Predicting Dengue Epidemic disease status using different exposure variables

Dawit Wolde

Department of Biostatistics
University of Kansas, USA
May 10, 2023

Title

To predict understand the association between dengue epidemics disease and several exposure variables on selected 196 people.

Abstract

It is important to study the relationship between dengue epidemics disease and exposure factors to prevent the disease. Dengue epidemics disease was a major problem in pacific coast of Mexico in 1980s. According to Journals Plos, dengue disease is a tropical mosquito born viral disease. The disease was widespread in the cost of Mexico, and it is transmitted by blood feeding mosquitoes. The most common symptom of this viral disease is mild or severe fever. Studying the association between the disease and exposure factors helps for prevention as it does not have a medicine for treatment. The blood feeding mosquitos are spread and breed easily in places where there a standing water.

The four predictor variables in this study are Age of a person, socioeconomic status, sector, and saving account status. The response variable for this study is disease status.

Introduction

Socioeconomic status, and sector are believed to have a direct correlation with having dengue epidemics disease. According to world health organization, Socioeconomic status is an important predictor variable as it can be directly associated to disease prevention. People with a low socioeconomic status are more likely to get the disease compared to people with high socioeconomic status. other important predictor variables are saving status and sector. It is important to find the association between the response and the predictor variable so that dengue disease prevention can be understood and improved. The main purpose of the study is to observe

the association between dengue disease (response variable) and 4 other predictor variables.

Primary Analysis Objective

Data source

To observe the association between the response variable dengue disease and the predictor exposure variables and get the best predictor variables.

Materials and Methods.

The dataset is obtained from a professor from the university of Kansas and the original data set is collected by a retrospective survey from a pacific coast of Mexico. The dataset is collected from 196 selected group of people using surveys. The dataset has a dependent variable **disease status** (**dis)** where 0 without disease and 1 with disease. Independent variables included in the dataset are age of a person, socio-economic status , sector, and saving account.

age – Age of a person
ses -- Socio-economic status
sector -- Sector
dis -- disease status.
save – saving account status.

Variable Name	Data Type	Description	Example
age	number	Age of a person	23
ses	number	Socio economic status	1 (upper)
sector	number	sector	2 (sector 2)
dis	number	Disease status	0 (without disease)
save	number	Saving account status	0(does not have saving account)

Table 1: Response and predictor variables table

Statistical analysis

Primary objective analysis

In order to have the model fit good, it is important understand and study the independent variables. By analyzing each variable, skewness and outliers in the data can be tested which are a great way to observe degree of asymmetry. Such step help achieve a better fitting model and suggest if any transformation is needed. The best way to perform the analysis is to use both boxplot and histogram.

Box plot analysis

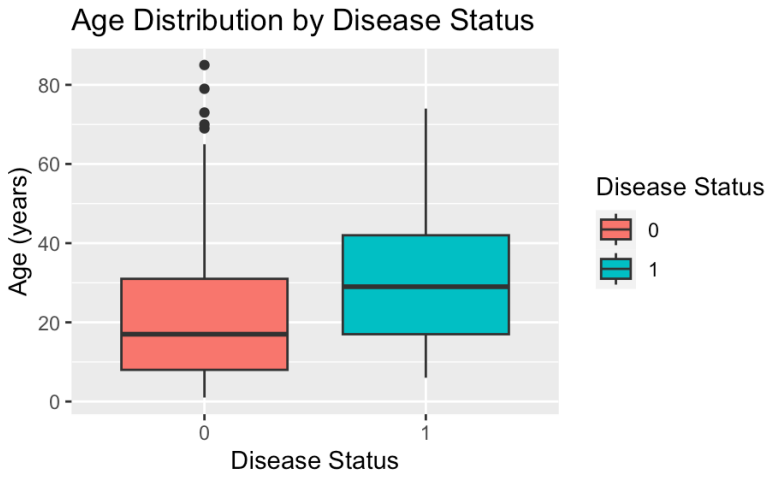
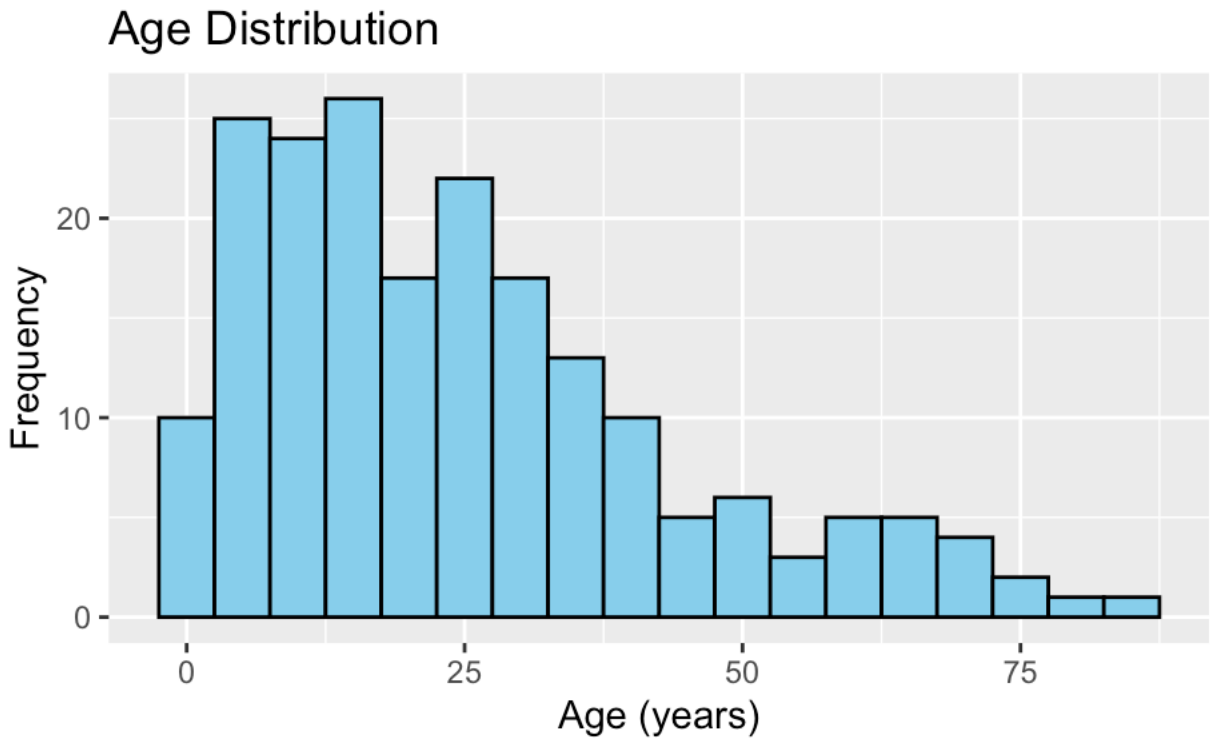
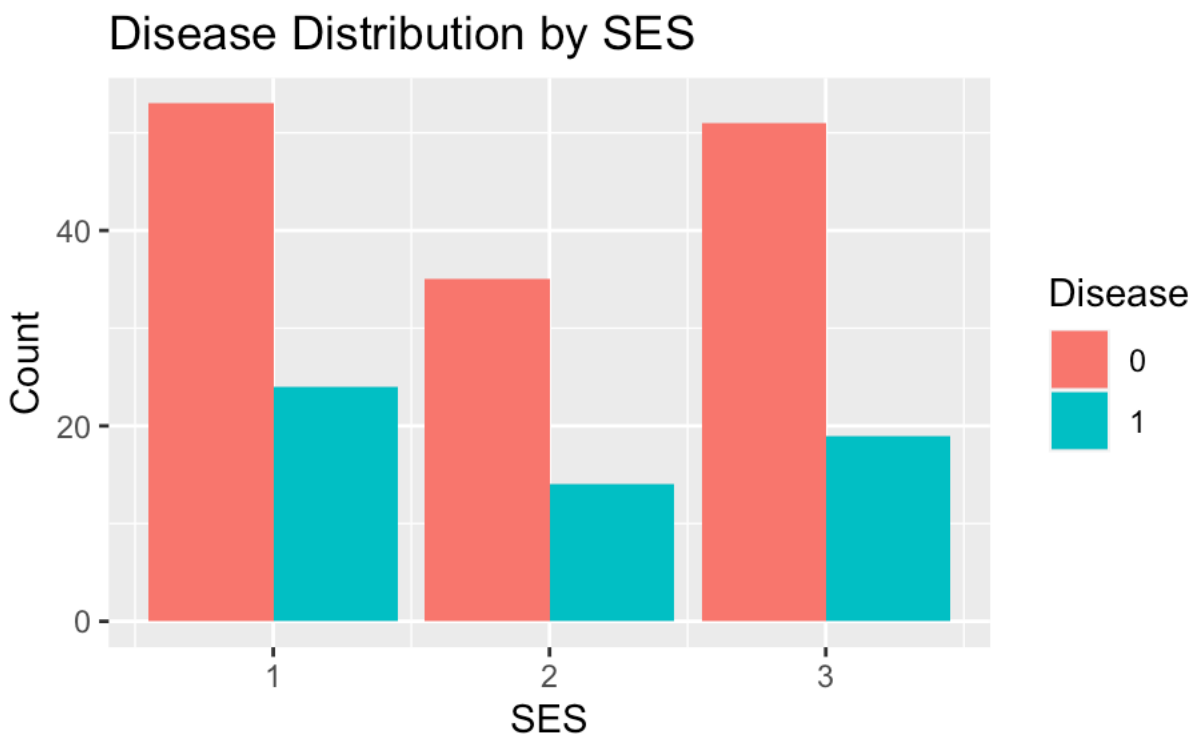


Figure 1: Boxplot analysis of variables

Analysis of potential predictors. Figure 1 shows boxplot for each of the predictor and response variable. Box plot of age seems to have a few outliers and have skewness.

Distribution analysis of the independent variables




Disease Distribution by SES
Count
40
20
0
1
2
3
SES
Disease
0
1

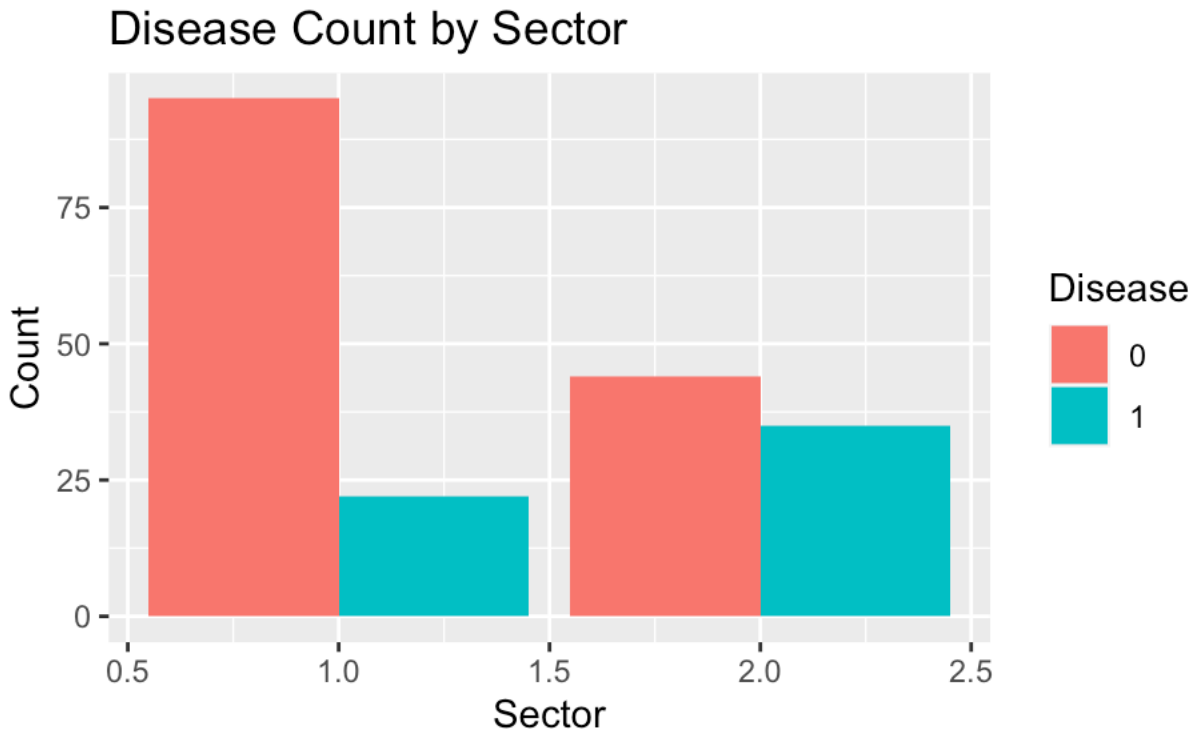
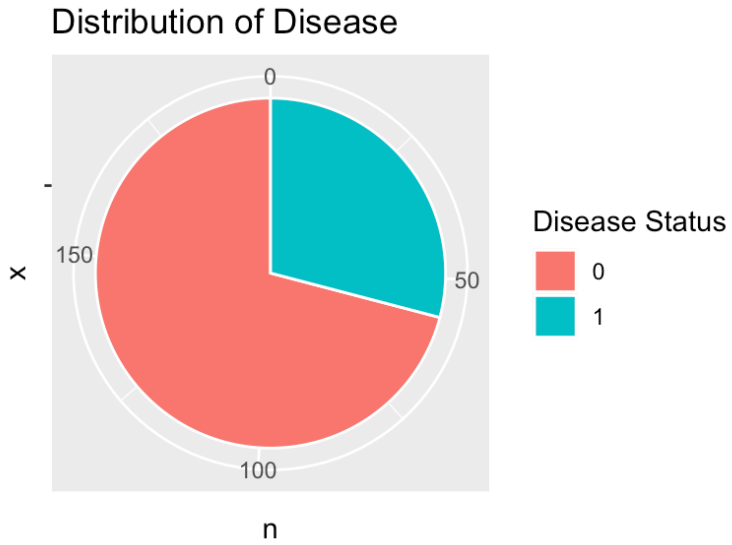

Disease Count by Sector
Count
75
50
25
0
0.5
1.0
1.5
2.0
2.5
Sector
Disease
0
1

Figure 2 Distribution plot of each predictor variables

Figure 1: shows bar charts and distribution for each independent variables.

Analysis of the dependent variable disease status

(a) Pie chart for disease variable



(b) Heatmap correlation map

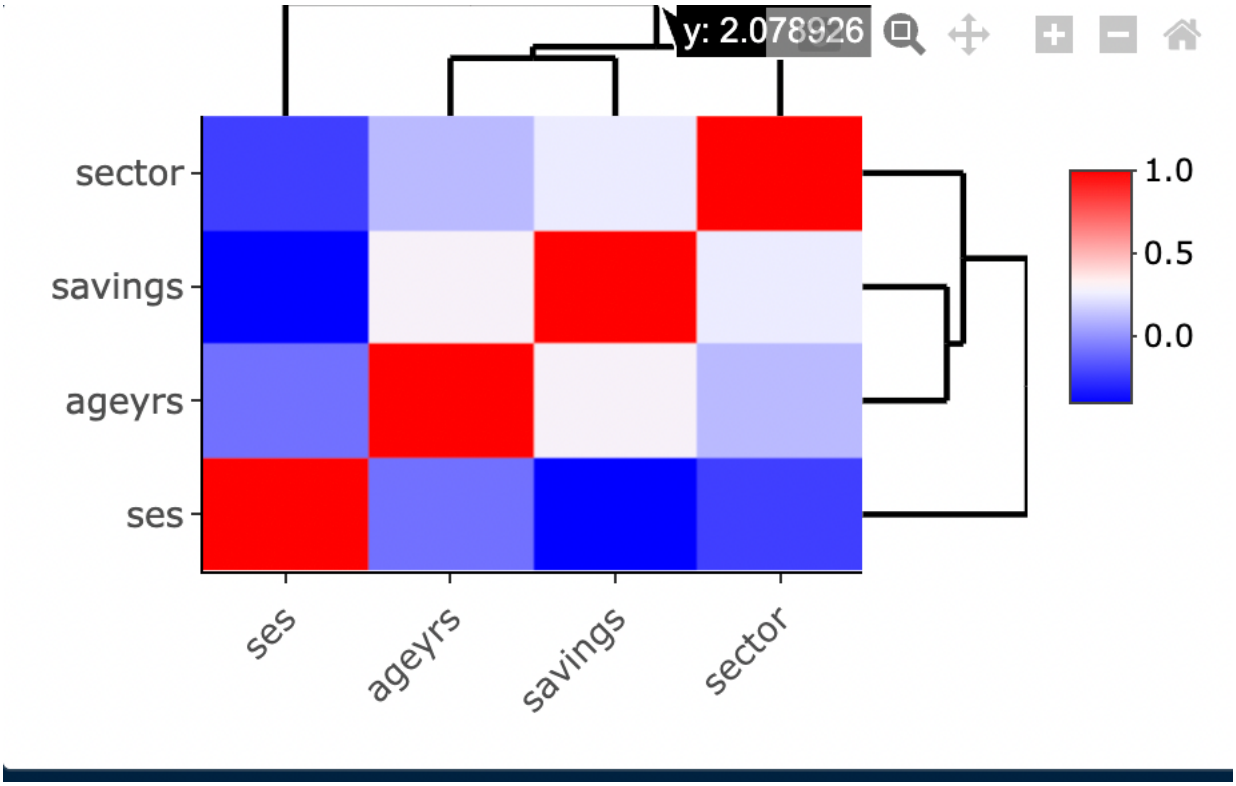


Figure 2: distribution of response variable disease

The pie chart and heatmap are the graphical representation for the response and predictor variable.

Method and Goodness of Fit Test

Generalized linear model (GLM), Boxplot, Histogram, AIC, BIC, p-value and chi-square were used to perform the analysis.

All of the statistical analysis was done using statistical software R version 4.2.2(2022-07-02).

Results

The association between the response variable disease and each of the independent variable, chi-square test and generalized linear model were used.

The hypothesis in this study:

Null hypothesis: There is no significant relationship between the response variable(disease) and the predictor variable (age, ses, sector, and save).
Alternative hypothesis: There is a significant relationship between the response and predictor variables.

The level of significance(α) used is 0.05.

If the p-value from chi-square test is less than 0.05, there is enough evidence to reject the null hypothesis. If the test statistics that is denoted as X^2 value is large, it indicates that there is a strong association between the response and predictor variables. Both X-squared and p-value are important values to use for conclusion of association between the response and predictors.

In generalized linear model summary, z value represents a test statistic for null hypothesis and a large z value indicates that there is strong evidence to reject the null hypothesis. P-value implies if each predictor variables are statistically significant for predicting the response variable. In order to study goodness of fit, AIC values are explored where a small value is a good fit.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.650188   0.838074  -4.355 1.33e-05 ***
ageyrs       0.027405   0.009115   3.007 0.002642 **
ses          0.116707   0.217156   0.537 0.590966
sector       1.240703   0.354229   3.503 0.000461 ***
savings     -0.038304   0.395948  -0.097 0.922933
---
```

Table 2: important statistical numerical values for variables

Effect of age on predicting dengue Epidemic disease

In order to study the association, alpha significance level (0.05) is used. The p-value is then compared to the alpha significance level. The p-value for age from generalized linear model summary is 0.0026 which suggests that we have enough

evidence to reject the null hypothesis implying that the variable age is statistically significant for predicting dengue epidemic disease. The z value is 3.0 indicating there is a strong association. Chi-square test is made for each predictor variable to study individual association. Although doing chi-square test for each variable does not account other variables, it is important to see the association between the response variable and each predictor variable individually. The p-value from chi-square test is 0.326 and the X-squared is 65.42.
While X-squared suggests that there is a strong association, the p-value indicates that the variable is not statistically significant. The odds ratio is 1.03 implying that the odds of having dengue epidemic disease increases by a factor of 1.03 for one unit increase in age.

Effect of social economic status(ses) on predicting dengue Epidemic disease

Alpha significance level (0.5) is used to analyze the results in this study. The p-value from generalized linear model summary for social economic status is 0.59. The p value is greater than the alpha significant level indicating that the predictor social economic status (lower, middle, and upper) is not a statistically significant. The z value (0.5) is also small indicating a weak association between the response variable and the predictor variable. The p-value from chi-square that is calculated independently for the predictor variable is 0.86 along with X-squared value of 0.29. The p-value show the predictor is not statistically significant and the X-squared shows a weak association between dengue disease and ses. 1.12379058. Odds ratio is 1.124 meaning that the middle class has 1.124 odds of having dengue epidemic disease compared to the upper-class status.

Effect of sector on predicting dengue Epidemic disease

Using alpha significant level 0.05, the association between dengue epidemic disease and sector within city (sector 1 and 2). The p-value for sector is 0.00046 which is very small and less than the alpha significance level indicating that there is enough evidence to reject the null hypothesis and sector and dengue epidemic disease are strongly associated. P-value from chi-square test is 0.0002 implying that there is enough evidence to reject the null hypothesis and therefore sector is statistically significant for predicting the disease. X-squared is 13.658 indicating

there is a strong association between response disease and sector. The odds ratio for sector is 3.45 meaning the odds of having a disease for sector one is 3.45 than sector two.

Effect of saving account status on predicting dengue Epidemic disease

The alpha significance level used for analyzing the p-value is 0.05. The p-value for savings account is 0.9 implying that there is no enough evidence to reject the null hypothesis and saving account status is not statistically significant for predicting the disease. The p-value from chi-square is 0.1662 hence we cannot reject the null hypothesis as the alpha significance level is less than the p-value. The X-squared (0.166) also suggests that there is no strong association between saving account status and dengue epidemic disease. The odds ratio for the predictor saving account status is 0.96 indicating that the odd of getting the disease is 0.96 lower for people without saving account compared to people with saving account.

Model Selection

Hypothesis for model selection

Null hypothesis: The reduced model gives a significantly better fit to the data.
Alternative hypothesis: The reduced model does not give a significantly better fit to the data.
Alpha significance level =0.05

In order to select a model, likelihood ratio test was performed. To perform the test, full model with all predictor variables and reduced model with reduced variables were made. After performing likelihood ratio tests for each reduced model, the full model and the reduced model were compared in ANOVA.

The p-value is 0.002 for reduced model with only socio-economic status, sector, and savings status. P-value suggests that there is enough evidence to reject the null hypothesis thus the reduced model does not give a significantly better fir to the data and age is an important predictor.

The p-value for reduced model with only age, sector and savings status is 0.59. This suggests that there is no enough evidence to reject the null hypothesis and

the reduced model gives a significantly better fit to the data which indicates that the predictor variable socio economic status is not an important predictor.

The p-value for reduced model with only age, socio economic status, and saving status is 0.00035. The p-value is very small and is less than the alpha significance level indicating that there is enough evidence to reject the null hypothesis. Therefore, the p-value implies that the predictor variable sector is an important predictor.

The p-value for reduced model with only age, socio economic status, and savings is 0.92 which is greater than the alpha significance level (0.05). The p-value suggests that there is enough evidence to reject the null hypothesis. Therefore, the reduced model does not give a significantly better fit to the data and indicates that the predictor variable saving status is not an important predictor.

AIC and BIC analysis for model selection

model	AIC	BIC
mull model	221.25	237.64
model without age	228.76	241.87
model without ses	219.54	232.65
model without sector	232.04	245.15
model without savings status	219.26	232.37

Table 3: AIC and BIC values

The lowest AIC and BIC implies a better fit. Therefore, the model without saving account status and socio-economic status is a better fit model.

Discussion and Conclusion

In this study, many methods were used to analyze the association between age, socio-economic status, sector, and saving account status. In order to perform the analysis, methods such as bar chart, boxplot, generalized linear model, chi-square test, likelihood ratio, ANOVA were used. Statistical values such as P-value, z-value,

X-square, AIC and BIC were used to perform the analysis and select the best models.
The best fit model for this study is model with only age and sector. A separate model that only contains age and sector was made. The p-value for age is 0.0019 and 0.00045 for sector indicating that there is enough evidence to reject the null hypothesis, and both are statistically significant for predicting the response variable dengue disease. Finally, important statistical methods and analysis techniques were used to select age and sector as best predictors for dengue disease.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.34135	0.59206	-5.644	1.66e-08	***
ageyrs	0.02681	0.00865	3.100	0.001936	**
sector	1.18169	0.33696	3.507	0.000453	***

Table 4: Statistics table for selected models

```r
library(psych)
library(readxl)
library(lmtest)
library(rcompanion)
library(car)
library(faraway)
library(epiDisplay)
library(TraMineR)
library(dplyr)
library(leaps)
library(tidyverse)
library(questionr)
library(ggplot2)
library(readr)
library(caret)
library(lattice)
```

```
---
title: "Untitled"
author: "Dawit Wolde"
date: "2024-01-28"
output: word_document
---

```{r}
library(readr)
disease <- read_csv("~/Desktop/stat 835/disease.csv")
head(disease)
disease<-disease[,-1]
```

```{r}
library(ggplot2)
library(dplyr)

# Visualization 1: Age Distribution
ggplot(disease, aes(x = ageyrs)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Age Distribution", x = "Age (years)", y = "Frequency")

```
```

```r
# Visualization 2: Disease by SES
ggplot(disease, aes(x = ses, fill = as.factor(disease))) +
  geom_bar(position = "dodge") +
  labs(title = "Disease Distribution by SES", x = "SES", y = "Count", fill = "Disease")

# Visualization 3: Sector-wise Disease Count
ggplot(disease, aes(x = sector, fill = as.factor(disease))) +
  geom_bar(position = "dodge") +
  labs(title = "Disease Count by Sector", x = "Sector", y = "Count", fill = "Disease")

# Visualization 4: Savings and Disease
ggplot(disease, aes(x = savings, fill = as.factor(disease))) +
  geom_bar(position = "dodge") +
  labs(title = "Disease Distribution by Savings", x = "Savings", y = "Count", fill = "Disease")
# Box plot for Age distribution by Disease
ggplot(disease, aes(x = as.factor(disease), y = ageyrs, fill = as.factor(disease))) +
  geom_boxplot() +
  labs(title = "Age Distribution by Disease Status",
       x = "Disease Status",
       y = "Age (years)",
       fill = "Disease Status")
```

```r
#pie chart for disease distribution
disease %>%
  count(disease) %>%
  ggplot(aes(x = "", y = n, fill = as.factor(disease))) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y") +
  labs(title = "Distribution of Disease", fill = "Disease Status")
#
install.packages("heatmaply")
library(heatmaply)

heatmaply(cor(disease[, c("ageyrs", "ses", "sector", "savings")]),
          col = colorRampPalette(c("blue", "white", "red"))(20))

##
head(disease)

disease$ses<-as.factor(disease$ses)
disease$sector<-as.factor(disease$sector)
disease$disease<-as.factor(disease$disease)
disease$savings<-as.factor(disease$disease)
```

```r
#statistical analysis
glm.fit<-glm(disease~ageyrs + ses + sector + savings, data =
disease,family = binomial(link = "logit"))
sm_all.fit<-summary(glm.fit)
sm_all.fit
#odds ratio
exp(coef(glm.fit))
#chi-square test
chisq.test(disease$disease,disease$ageyrs,disease$ses,disease
$sector,disease$savings)
chisq.test(disease$disease,disease$ageyrs)
chisq.test(disease$disease,disease$ses)
chisq.test(disease$disease,disease$sector)
chisq.test(disease$disease,disease$savings)
```

```r
## model selcetion
# Fit the full model with all predictor variables
full.model <- glm(disease~ageyrs + ses + sector + savings,
data = disease, family = "binomial")
# likelihood ratio tests for  reduced models
reduced.model1 <- glm(disease~ses + sector + savings, data =
disease, family = "binomial")
anov1 <- anova(reduced.model1, full.model, test = "Chisq")
reduced.model2 <- glm(disease~ageyrs + sector + savings, data
= disease, family = "binomial")
anov2 <- anova(reduced.model2, full.model, test = "Chisq")
reduced.model3 <- glm(disease~ageyrs + ses + savings, data =
disease, family = "binomial")
anov3 <- anova(reduced.model3, full.model, test = "Chisq")
reduced.model4 <- glm(disease~ageyrs + ses + sector, data =
disease, family = "binomial")
anov4 <- anova(reduced.model4, full.model, test = "Chisq")
# likelihood ratio test results
anov1
anov2
anov3
anov4
```

```r
#AIC for model selection
full.model$aic
AIC(reduced.model1)
AIC(reduced.model2)
AIC(reduced.model3)
AIC(reduced.model4)
#BIC for model selection
BIC(full.model)
BIC(reduced.model1)
BIC(reduced.model2)
BIC(reduced.model3)
BIC(reduced.model4)
#glm for selected models
selectedmodel <- glm(disease~ageyrs + sector, data = disease,
family = "binomial")
summary(selectedmodel)
```

References

Dantés, H. G., Farfán-Ale, J. A., & Sarti, E. (n.d.). *Epidemiological trends of Dengue disease in Mexico (2000–2011): A systematic literature search and analysis*. PLOS Neglected Tropical Diseases. Retrieved April 11, 2023, from https://journals.plos.org/plosntds/article?id=10.1371%2Fjournal.pntd.0003158

World Health Organization. (1970, January 1). *World health statistics 2019: Monitoring Health for the sdgs, sustainable development goals*. World Health Organization. Retrieved April 11, 2023, from https://apps.who.int/iris/handle/10665/324835

Carreto, C., Gutiérrez-Romero, R., & Rodríguez, T. (2022, October 27). *Climate-driven mosquito-borne viral suitability index: Measuring risk transmission of Dengue, chikungunya and Zika in mexico - international journal of health geographics*. BioMed Central. Retrieved April 11, 2023, from https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-022-00317-0