 2002].

Figure 1.1 depicts a generic example of the difference between a simple linear regression line and a logistic regression curve. The logistic regression model [Equation 1.2] converts the nonlinear relationship between Pr(Y=1) and the independent variables (X) to a linear equation that explains the effect they may have on the dependent variable. This linear form gives one the opportunity to interpret the coefficients of the proposed model. However, because of such conversion, the or-

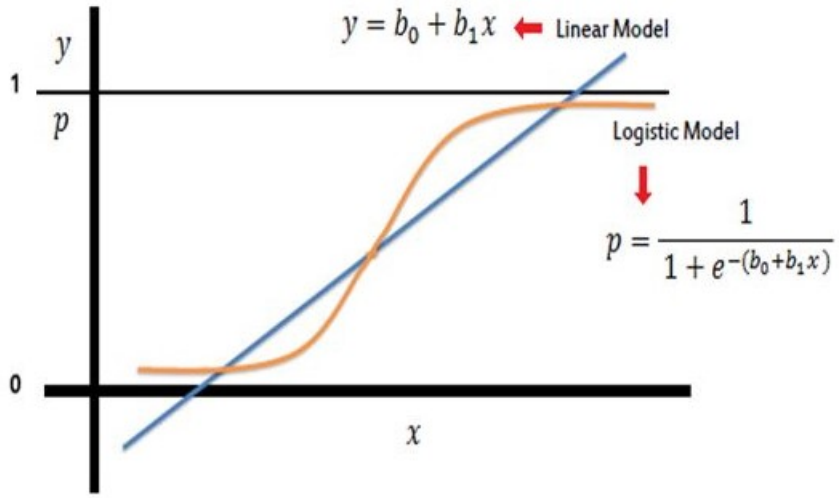

Figure 1.1: Linear Regression vs Logistic Regression [Genesis, 2018].

dinary least square method can not be used to estimate the parameters. Thereby, the parameters, standard error of the parameters and the measures of goodness of fit are generally estimated using the methods of maximum likelihood estimation [Greene, 1993, Peng et al., 2002].

Various calculation using the β's have been proposed to aid interpretation: 1) odds ratio[Morgan and Teachman, 1988a, DeMaris, 1992, Long, 1997], 2) Instantaneous slopes [Aldrich and Nelson, 1984, Morgan and Teachman, 1988b][Petersen, 1985, Roncek, 1991] and, 3) Predicted changes in probability [Petersen, 1985, Roncek, 1991].

The interpretation of results is rendered using the widely used odds ratio technique for both categorical and continuous predictors [Peng et al., 2002]. Odds ratio can be calculated by taking the exponential of the coefficient β_j, where β_j is the coefficient of the j^{th} explanatory variable

$$OddsRatio = \frac{Odds[Y = 1]}{Odds[Y = 0]} = e^{\beta_j}. \tag{1.4}$$

Despite the fact that the odds ratio can give an idea of the direction of the relationship between the response variable and explanatory variables, it is not enough to explain the overall extent of how they are related and also it falls short of comparing over models [Allison, 1999]. Different calculations of the β coefficients have been proposed to make more meaningful interpretations.

1.2 Research Purpose

Our main purpose in this thesis is to do an in-depth review of logistic regression and its application. The primary aims are –

- A review of four different methods of standardization of the β – coefficients of logistic regression with the help of Heart Disease Dataset [Detrano, 1989], based on the resulting testing accuracy, training accuracy, AUC (Area under the curve). The goal is to inspect whether applying standardization on the coefficients have effects on the overall prediction accuracy of the model.
- To apply logistic regression to a complex survey dataset to analyze the association between the response variable and the independent variables.

The rest of the thesis is organized as follows: In Chapter 2, discussion on related research topic is presented. In Chapter 3, standardization methods for the coefficients will be studied and their effects on the overall accuracy of the models will be compared and discussed with a numerical example on the heart disease dataset. In Chapter 4, an example of using logistic regression for survey data analysis will be demonstrated using Add Health Survey data and the results will be studied. Finally, Chapter 5 presents the conclusions drawn from both of these analysis.

CHAPTER 2

LITERATURE REVIEW

There are debates about the necessity of standardizing coefficient for logistic regression akin to the linear regression but there are few situations when it becomes vital. For instance, in cases when it is needed to compare the effects between different explanatory variables on the response variable [Kaufman, 1996].

Standardized coefficients become invariant to the change in scale of measurement that enables one to compare the relative influence of different explanatory variables within logistic regression [Agresti, 2018] [Agresti and Finlay, 1997].

However, even though there are some proposed standardized, semi-standardized coefficients for logistic regression none of them can be universally defined. Furthermore the implementation of these standardization has been strictly used in case of interpretation. There is not enough published research work that explains whether these standardization may effect the prediction ability of a logit model.

Numerous studies have been done applying different techniques to standardize the coefficients. Robert L. Kaufman [Kaufman, 1996] in his study found that semi-standardized coefficients measuring the change in predictive probability of outcomes are preferable because they are intuitively appealing and as they are bounded in the interval [-1, +1] interpretation of their magnitude becomes easier. Six approaches of standardizing the coefficients were analyzed using a practical example by Scott Menard, which included both semi-standardized and completely standardized techniques [Menard, 2004]. The approaches included the one currently most readily available in logistic regression software, the unstandardized coefficient divided by its standard error (which is actually the normal distribution version of the Wald statistic). Followed by four different type of adjustments made to the formula in order to make it fully standardized. The sixth approach was an alternative to these

methods as it utilized the information theory which was conceptually superior to the other approaches but lacked simplicity which made it only possible to apply in the case of simple logistic regression.

To aid comparison across models for the same sample, Winship and Mare suggested to divide the coefficients with the estimated standard deviation of the dependent variable for each model [Winship, 1984]. This method is known as y standardization. The estimated value of dependent variable Y is calculated by adding the standard deviation of the predicted logit to the estimated standard deviation of the error term. This estimated error is considered fixed which leaves only the standard deviation of the predicted logit to be accountable for all variations across models [Mood, 2010]. Dividing the coefficients by the estimated standard deviations, rescales them and nullifies the increase of the standard deviation of the logit that occurs due to the addition of explanatory variables which is included to enhance the prediction of the response variable [Mood, 2010].

Logistic regression has a wide range of applications in various fields and its functionality has increased dramatically in the past several decades. While multiple linear regression falls short in analyzing data with response variable that is not continuous, logistic regression gives an essential tool is such cases. Application of this method is not limited to only binary cases as it can be easily modified for cases where response variables have more than two categories. Risk factor analysis and predictive modeling is one of the main implementations of logistic regression. It is broadly used in medical research fields to examine the association between risk factors and diseases. Logistic regression can also be used in survival analysis by grouping event times into intervals and converting them to categories. It has been seen that by doing so it is possible to get an estimation similar to the proportional hazard model which is generally used for such types of data [Abbott, 1985]. The use standard

logistic regression techniques can also be extended in case the response variable is of ordinal scale [Kleinbaum and Klein, 2010]. In analyzing complex surveys data logistic regression is found to be greatly useful as it gives an essential tool to deal with categorical response variables. A similar example is demonstrated in the 4 th chapter of this thesis.

In the next chapter a review of the standardized logistic regression coefficients will be presented along with a numerical example.

CHAPTER 3

STANDARDIZED REGRESSION COEFFICIENTS

In this chapter some of the established methods of standardizing the coefficients will be discussed with the help of a numerical example for an in-depth review. In addition to the standardization of the coefficients, two techniques of scaling were also applied to the data to compare the results. The first technique includes scaling the data by mean and standard deviation of the corresponding predictors and the later includes scaling the data using median and mean absolute deviation.

The simplest method of partial standardization is to multiply the coefficients by their individual standard deviation. This method was mentioned by Menard [Menard, 1995]

$$b_1 = b * S_x, \tag{3.1}$$

where, the standard deviation of the explanatory variable X (S_x) is multiplied with the unstandardized estimated coefficient of the corresponding variable b. This can be considered as the only predictor based standardization technique. Another similar approach is to change the scale of both the dependent variable and the predictors using the standard deviation of the standard logistic distribution. That is,

$$b_2 = \frac{b * S_x}{\frac{\pi}{\sqrt{3}}}, \tag{3.2}$$

where, $(\pi\sqrt{3}) = 1.8$. This method has been adapted in SAS to standardize the coefficients in the PROC LOGISTIC procedure. Long suggested another approach for standardization which includes the standard deviation of the standard normal distribution [Long, 1997]. Calculation of this method is similar to the previous one, the only difference is the standard deviation of the standard normal distribution is

added with the standard deviation of the logistic distribution. Hence Equation (3.2) becomes,

$$b_3 = \frac{b * S_x}{\frac{\pi}{\sqrt{3}} + 1}. \tag{3.3}$$

All of these standardized coefficients only take into account the variation of the independent variable. Hence, they cannot be considered as fully standardized. To standardize the response variable standard deviation of logit (y) needs to be calculated, which is tricky. A way out of this is to use the standardization followed in OLS, which is defined as follows,

$$b^{**} = b * \frac{Sx}{Sy}. \tag{3.4}$$

Again, from the definition of Coefficient of Determination (R^2), we get

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2}, \tag{3.5}$$

where, $\widehat{y}$ is the estimated value of y. Adjusting the equation for OLS we get,

$$S_y^2 = \frac{S_{\hat{y}}^2}{R^2}, \tag{3.6}$$

Substituting *logit(y)* in case of y and *logit($\hat{y}$)* in the place of $\hat{y}$ we get for logistic regression,

$$S_{logit(y)}^2 = \frac{S_{logit(\hat{y})}^2}{R^2}. \tag{3.7}$$

[Menard, 1995]. Hence, using the similar strategy used in OLS the estimated coefficients can be standardized as follows

$$b_4 = \frac{(b * S_x)\,(R)}{S_{logit(\hat{y})}} \tag{3.8}$$

This coefficient can be considered as fully standardized as it also takes into account the variance of the response variable in contrast to the other coefficients discussed before where only the variation of the predictor was studied.

3.1 A Numerical Example

In order to illustrate the calculation of the standardization techniques and to review the outcomes, Cleveland Heart disease dataset was used. It is a widely used database publicly available online [Detrano, 1989]. The aim was to apply logistic regression to develop a predictive model for heart diseases using the predictors. The four different coefficient standardization methods were applied to the coefficients of the customary model. After that the resultant models were compared based on their prediction accuracy.

3.1.1 Dataset Details

Originally, the data set contained 76 attributes, but a subset of 14 variables are generally used by the researchers in all published experiments with a total of 313 observations.

The 14 variables include a response variable "target" which refers to to the presence of heart disease in the patient. It is integer valued, 0 = no/less chance of heart attack, and 1 = yes/ more chance of heart attack.

The 13 predictors considered in the dataset are as follows [Detrano, 1989]:

1. *AGE*: Continuous
2. *SEX*: Categorical (1 : Male, 0 : Female)
3. *Chest Pain Type(CP)*: Categorical (4 values) 0: typical angina 1: atypical angina 2: non - anginal pain 3: asymptomatic

4. *Trestbps*: Continuous, represents resting blood pressure on admission
5. *Chol*:Continuous, represents Serum cholesterol in mg/dl
6. *Fbs*: Categorical , represents fasting blood sugar level, (2 values) 1: True - fasting blood sugar is greater than 120 mg/dl 0: False - fasting blood sugar is less than 120 mg/dl
7. *Restecg*: Categorical ,represents resting electrocardiographic outcomes (4 values) 0: normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. *Thalach*: Continuous, represents maximum heart rate achieved
9. *Exang*: Categorical, represents existence of exercise induced angina (2 values Yes/No)
10. *Oldpeak*: Continuous, ST depression induced by exercise relative to rest
11. *Slope*: Categorical, represents the slope characteristics of the peak exercise ST segment
12. *Ca*: Discrete, represents number of fluoroscopy colored major vessels (values 0-3);
13. *Thal*: Categorical, (3 values) 0: normal 1: fixed defect 2: reversible defect

Table 3.1 and Table 3.2 shows the descriptive statistics of the categorical and continuous variables mentioned above.

Table 3.1: Descriptive statistics for categorical variables.

Variables	**Categories**	**N**	**Percentage(%)**
Sex	Male	207	68.31
	Female	96	31.68

Table 3.1 continued from the previous page

Variables	Categories	N	Percentage(%)
CP	0	143	47.19
	1	50	16.50
	2	87	28.71
	3	23	7.59
fbs	0	258	85.15
	1	45	14.85
restecg	0	147	49.51
	1	152	50.17
	2	4	1.32
exang	0	204	67.33
	1	99	33.67
slope	0	21	6.93
	1	140	46.20
	2	142	46.86
ca	0	175	57.76
	1	65	21.45
	2	38	12.54
	3	20	6.60
	4	5	1.65
thal	0	2	0.66
	1	18	5.94
	2	166	54.79
	3	117	38.61
target	0	138	45.54
	1	165	54.46

Among the total population, 68.31 % were male and the rest were female. The average age of the respondents was 54.37 (SD = 9.08) [Table 3.2].

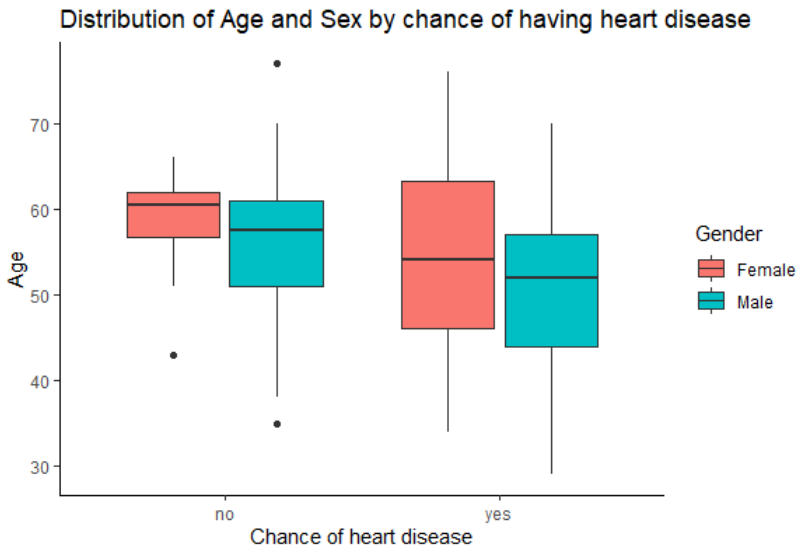

Figure 3.1: Boxplot showing distribution of Age and Sex by chance of having heart disease.

Figure 3.1 depicts a boxplot of the distribution of age and gender based on their probability of developing heart disease.

Table 3.2: Descriptive statistics for continuous variables.

Variables	**Mean (SD)**	**Median**	**Mean Absolute Deviation**
Age	54.37 (9.08)	55.0	10.38
Trestbps	131.6 (17.54)	130.0	14.83
Chol	246.3 (51.83)	240.0	47.44
Thalach	149.6 (22.91)	153.0	22.24
Oldpeak	1.04 (1.16)	0.8	1.19

The average age at which women may have a high probability of having a heart disease is seen to be higher than men in the study population [Figure 3.1]. However, men are seen to develop a higher risk of heart disease earlier in life than women.

3.1.2 Implementation of Standardization Methods

Primarily, logistic regression was applied to the complete dataset. Table [3.3] presents the results for the logistic regression analysis of the predictors of the dependent variable that refers to the presence of heart disease in a patient.

Table 3.3: Logistic Regression Results (Unstandardized data)

Predictor	**Coefficients**	**SE**	**P-value**	**SD**
AGE	-0.0049	0.0232	0.8323	9.0821
SEX	-1.7582	0.4688	0.0002 ***	0.4660
CP	0.8599	0.1854	0.0000 ***	1.0321
Trestbps	-0.0195	0.0103	0.0596 .	17.5381
Chol	-0.0046	0.0038	0.2209	51.8308
Fbs	0.0349	0.5295	0.9475	0.3562
Restecg	0.4663	0.3483	0.1806	0.5259
Thalacht	0.0232	0.0105	0.0265 *	22.9052
Exang	-0.9800	0.4099	0.0168 *	0.4698
Oldpeak	-0.5403	0.2139	0.0115 *	1.1610
Slope	0.5793	0.3498	0.0977 .	0.6162
Ca	-0.7734	0.1909	0.0000 ***	1.0226
Thal	-0.9004	0.2901	0.0019 **	0.6123

Note: Dependent Variable: Presence of heart disease in the patient (reference category= (0) No). Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The coefficients of the model along with their standard error and significance values are given here. From the results it can be seen that predictor *Sex*, *cp*,

Thalacht, *Exang*, *Oldpeak*, *Slope*, *Ca*, *Thal* were found to have a significant effect on the presence of the heart disease at 0.05 level of significance.

In the next step, the four standardization techniques of the coefficients discussed in the previous section were applied to this result. Calculation of b_1 is done by simply multiplying the standard deviation of each explanatory variable with their corresponding coefficients. For instance, for *Age* b_1 = (-0.004908) * (9.0821010) = -0.04457922 and so on. To get b_2 [Equation 3.2] above result has to be divided by $\pi\sqrt{3}$, the numerical value of which is approximately 1.814. Hence the for *Age* the standardized coefficient becomes b_2 = (-0.004908) * (9.0821010)/ 1.814 = -0.02457. To obtain the standardized coefficient by the third method [Equation 3.3] discussed in the previous section, the calculation is similar but instead of dividing by $[\pi\sqrt{3}]$ the unstandardized coefficients are divided by $[\pi\sqrt{3}+1]$ which is equal to 2.814 approximately. Therefore, for Age the calculation of the standardized coefficients is as follows: b_3 = (-0.004908) * (9.0821010)/ 2.814 = -0.01584. The fully standardized fourth approach utilizes the coefficient of determination (R^2) value to calculate the modified coefficients. In this method, it multiplies the first approach explained in equation 3.1 by $R/S_{logit(\hat{y})}$. In this example, the value of the square root of R^2 divided by the the standard deviation of the $logit(\hat{y})$ was calculated to be 0.246434. So the modified coefficient for predictor Age changed in to b_4 = (-0.004908) * (9.0821010) * (0.246434) = -0.01098. Similar calculations have been done for all other variables and are presented in Table [3.4].

The column 'Customary model' in Table [3.4] refers to the calculated unstandardized coefficients from the logistic regression model. 'Method 1', 'Method 2', 'Method 3', and 'Method 4' represent the standardized coefficients computed using Equation 3.1, Equation 3.2, Equation 3.3, Equation 3.7 respectively.

Table 3.4: Modified coefficients using different standardization methods

	Customary model	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Intercept	3.4505				
Age	-0.0049	-0.0446	-0.0246	-0.0158	-0.0110
Sex	-1.7582	-0.8193	-0.4517	-0.2912	-0.2019
Cp	0.8599	0.8874	0.4893	0.3154	0.2189
Trestbps	-0.0195	-0.3416	-0.1883	-0.1214	-0.0842
Chol	-0.0046	-0.2400	-0.1323	-0.0853	-0.0591
Fbs	0.0349	0.0124	0.0069	0.0044	0.0031
Restecg	0.4663	0.2452	0.1352	0.0871	0.06043
Thalach	0.0232	0.5317	0.2931	0.1889	0.1310
Exang	-0.9800	-0.4604	-0.2538	-0.1636	-0.1135
Oldpeak	-0.5403	-0.6273	-0.3458	-0.2229	-0.1546
Slope	0.5793	0.3570	0.1968	0.1269	0.0880
Ca	-0.7733	-0.7908	-0.4360	-0.2811	-0.1949
Thal	-0.9004	-0.5513	-0.3040	-0.1959	-0.1359

From the results in Table 3.4 it is evident that as the coefficients start from being partially standardized using method 1 to fully standardized in method 4, they seem to decrease in term of magnitude. Techniques used in SAS have the closest values to the method suggested by Long. Predictor cp (chest pain) seems to have the comparatively higher relative effectiveness among the the significant variables.

In the next step, the target was to set up four different models using standardized coefficients calculated by these approaches and compare their performance based on prediction accuracy. To measure the prediction accuracy, dataset was randomly divided into two sets; the testing set which contains 20% and training set which contains rest of the 80 % of the entire population. The models were developed using the training set and the testing set was used to verify the overall accuracy, sensitivity, and specificity. One of the major hurdles faced while setting up models to calculate their accuracy's is that the predictors were measured in different scale. Hence to avoid any kind of measurement error, the predictors were scaled before any kind of analysis was done. Two different types of scaling was adapted to scale the dataset. These are

- Standardization with mean/SD : Mean of each variable was subtracted from individual observation and divided by their corresponding standard deviation.
- Standardization with median/MAD : Median of only the continuous variables were subtracted from individual observation and divided by their corresponding Mean Absolute Deviation (MAD), where Mean Absolute Deviation (MAD) $= Mean \mid x - Median \mid$. Categorical variables were standardized using the mean and standard deviation.

Table [3.5] and Table [3.6] shows the results of the logistic regression analysis of heart disease based on the 13 predictors for scaled data using mean/standard deviation and median/ mean absolute deviation respectively.

Table 3.5: Unstandardized logistic regression coefficients (Mean/SD Scaled data)

Predictor	**Coefficients**	**Standard error**	**P-value**
AGE	-0.0446	0.2105	0.8323
SEX	-0.8193	0.2185	0.0002 ***
CP	0.8874	0.1913	0.0000 ***
Trestbps	-0.3416	0.1813	0.0596 .
Chol	-0.2400	0.1960	0.2209
Fbs	0.0124	0.1886	0.9475
Restecg	0.2452	0.1831	0.1806
Thalacht	0.5317	0.2396	0.0265 *
Exang	-0.4604	0.1925	0.0168 *
Oldpeak	-0.6273	0.2483	0.0115 *
Slope	0.3570	0.2156	0.0977 .

Table 3.5 continued from the previous page

Predictor	**Coefficients**	**Standard error**	**P-value**
Ca	-0.7908	0.1952	0.0000 ***
Thal	-0.5513	0.1776	0.0019 **

Note: Dependent Variable: Presence of heart disease in the patient (reference category= (0) No). Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3.6: Unstandardized logistic regression coefficients

(Median/MAD scaled data)

Predictor	**Coefficients**	**Standard error**	**P-value**
AGE	-0.0126	0.1814	0.9446
SEX	-0.7565	0.2233	0.0007 ***
CP	0.8374	0.2104	0.0000 ***
Trestbps	-0.0870	0.1577	0.5814
Chol	0.0057	0.1481	0.9692
Fbs	0.0549	0.1985	0.7820
Restecg	0.0676	0.2014	0.7370
Thalacht	0.6214	0.1756	0.0004 ***
Exang	-0.7495	0.2117	0.0004 ***
Oldpeak	-0.7135	0.1824	0.0000 ***
Slope	0.7701	0.2062	0.0002 ***
Ca	-0.8306	0.1976	0.0000 ***
Thal	-0.6134	0.1874	0.0011 **

Note: Dependent Variable: Presence of heart disease in the patient (reference category= (0) No). Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From Table [3.5] and Table [3.6] it can be noticed that based on the standardization technique used for the analysis, the resultant p-values differed for some variables. For example, for the first dataset, three variables *Trestbps*, *sex*, and *CP* were found to be significant at 5% level of significance whereas for the median scaled dataset 7 variables including *Thalacht*, *Exang*, *Slope*, and *Ca*, were found to be significant.

Previously explained four methods of standardizing the coefficients were then applied to both of these scaled dataset to standardize the coefficients. Thus, eight set of modified coefficients from the coefficients from Table [3.5] and Table [3.6] were computed. All these calculations has been done with the help of statistical software R. One predicament in this process is that, as the training and testing set are divided randomly using R, there is a chance of getting different results for different subsets which may result in bias. To solve this issue the complete process was repeated 1000 times and average of these repetitions was taken for all further calculations. Outcomes of standardization of the coefficients are given in Table 3.7 and Table 3.8.

Here in Table 3.7 column 'Customary model' refers to the unstandardized coefficients of the dataset scaled by mean and standard deviation along with the four standardization methods for the coefficients in the following columns. Similarly, in Table 3.8 column 'Customary model' refers to the unstandardized coefficients of the dataset scaled by median and mean absolute deviation along with the four standardization methods for the coefficients in the following columns.

Resultant standardized coefficients were then used to set up eight logistic regression models. In the next stage to evaluate the performance of these models, parameters like training accuracy, testing accuracy, overall prediction accuracy, sensitivity, and specificity probability, corresponding to each model, have been calculated by

Table 3.7: Logistic regression coefficients (Mean/SD scaled data)

	Customary model	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Intercept	0.2319				
Age	-0.0419	-0.0419	-0.0231	-0.0365	-0.0101
Sex	-0.8188	-0.8172	-0.4505	-0.7106	-0.1966
cp	1.0425	1.0317	0.5688	0.8972	0.2483
trestbps	-0.2409	-0.2340	-0.1323	-0.2087	-0.0577
chol	-0.2510	-0.2297	-0.1266	-0.1997	-0.0553
fbs	-0.0730	-0.0755	-0.0416	-0.0657	-0.0182
restecg	0.3668	0.3711	0.2046	0.3228	0.0893
thalach	0.3420	0.3385	0.1866	0.2944	0.0815
exang	-0.4276	-0.4304	-0.2373	-0.3743	-0.1036
oldpeak	-0.5950	-0.6236	-0.3438	-0.5423	-0.1501
slope	0.5568	0.5641	0.3110	0.4905	0.1357
ca	-0.7673	-0.7983	-0.4402	-0.6943	-0.1921
thal	-0.5539	-0.5676	-0.3129	-0.4936	-0.1366

Table 3.8: Logistic regression coefficients (Median/MAD scaled data)

	Customary model	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Intercept	0.6920				
Age	0.0650	0.0844	0.0465	0.0734	0.0177
Sex	-0.8415	-0.8398	-0.4630	-0.7303	-0.1760
cp	1.1343	1.1226	0.6189	0.9762	0.2353
trestbps	-0.1610	-0.2322	-0.1280	-0.2019	-0.0487
chol	0.1072	0.1527	0.0842	0.1328	0.0320
fbs	0.0673	0.0696	0.0384	0.0606	0.0146
restecg	0.1261	0.1276	0.0703	0.1109	0.0267
thalach	0.6672	0.9032	0.4979	0.7854	0.1893
exang	-0.7237	-0.7284	-0.4016	-0.6335	-0.1527
oldpeak	-0.8915	-1.1391	-0.6280	-0.9906	-0.2387
slope	0.9685	0.9811	0.5409	0.8532	0.2056
ca	-0.7982	-0.8305	-0.4579	-0.7222	-0.1741
thal	-0.6699	-0.6864	-0.3784	-0.5969	-0.1439

making use of the following equation

$$p\,(y=1) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}. \tag{3.9}$$

By setting up the probability threshold to 0.5 the outputs have been divided in to two groups indicating the presence of heart disease to 'Yes' (= 1) if the probability is greater than 0.5 and 'No' (= 0) if the value is found to be less than 0.5. These re-coded values make it possible to build up separate confusion matrices for each model and afterwards enumerate training accuracy, testing accuracy, overall prediction accuracy, sensitivity, specificity, and area under the ROC (Receiver operating characteristic) curve commonly known as AUC.

Table 3.9 shows the testing accuracy and training accuracy of the models constructed by applying each of the four coefficient standardization methods along with the model of unstandardized coefficients, which is represented by the 'Customary model' column.

Results indicate that testing accuracy of the customary model was slightly higher than all standardized models for median/MAD scaled data. However, for the mean/SD scaled data, the testing accuracy for the customary model and the models for the 4 methods were similar. Similarly, the training accuracies were somewhat similar for the unstandardized and standardized coefficients. Moreover, method 4 was seen to have the lowest prediction accuracy among all four methods. On the other hand, by comparing the testing and training accuracies for mean/SD scaled data and median/MAD scaled data it can be seen that median/MAD scaled data has approximately 4% to 5% higher accuracy overall.

However, by taking a look at the AUC's for these models [Table 3.10] it can be seen that even though the unstandardized model had slightly different AUC, there was no difference in AUC's of the models constructed from different standardization

Table 3.9: Table for Testing and Training Accuracy.

Data	**Customary model**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Mean Standardized (Test set)	0.8193	0.8218	0.8193	0.8221	0.8126
Median Standardized (Test set)	0.8754	0.8767	0.8646	0.8766	0.7813
Mean Standardized (Train set)	0.8576	0.8574	0.8540	0.8569	0.8472
Median Standardized (Train set)	0.9005	0.9000	0.8859	0.8987	0.8083

Table 3.10: AUCs for Testing and Training data.

Data	**Customary model**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Mean Standardized (Test set)	0.8895	0.8899	0.8899	0.8899	0.8899
Median Standardized (Test set)	0.9230	0.9210	0.9210	0.9210	0.9210
Mean Standardized (Train set)	0.9262	0.9261	0.9261	0.9261	0.9261
Median Standardized (Train set)	0.9280	0.9279	0.9279	0.9279	0.9279

techniques. This indicates that in terms of distinguishing between the two diagnostic groups, all of these models show similar performance. In addition to the AUC, the sensitivity and specificity of these models has to be computed. Table 3.11 presents the sensitivity and specificity of the customary model along with the modified models.

In terms of improving the sensitivity or specificity of the models the standardization techniques seem to have no significant effect. As the overall accuracy for the standardized models were lower than the un-standardized one, evidently the sensitivity ans specificity was also found to be less than the prior. Moreover, method 4

Table 3.11: Sensitivity and Specificity

Data	Customary model		Method 1		Method 2		Method 3		Method 4	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
Mean Standardized (Test set)	0.8801	0.7462	0.8836	0.7470	0.8911	0.7345	0.8852	0.7463	0.9060	0.7046
Median Standardized (Test set)	0.8935	0.8569	0.8929	0.8597	0.9109	0.8125	0.8994	0.8519	0.9459	0.5934
Mean Standardized (Training set)	0.9117	0.7925	0.9114	0.7925	0.9178	0.7893	0.9131	0.7893	0.9294	0.7477
Median Standardized (Training set)	0.9110	0.8877	0.9099	0.8880	0.9221	0.8421	0.9132	0.8811	0.9621	0.6223

seems to have the higher sensitivity than all other models, which also means lower specificity than others.

It is worth mentioning that the techniques used to scale the dataset seem to have some effect on improving the overall accuracy of the models. Test sets taken from the dataset for which the numerical variables were scaled using median standardization performed better than the one which was scaled using mean and standard deviation.

For instance, for the customary model and the first three models, the testing accuracy were approximately 4% higher in the case of the dataset standardized by median/MAD [Table 3.9]. Additionally, from Table 3.10 it can be seen that the AUC are slightly higher for the data which was standardized using median/MAD.

3.1.3 Discussion

The primary purpose of standardizing logistic regression coefficients is to set a ground on the basis of which the predictors can be ranked. The absolute value of the standardized coefficients enables one to order the independent variables in terms of importance. According to Menard [Menard, 1995] standardized coefficients render a more precise idea than the un-standardized logistic regression coefficients. However, adapting such measures for the sake of interpretation may effect the overall performance of the model. In this section, the goal was to investigate how different

standardization techniques effect the accuracy of the logistic regression model under study.

By taking a closer look at results it can be realized that standardizing the coefficients did not affect the overall prediction accuracy of the predictive logistic regression model. Similarly, no evidence was found that following a certain type of standardization technique would show better performance than the others, the un-stardardized regression model in general had higher accuracy.

In essence, standardizing facilitates better interpretation and does not effect the predictive capacity of the model. This is evident from the AUC's computed for both un-standardized and standardized regression coefficients showed in Table 3.8. As the AUC's calculated from taking the average of multiple iterations, they turned out to be exactly equal for all standardization techniques, which was also similar to the un-standardized logistic regression model. Hence, if the primary goal of conducting a logistic regression analysis is building up a predictive model which can also be used for comparing the predictor effects and does not affect the overall accuracy of the model, standardizing the regression coefficients may be advisable.

CHAPTER 4

LOGISTIC REGRESSION ON SURVEY DATA

In this chapter, the application of logistic regression is demonstrated by analyzing a survey dataset.

4.1 Survey Data

A sample survey is a method for collecting data from a population to obtain inferences about the entire population from a subset, or a sample of population individuals [Williams, 2014]. This kind of complex samples can be used for both descriptive and analytical procedures. Sample survey utilizes different sampling design tools to make the result to represent the complete population. Sampling techniques such as probability sampling, stratified sampling are used which increases the complexity of analyzing data. These techniques enable one to draw valid inferences from a smaller representative part about the population parameter of a large population. This is done by designating a sampling weight for each sample point that is the inverse of its sample selection probability. This is done by designating a sampling weight for each sample point that is the inverse of its sample selection probability [Williams, 2014].

4.2 Analysis of Survey Data

The challenge in analyzing survey data is to select a suitable study design that will deliver satisfactory inferences at a reduced cost. Sampling weights are the values assigned to each of the data points in a data set to make the data set a more representative sample of the entire population. To be specific, this value indicates how many times each of the observation of the sample should be counted in any

statistical analysis. When units are sampled with unequal probability, it is necessary to give them corresponding unequal weights in the analysis [Lumley et al., 2004].

Two most common types of weights are: Design weights and post stratification weight or non-response weight [Johnson, 2008]. If the data set in disproportionately stratified then setting equal weights for all observations may cause over or under-sampling and produce biased results. Design weights are used to solve this issue. It enables the results from the analysis to represent the complete population. Another type of weight that is used in survey data analysis is known as post-stratification weight or in other words, non-response weight. Including this type of weight paves a way to solve the complications due to the non-response of persons with certain characteristics such as age, race, gender, etc.

Although weighed survey data analysis is a useful technique to convert a sample survey to represent the whole population, but this approach is not without problems. The primary purpose of the weights is to adjust mean and proportions which does not affect the descriptive statistics. However, in case of inferential statistics, it has been seen that the weights tend to increase the standard error of the estimates. There are some statistical procedures which can be used to reduce this consequence of using weights. Another way is the normalization of the weights which lessens the bias. In the next section, a practical example will be presented on the application of logistic regression in analyzing a complex survey dataset.

4.3 A Numerical Example

The data used for the practical example is a nationally representative longitudinal study known as Add Health.

4.3.1 Study Population

Add Health is a school-based longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States in 1994-95 [Harris, 2013, Harris et al., 2009].For more than 20 years data have been collected from adolescents, their fellow students, school administrators, parents, siblings, friends, and romantic partners through multiple data collection components. In addition, existing databases with information about respondents' neighborhoods and communities have been merged with Add Health data, including variables on income and poverty, unemployment, availability and utilization of health services, crime, church membership, and social programs and policies.

Add Health is a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. No direct support was received from grant P01-HD31921 for this analysis. Add Health was mandated by the United States Congress for the purpose of measuring the role the social environment plays on adolescent health issues and to study adolescents in the United States on a national scale as they mature into adulthood [Harris, 2013]. The survey was conducted in five stages known as Wave I, Wave II, Wave III, Wave IV, Wave V. The first wave was conducted in the year 1994-1995, which sampled adolescent students in grades 7-12 in United States. In 1996 Wave II was collected when participants were enrolled in grades 8-12, Wave III was collected in 2001-2002 when participants were aged 18-26, followed by Wave IV which was collected in 2008 when the participants were 24 to 32 years old and settling into young adulthood. Through Wave IV data collection, the researchers

obtained longitudinal survey data on the social, economic, psychological, and health circumstances of the respondents, as well as longitudinal geographic data. Survey questions were expanded on educational transitions, economic status and financial resources and strains, sleep patterns and sleep quality, eating habits and nutrition, illnesses and medications, physical activity, emotional content and quality of current or most recent romantic/cohabiting/marriage relationships, and maltreatment during childhood by caregivers [Harris, 2013]. A part of the complete data set was made public to facilitate independent researchers to investigate from a wide range of data relating health and social behavior of the participants. The findings from the data set enables researchers to follow the cohort throughout the years and analyze adolescence behavior to future health outcomes. The public use data set is available from different sources. The dataset [Harris and Udry, 2014] used in this analysis was collected from the public use dataset from the Data Sharing for Demographic Research (DSDR) project website and is housed within Inter-university Consortium for Political and Social Research (ICPSR) and is fully funded through the Population Dynamics Branch (PDB) of NICHD.

4.3.2 Study Design

The Add health study design used clustered study design and the clusters were selected with unequal probabilities to reduce the cost of data collection and management. However, this compromises with the characteristics of the observations being independently and identically distributed. Special statistical software is needed to run logistic regression analysis in such cases. To facilitate that, public-use data sets of Add Health made available have been properly weighted to account for the cluster sampling with unequal probabilities.

4.3.3 Research Question

The main objective of using The National Longitudinal Study of Adolescent to Adult Health (Add Health) dataset in this thesis is to examine whether a history of childhood abuse, medical conditions such as ADHD, PTSD, or socio-economic conditions, and addiction has any association with developing anxiety and panic disorder later in their life.

4.3.4 Variables and corresponding measurements

Multiple logistic regression was performed on the selected survey dataset in order to find the significant effect of the explanatory variables on the response variable.The variables under consideration are briefly discussed below.

Response variable

Anxiety and panic disorder has been considered as the study variable in this analysis. Based on the self-reported positive response to the question "Has a doctor, nurse, or other health care provider ever told you that you have or had: anxiety or panic disorder?" , the binary dependent variable was created.

Explanatory variables

To create the variable for abuse history in childhood, a six-item questionnaire in the Wave IV data set called "Mistreatment by adults" was used. The questionnaire includes questions that inquire the respondents about facing emotional, sexual or physical abuse in their earlier life. To measure emotional abuse, the corresponding question was "Before your 18^{th} birthday, how often did a parent or other adult caregiver say things that really hurt your feelings or made you feel like you were

not wanted or loved?".Sexual abuse was assessed by the question, "How often did a parent or other adult caregiver touch you in a sexual way, force you to touch him or her in a sexual way, or force you to have sexual relations?" Similarly to measure physical abuse, the question used was, "Before your 18^{th} birthday, how often did a parent or adult caregiver hit you with a fist, kick you, or throw you down on the floor, into a wall, or down stairs?" All the above questions have the six options which were "this never happened," "1 time," "2 times," "3–5 times," "6–10 times," and "more than 10 times. These options were recoded into two options "Happened at least once", "Never happened" where the incidents of facing any kind of abuse at least one time were re coded as "Happened at least once" and all other options were coded as "Never Happened". After that, all these three variables were merged into a single variable representing history of any kind of abuse that may have happened to the respondents as a child.

The measurement scales used to measure the level of childhood abuse was validated by the database administrators. The item measuring childhood emotional abuse was developed by the Add Health database administrators based on similar questions form the Childhood Trauma Questionnaire. Childhood physical abuse was measured using Traumatic Events Scale [Bernstein et al., 2003] and the scale used to measure childhood sexual abuse [Straus et al., 1998] was adapted from the Conflict Tactics Scale. These scales are used by numerous researchers for studies involving childhood abuse. Childhood abuse was linked to developing migraine in later life based on the data set of Wave IV in the published literature [Tietjen, 2016]. Equivalently, the childhood abuse items from Wave-IV dataset have been broadly used in research. For instance, based on Add Health data it has been found that childhood maltreatment has a direct and indirect effect in victimization by dating partners during adolescence and victimization by romantic and marital partners during adult-

hood, and whether dating victimization mediates the relationship between child abuse and intimate partner victimization in adulthood [Cohn and Iratzoqui, 2016]

In order to create the variable addiction some commonly known form of addiction such as smoking, drinking, marijuana and some other drugs were considered. These data were collected from a 119 item questionnaire titled tobacco, alcohol, and drugs. The questions were self- administered by the respondent. The four questions that were used to create the variable addiction were, "Currently, how soon after you wake up do you have your first cigarette? ", "Did you continue to drink after you realized drinking was causing you problems with family, friends, or people at work or school?", "Did you continue to use marijuana after you realized using it was causing you problems with family, friends, or people at work or school?" Did you continue to use favorite drug after you realized using it was causing you problems with family, friends, or people at work or school?" Addiction to smoking can be identified by the habit of smoking early in the morning. According to the Fagerström Test [Heatherton et al., 1991], which evaluates nicotine dependence, if you have your first cigarette of the day within five minutes of waking up, your addiction is pretty strong. If it's within 30 minutes, it's moderate, and if it's within 60 minutes or later, it's somewhat lower [Ratjen et al., 2017]. The question in the Add health data set has the options "within 5 minutes", "within 6 to 30 minutes", "within 31 to 60 minutes", "after 60 minutes". To convert this in to a binary response "within 5 minutes", "within 6 to 30 minutes" have been merged to be considered as the number of addicted respondents and the other options are considered to be the number of patients who are not severely addicted to nicotine. All other question concerning addiction such as alcohol, marijuana, and other drugs were recorded as a binary yes/ no response. These four types of addiction were then merged into a single variable which create the variable addiction which was used for the final

analysis. Among the other explanatory variables the history of being diagnosed with conditions such as ADHD (Attention deficit hyperactivity disorder), PTSD (Post traumatic stress disorder, Depression were considered. All of these questions had binary yes/no responses.

Age, sex, household income, race are the social and demographic characteristics that have been considered as other variables that may have a confounding effect in the model. In the Wave IV of the Add health data, the age cohort was between 25-34 years. To pin point the effect of age for causing anxiety disorder, this cohort was divided in to two age groups (less than 30 years, 30 and above). Based on annual household income respondents were classified into three classes such as Low income (less than 40,000 USD), mid-range income (between 40,000 USD and 100,000 USD), and high income class (100,000 USD or more).

4.3.5 Results and Discussions

In the public use dataset for wave IV, information of 5114 respondents are available with a mean age of 29 years ranged between 25 to 34. About 46% (n = 2354) of which are male (n = 2352), and the rest are Female. Among these 5114 respondents, 12% (n = 639) were diagnosed with anxiety or panic disorder. Table 4.1 depicts an overview of the overall characteristics of the study population.

Table 4.1: Descriptive Statistics of the study variables

Characteristics of the study population	Total Population	Diagnosed with anxiety disorder	
		Yes 639 (12.5 %)	No 4474 (87.5 %)
Age			
Mean	29 (±0.02)	29.01	28.94
Range	25-34	25-34	25-32
Less than 30, n(%)	2992 (58.52%)	375 (12.5%)	2617 (87.5 %)
More than 30, n(%)	2121 (41.48%)	1857 (87.6%)	264 (12.4%)
Sex			
Male, n(%)	2352 (46 %)	167 (26.2%)	2185 (48.8%)
Female, n(%)	2761 (54%)	472 (73.9 %)	2289 (51.2%)
Race			
White, n(%)	3671 (71.8%)	538 (84.3 %)	3132 (70.1%)
Others, n(%)	1438 (28.2 %)	138 (15.7 %)	1338 (29.9%)
BMI			
Underweight, n(%)	77 (1.5%)	18 (2.8%)	59 (1.3%)
Normal, n(%)	1578 (30.9%)	221 (34.6%)	1357 (30.3%)
Overweight, n(%)	3458 (67.6%)	400 (62.6%)	3058 (68.4%)
Income			
Low, n(%)	1608 (33.8%)	246 (40.9%)	1362 (32.7%)
Medium, n(%)	2430 (51%)	294 (48.9%)	2135 (51.3%)
High, n(%)	723 (15.2%)	61 (10.1%)	662 (15.9%)
Physical Abuse			

Table 4.1 continued from the previous page

Characteristics of the study population	Total Population	Diagnosed with anxiety disorder	
		Yes 639 (12.5 %)	No 4474 (87.5 %)
Never happened, n(%)	4170 (82.6%)	470 (74.1%)	3700 (83.8%)
Happened at least once, n(%)	880 (17.4%)	164 (25.9%)	716 (16.2%)
Emotional Abuse			
Never happened, n(%)	2667 (53.0%)	1956 (44.5%)	2442 (55.5%)
Happened at least once, n(%)	2366 (47.0%)	410 (64.6%)	225 (35.4%)
Sexual Abuse			
Never happened, n(%)	4800 (94.9%)	569 (89.7%)	4231 (95.7%)
Happened at least once, n(%)	257 (5.1%)	65 (10.3%)	192 (4.3%)
PTSD			
No, n(%)	4952 (96.9%)	552 (86.4%)	4400 (98.3%)
Yes, n(%)	161 (3.1%)	87 (13.6%)	74 (1.7%)
ADHD			
No, n(%)	4838 (94.6%)	572 (89.5%)	4266 (95.4%)
Yes, n(%)	275 (5.4%)	67 (10.5%)	208 (4.6%)
Depression			
No, n(%)	4286 (83.8%)	423 (9.5%)	4051 (90.5%)
Yes, n(%)	827 (16.2%)	404 (63.2%)	423 (9.5%)
Addiction			
No, n(%)	957 (18.8%)	163 (25.6%)	794 (17.8%)
Yes, n(%)	4140 (81.2%)	473 (74.4%)	3667 (82.2%)

The mean age of the respondents diagnosed with anxiety or panic disorder was 29.01 which was not much different than the others who were not. The percentage of Female respondents having such a disorder was seen to be higher than Male respondents. However, female to male ratio is not equal in the study population hence further test needs to be carried out to check if there is any statistically significant difference. White / Caucasians constituted a major portion of the population (71.8%). Distributing the study population on the basis of their reported BMI, it can be seen that more than 50% of them were overweight. The income of individual respondents was categorized in to three categories where about 50% (n=2430) people belonged to the middle class income range (40k-100k per year) and 15.2% (n = 723) were in the high income class.

About (49.4%) of the total respondents faced either physical, emotional or sexual abuse at least once before they turned to age 18. Approximately 5% respondents said that they experienced sexual abuse and 17 % admitted experiencing physical abuse at least once before the age of 18 whereas emotional abuse count exceeded 45 % of the total population. Among the individuals diagnosed with anxiety or panic disorder 13.6% (n = 87) also were recognized to have PTSD. Similarly 10.5% (n = 67) reported to have ADHD by a health care professional. It is worth to mention that more than half of the respondents who had a history of panic disorder had at least once suffered from depression in their lifetime.

The Characteristic addiction was created by merging the history of being addicted to either smoking, drinking, marijuana, or some other drugs. Among the respondents who reported to have at least one type of addiction (n = 4140), only 11.4% (n = 473) of them were diagnosed to have anxiety or panic disorder.

Table 4.2 illustrates the outcome of logistic regression taking the binary variable anxiety disorder as the dependent variable against all other characteristics of study

population. It shows the P-value of each explanatory variable along with the odds ratio of the corresponding category.

It is noticeable that the effect of gender was significant and female respondents were 2 times (95% CI 1.627-2.7) more likely to be diagnosed with panic disorder than Male respondents. White / Caucasians have higher odds of reporting such disorder compared to individuals of other races such as African American, American Indian or Alaska Native, Asian or Pacific Islander. Amid the other socio-demographic factors Income was seen to be borderline significant with higher income having higher odds or being diagnosed with anxiety, whereas effect of BMI was not found to be significant.

Table 4.2: Results showing tests of model effects and corresponding odds ratio

Characteristics of the study population	Significance	Odds Ratio	95% CI	
			Lower	Upper
Age				
Less than 30 n(%)	0.088	Reference		
More than 30 n(%)		1.224	0.97	1.544
Sex				
Male n(%)	0.000*	Reference		
Female n(%)		2.096	1.627	2.7
Race				
White n(%)	0.000*	Reference		
Others n(%)		0.513	0.38	0.694
BMI				
Underweight	0.194	Reference		

Table 4.2 continued from the previous page

Characteristics of the study population	Significance	Odds Ratio	95% CI	
			Lower	Upper
Normal		0.853	0.394	1.844
Overweight		0.676	0.306	1.489
Income				
Low		Reference		
Medium	0.057	0.733	0.564	0.953
High		0.745	0.481	1.156
Abuse				
No	0.001*	Reference		
Yes		0.665	0.527	0.841
PTSD				
No	0.000*	Reference		
Yes		2.087	1.811	4.35
ADHD				
No	0.192	Reference		
Yes		1.334	0.864	2.062
Depression				
No	0.000*	Reference		
Yes		9.857	7.752	12.535
Addiction				
No	0.325	Reference		
Yes		0.875	0.669	1.144

Table 4.2 continued from the previous page

Characteristics of the study population	Significance	Odds Ratio	95% CI	
			Lower	Upper

Note: Dependent Variable: Ever been diagnosed with Panic or Anxiety Disorder (reference category= (0) No). Model: Intercept, Age, Sex, Race, BMI, Income, Abuse, PTSD, ADHD, Depression, Addiction.
(*) significant at 5% level of significance, (**) significant at 10% level of significance.

Individuals who self reported to be identified of having PTSD or Clinical Depression also were significantly more likely to suffer from anxiety. Undergoing PTSD doubles the likelihood of having anxiety or panic disorder and Depression multiplies the chances by 10 times (95% CI 7.752-12.535) approximately [Table 4.2]. On the contrary, the effect of ADHD and Addiction did not seem to have any significant relationship with the dependent variable.

The predictor Abuse, which accounts for all three types of abuses taking 'Yes' if the respondents faced at least one type of abuse before the age of 18 was found to have a significant effect on anxiety at a level of 0.01. That is, an individual who experienced any kind of abuse in their childhood is 0.7 times (95% CI 0.527-0.841) [Table 4.2] more likely to develop panic disorder later in their life.

In order to investigate if any one of the abuses is more significant than the other, instead of using the merged abuse variable, three separate binary variables for each type of abuse were included, where 0 being the absence of abuse and 1 means the presence of any maltreatment, controlling for the other variables in the model. Results from the logistic regression are given in the Table 4.3.

Table 4.3: Relationship between Abuse with Anxiety or Panic Disorder.

Explanatory variable	Significance	Odds Ratio (95% CI)
Considering Abuse types individually		
Physical Abuse (0 = reference group)		
Happened at least once	0.023	1.33 (1.041-1.699)
Emotional Abuse (0 = reference group)		
Happened at least once	0.001	1.496 (1.176 - 1.903)
Sexual Abuse (0 = reference group)		
Happened at least once	0.189	1.334 (0.866 - 2.056)
Number of Abuse (0 = reference group)		
1 type	0.003	1.462 (1.128 - 1.895)
2 types		1.442 (1.054 - 1.972)
3 types		2.27 (1.32 - 3.905)

Note: Dependent Variable: Ever been diagnosed with Panic or Anxiety Disorder (reference category= (0) No). (*) significant at 5% level of significance.

Results suggest emotional abuse (p = 0.001) and physical abuse (p = 0.023) to have a significant effect on anxiety or panic disorder, whereas sexual abuse was found to be insignificant with the odds of having the condition is higher for emotional abusive history compared to the other types of abuse. Moreover, number of abuses was also found to be significant (p = 0.003) when regressed on the dependent variable controlling for the other variables present in the model.

Furthermore, to inspect whether any one type of abuse is affecting more than the other, three different bivariate logistic models were utilized to get the odds ratio

by regressing each type of abuse against the dependent variable.

Table 4.4: Relationship between Abuse with Anxiety or Panic Disorder.

Explanatory variable	Overall Odds Ratio (95% CI)	Males (95 % CI)	Females (95% CI)
Considering Abuse types individually			
Physical Abuse (0 = reference group)			
Happened at least once	1.799 (1.473 - 2.197)	1.449 (0.974 - 2.156)	2.039 (1624 - 2.560)
Emotional Abuse (0 = reference group)			
Happened at least once	1.496 (1.176-1.903)	1.971 (1.359 - 2.856)	2.080 (1.656 - 2.612)
Sexual Abuse (0 = reference group)			
Happened at least once	2.359 (1.664 - 3.346)	0.870 (0.317 - 2.385)	2.272 (1.611 - 3.205)
Number of Abuse (0 = reference group)			
1 type	1.886 (1.512 -2.354)	1.788 (1.155-2.770)	1.722 (1.343 - 2.209)
2 types	2.502 (1.930 - 3.244)	2.037 (1.280 - 3.240)	2.512 (1.828 - 3.415)
3 types	4.287 (2.659 - 6.912)	1.974 (0.626- 6.225)	4.432 (2.738 - 7.171)

Note: Dependent Variable: Ever been diagnosed with Panic or Anxiety Disorder (reference category= (0) No).

Accompanied by another model where the predictor indicated the number of types of abuse experienced by a respondent before the age of 18. Additionally, the same models were repeated in the case of female and male respondents separately to investigate if gender creates any discrepancy with the outcome. Results from the model which included physical abuse indicated that individuals who reported the history of childhood physical abuse have 1.799 times (95% CI 1.473-2.197) higher odds of being diagnosed with anxiety disorder. Interestingly, women have higher odds (2.039 95% CI 1.624-2.560) of having anxiety disorder than men if they were victim to childhood abuse. In the same manner, the odds for the entire sample in case of emotional abuse and sexual abuse were 1.496 (95% CI 1.176-1.903), 2.359 (95% CI 1.664-3.346). In both of this cases, females have higher odds in comparison with the male respondents. For example, the odds of having anxiety or panic disorder is more than double in the case of women who faced sexual abuse compared to men in the same category.

Additionally, respondents those who reported to have faced a certain type of abuse have approximately 1.886 times (95% CI 1.512-2.354) greater odds of having anxiety than those who didn't. Consecutively, experiencing the two types of abuse increased the odds to 2.502 (95% CI 1.930-3.244) finally undergoing all three types of abuses increased the odds by more than double in comparison to those who faced a single kind.

CHAPTER 5

CONCLUSION

Logistic regression facilitates a wide range of techniques in conducting statistical analyses. In logistic regression like any other regression technique, the primary aim is to construct an equation based on the set of explanatory variables, which as a whole would explain the variation and predict the dependent variable better.

Different methods of standardizing the coefficients assist in explaining the variation in the dependent variable and allow one to compare their contributions. In chapter 3, four different methods of standardizing the coefficients were analyzed using a practical example. From the results of the analysis, it can be seen that if the standardized values are only used in case of relative comparison of the predictors, there is not much difference between the four methods. The overall magnitude of the influence is comparatively lower for the 4^{th} method but if the influences of the predictors were ranked, the ranking was found to be the same for all four methods. It was also studied to see if standardizing the coefficients would change the performance of the model.

Results indicated that using any kind of technique to standardize the coefficient did not effect the overall accuracy of the model regardless of the method used. Although the training and testing accuraciess differed by a small amount, the AUC's were similar to the unstandardized model for all methods.

Therefore, it could be inferred that standardized coefficients can also be used for predictive modeling. Similarly, selecting any specific method for standardizing the coefficients for interpretation is completely based on how one wants to interpret it, but a method 4 would be a better approach, as method 2 which was suggested by Long and method 3, which is used in SAS, partially standardizes by only using

predictors. Both of these methods make little difference to the outcomes thus are not recommended.

Based on the findings of the logistic regression applied to a study on Add Health; survey data, the following can be inferred:

- Childhood emotional abuse was found to be the more significant contributor of anxiety or panic disorder than other types of abuse.
- Any kind of childhood abuse experience seemed to have greater affect on the female portion of the respondents in comparison to the males.
- The number of types of abuse experienced is also significant.
- Women have higher odds of developing the condition later in life. and finally.
- Having mental disorders like PTSD and depression also amplifies the chance of developing anxiety or panic disorder.

Even though a direct relationship between the predictors and the response variable can not be established from this analysis, the association present can not be disregarded. Nevertheless, a further thorough investigation could be effective in understanding the magnitude of these factors causing the discussed disorders and their remedies.

To summarize, implications of logistic regression model are wide spread in numerous sectors including Health, Demography, Economics and much more. It can also be modified to accommodate various types of data acting as a beneficial tool for the researchers.

BIBLIOGRAPHY

[Abbott, 1985] Abbott, R. D. (1985). Logistic regression in survival analysis. *American journal of epidemiology*, 121(3):465–471.

[Agresti, 2018] Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

[Agresti and Finlay, 1997] Agresti, A. and Finlay, B. (1997). Statistical methods for the social sciencesffd ed.

[Aldrich and Nelson, 1984] Aldrich, J. H. and Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Number 45. Sage.

[Allison, 1999] Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological methods & research*, 28(2):186–208.

[Bernstein et al., 2003] Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., Stokes, J., Handelsman, L., Medrano, M., Desmond, D., et al. (2003). Development and validation of a brief screening version of the childhood trauma questionnaire. *Child abuse & neglect*, 27(2):169–190.

[Berry, 1993] Berry, W. D. (1993). *Understanding regression assumptions*, volume 92. Sage.

[Berry et al., 1985] Berry, W. D., Feldman, S., and Stanley Feldman, D. (1985). *Multiple regression in practice*. Number 50. Sage.

[Cohn and Iratzoqui, 2016] Cohn, E. G. and Iratzoqui, A. (2016). The most cited scholars in five international criminology journals, 2006–10. *British Journal of Criminology*, 56(3):602–623.

[DeMaris, 1992] DeMaris, A. (1992). *Logit modeling: Practical applications*, volume 86. Sage.

[Detrano, 1989] Detrano, R. (1989). Cleveland heart disease database. *VA Medical Center, Long Beach and Cleveland Clinic Foundation*.

[Genesis, 2018] Genesis (2018). Step by step explanation of linear regression.

[Greene, 1993] Greene, W. H. (1993). and 2000. econometric analysis, 2 nd and 4 th edition.

[Harris, 2013] Harris, K. M. (2013). The add health study: Design and accomplishments. *Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill*, pages 1–22.

[Harris et al., 2009] Harris, K. M., Halpern, C. T., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and Udry, J. R. (2009). The national longitudinal study of adolescent to adult health: Research design.

[Harris and Udry, 2014] Harris, K. M. and Udry, J. R. (2014). National longitudinal study of adolescent to adult health (add health), 1994-2008 [public use](icpsr21600).

[Heatherton et al., 1991] Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., and Fagerstrom, K.-O. (1991). The fagerström test for nicotine dependence: a revision of the fagerstrom tolerance questionnaire. *British journal of addiction*, 86(9):1119–1127.

[Johnson, 2008] Johnson, D. R. (2008). Using weights in the analysis of survey data. *Presentation prepared for the Population Research Institute, Pennsylvania State University, November*.

[Kaufman, 1996] Kaufman, R. L. (1996). Comparing effects in dichotomous logistic regression: A variety of standardized coefficients. *Social Science Quarterly*, pages 90–109.

[Kleinbaum and Klein, 2010] Kleinbaum, D. G. and Klein, M. (2010). Ordinal logistic regression. In *Logistic regression*, pages 463–488. Springer.

[Long, 1997] Long, J. S. (1997). *Regression models for categorical and limited dependent variables*, volume 7. Sage.

[Lumley et al., 2004] Lumley, T. et al. (2004). Analysis of complex survey samples. *J Stat Softw*, 9(1):1–19.

[Menard, 2000] Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1):17–24.

[Menard, 2002] Menard, S. (2002). *Applied logistic regression analysis*, volume 106. Sage.

[Menard, 2004] Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician*, 58(3):218–223.

[Menard, 1995] Menard, S. W. (1995). Applied logistic regression analysis. Technical report.

[Mood, 2010] Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review*, 26(1):67–82.

[Morgan and Teachman, 1988a] Morgan, S. P. and Teachman, J. D. (1988a). Logistic regression: Description, examples, and comparisons. *Journal of Marriage and Family*, 50(4):929–936.

[Morgan and Teachman, 1988b] Morgan, S. P. and Teachman, J. D. (1988b). Logistic regression: Description, examples, and comparisons. *Journal of Marriage and Family*, 50(4):929–936.

[Peng et al., 2002] Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14.

[Petersen, 1985] Petersen, T. (1985). A comment on presenting results from logit and probit models. *American Sociological Review*, 50(1):130–131.

[Ratjen et al., 2017] Ratjen, F., Hug, C., Marigowda, G., Tian, S., Huang, X., Stanojevic, S., Milla, C. E., Robinson, P. D., Waltz, D., Davies, J. C., et al. (2017). Efficacy and safety of lumacaftor and ivacaftor in patients aged 6–11 years with cystic fibrosis homozygous for f508del-cftr: a randomised, placebo-controlled phase 3 trial. *The Lancet Respiratory medicine*, 5(7):557–567.

[Roncek, 1991] Roncek, D. W. (1991). Using logit coefficients to obtain the effects of independent variables on changes in probabilities. *Social Forces*, 70(2):509–518.

[Straus et al., 1998] Straus, M. A., Hamby, S. L., Finkelhor, D., Moore, D. W., and Runyan, D. (1998). Identification of child maltreatment with the parent-child conflict tactics scales: Development and psychometric data for a national sample of american parents. *Child abuse & neglect*, 22(4):249–270.

[Tietjen, 2016] Tietjen, G. E. (2016). Childhood maltreatment and headache disorders. *Current pain and headache reports*, 20(4):26.

[Williams, 2014] Williams, R. (2014). Survey sampling and weighting. In Culyer, A. J., editor, *Encyclopedia of Health Economics*, pages 371–374. Elsevier, San Diego.

[Winship, 1984] Winship, Christopher, M. R. D. (1984). Regression models with ordinal variables. *American sociological review*, pages 512–525.

[Wright, 1995] Wright, R. E. (1995). Logistic regression.