Detection of Dengue using Statistical Modelling

STAT 823: Spring Class Project, 2021

Bibekananda Mishra

Department of Biostatistics and Data Science
University of Kansas, USA

May 16, 2021

Contents

List of Tables

List of Figures

Abstract

We investigate the parameters that plays the decisive role in predicting the incidence of Dengue disease in a certain population. We conclude using Logistic regression model that Age and part of the city (Sector) plays important role with 75.5% accuracy in deciding whether a person under consideration has Dengue or not.

Introduction

We are interested in analysing the data collected for the analysis of detection of Dengue epidemics in the pacific coast of Mexico. The data was collected for 196 persons on various factors including their socio-economic status, which part of the city they belong to and whether they have the disease. We fit the data in to various statistical models and choose the best among them. Our analysis shows that the 'sector' parameter i.e. which part of city the subjects belong to is the most important parameter. However, the final model also depends upon the age of the subjects.

The data was presented in the research paper "Dengue Epidemics on the Pacific Coast of Mexico." by H.G. Dantes, J.S. Koopman, C.L. Addy, et al. published in International Journal of Epidemology, 17 (1988), pp. 178-86. I have referred to the book "Applied Linear Statistical Models" by Kurtner et al. and the notes of Dr. Dong for analysis on this data.

Primary Analysis Objectives

Know the most important parameters that affects the disease incidence in a subject.

Secondary Analysis Objectives

Analyse the statsical models that are being used and check the validity of applying such models with the given data.

Materials and Methods

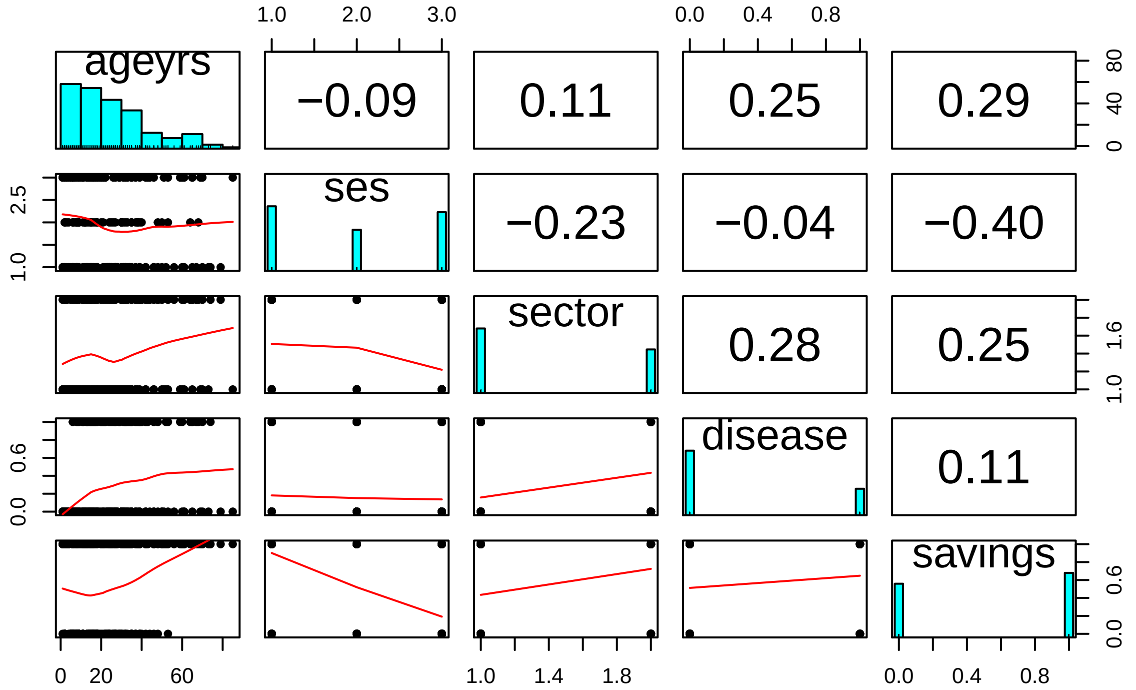
Statistical Analysis

We use two statistical models: Linear regresssion and Logistic regression model to verify the

Primary Objective Analysis

First we would like to know whether there is any significant relationship between any two variables/measurements considered in the study. For this, we compute the correlation coffiecients of the associated paarmeters in the analysis and plot the scatter plot for each variable vs another variable. See figure-1.

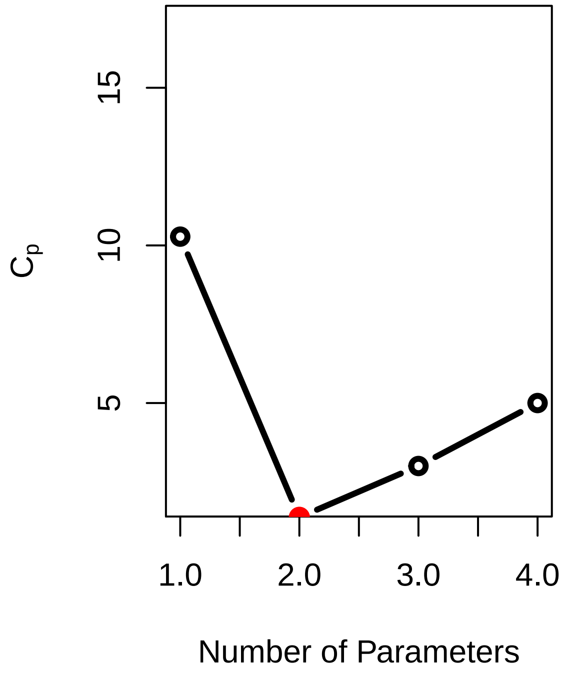
Figure−1



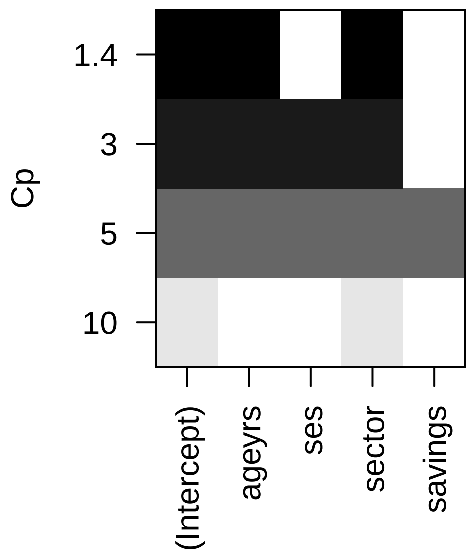
From the graphs above and correlation coefficients, it is apparent that there is no significant correlation between any of the two variables.

Next we use Cp values to find the best unbiased model or the models having the most effective variable useful for analysis. See figure 2(a) and 2(b) below.

Figure−2(a) Cp variable Selectio



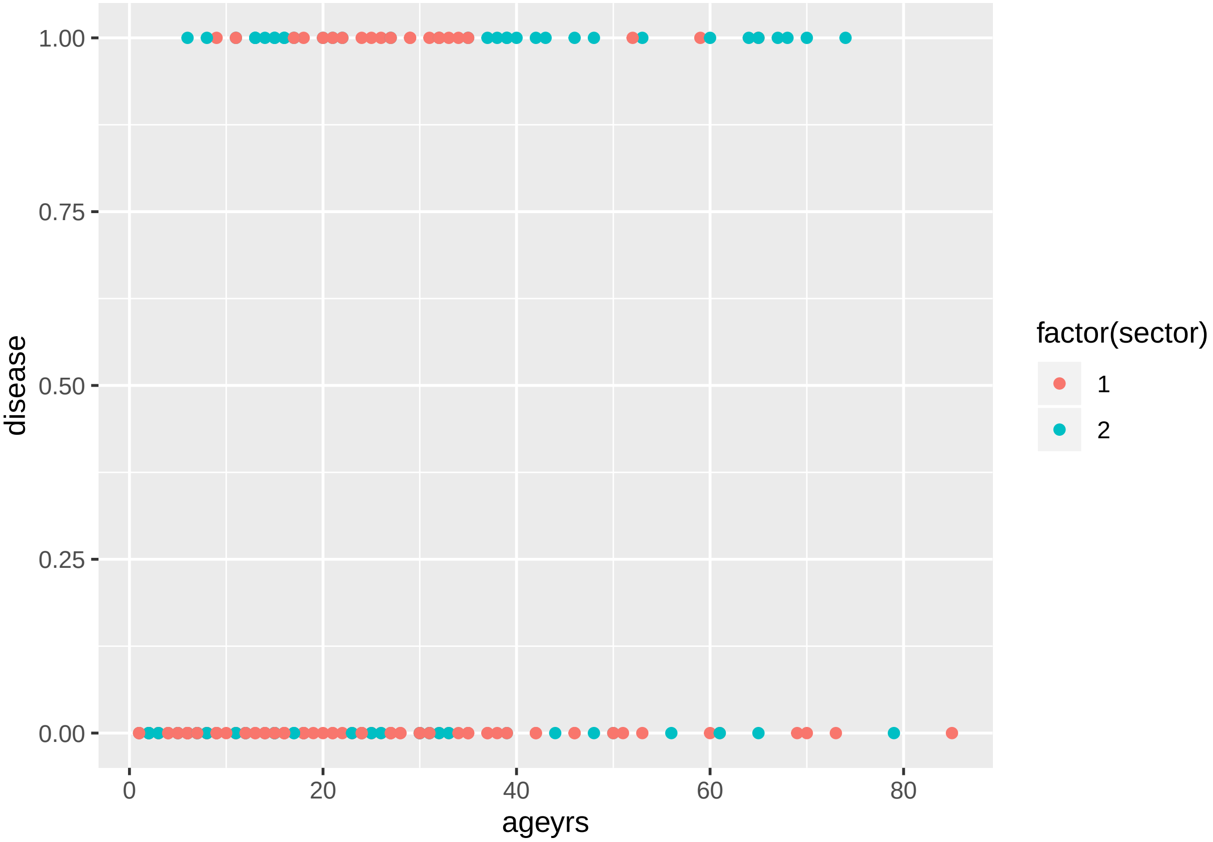
Figure−2(b)

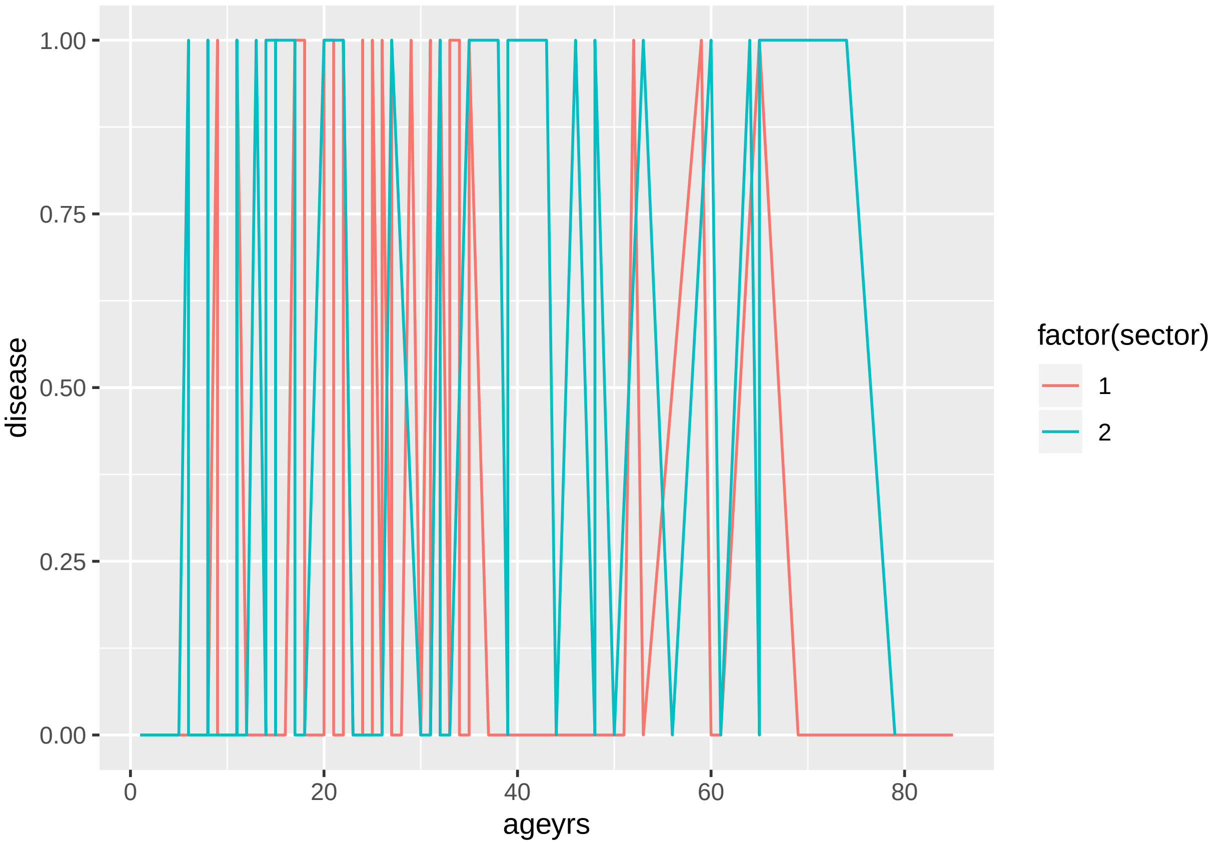


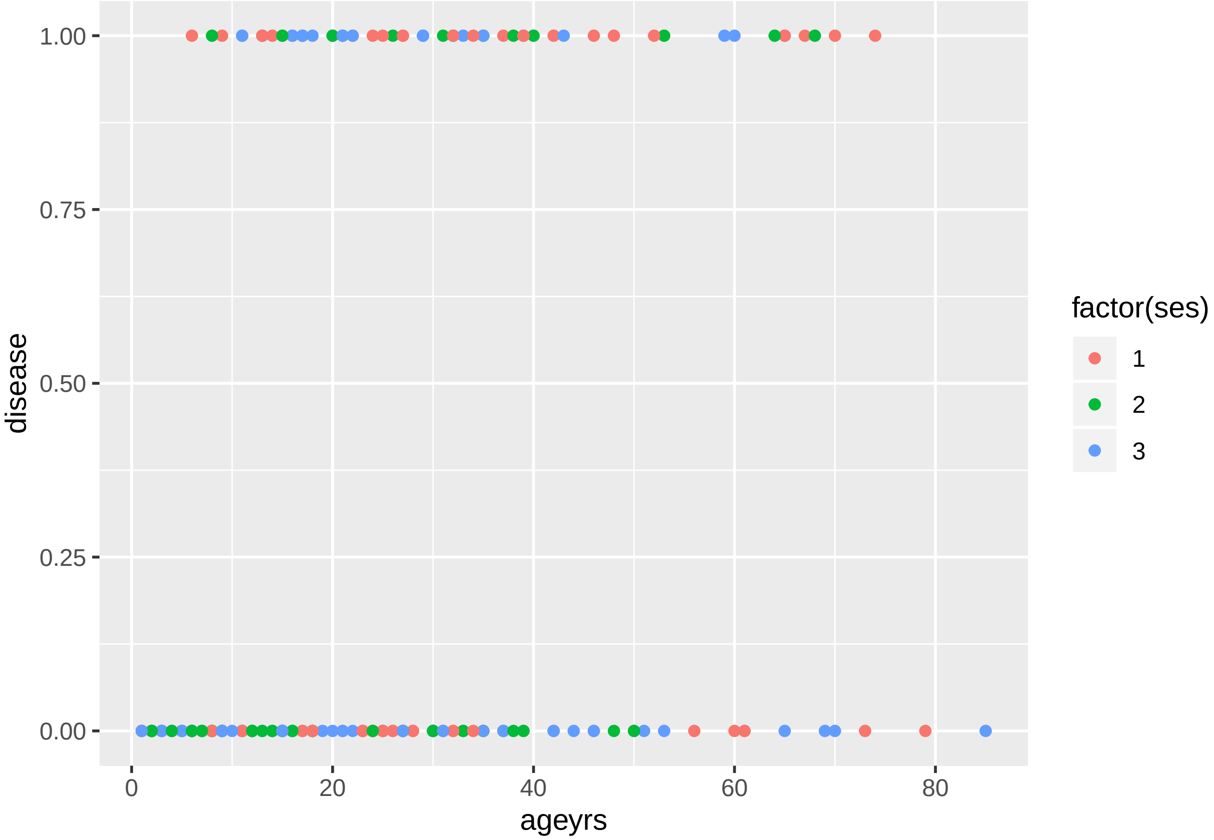
From the plots above, we choose the linear model having the C_p value closest to the (p+1)-value so that the resultant model will be least unbiased. We set three models:

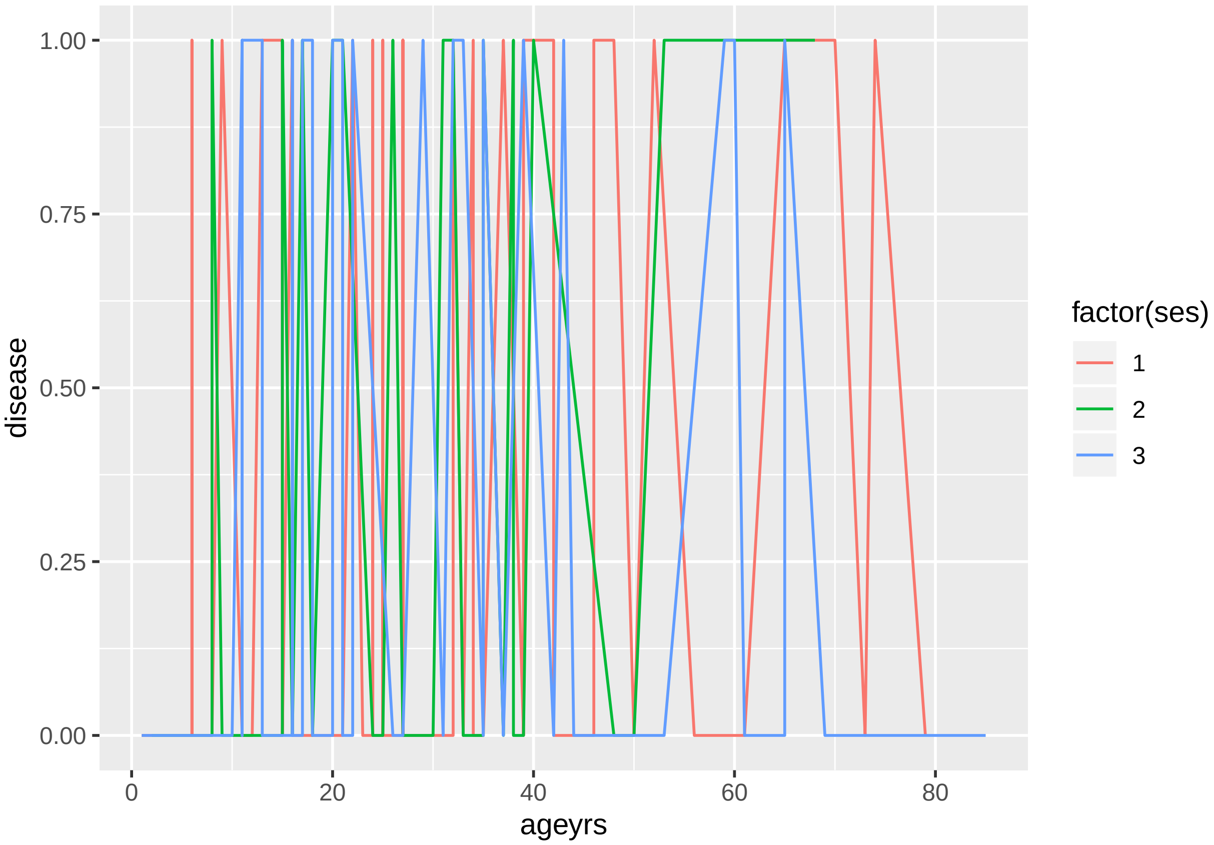
[a] Model-1 having two variables: age and sector. [b] Model-2 having three variables: age, sector and socioeconomic status. [c] Model-3 hahing all the variables.

See the following plots to figure out the relationship between some of the variables.




disease
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
ageyrs
factor(sector)
1
2


disease
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
ageyrs
factor(ses)
1
2
3



The linearity is not very clear from the plots. But we can see the sector variable playing an crucial role. With the increase in age, there is some increase in the incidence of the disease.

Now we will consider the above three models and find the validity of the linearity among them.

```
## [1] 2.353e-06
```

```
## [1] 8.5497e-06
```

```
## [1] 2.937e-05
```

```
## $r.squared
## [1] 0.12567
```

```
## $r.squared
## [1] 0.12737
```

```
## $r.squared
## [1] 0.12739
```

From the p-values of the three models, it is clear that the linear models are better. However, if we look at the R-values, it is hardly around 0.13 implying that the models explain only around 13% of the variability of the data. Since this is an extremely poor performance of the models, we have to explore some better models. Since the output is binary, a better model to consider is logistic regression model in which the outcomes are probabilities (of having the disease). We will describe the models below.

Before that, we will split the dataset into two subsets: trained and test datasets (with 3:1 split). We will fit various logistic regression models in to the train dataset and measure the efficacy in each case. And then we will measure the performance of these models against the test dataset to pick up the best model.

```
##
## glm.pred4  0  1
##         0 97 31
##         1  7 12
```

```
##
## glm.pred5  0  1
##         0 96 31
##         1  8 12
```

```
##
## glm.pred6  0  1
##         0 96 31
##         1  8 12
```

From the above tables, we see that the models performance is almost similar though performance of model-4 having two variables (age and sector) is slightly better than the other two. Since it also has the least number of variables, it is the most preferable one.

Secondary Analysis

test.pred4 0 1 0 34 11 1 1 3

test.pred5 0 1 0 33 11 1 2 3

test.pred6 0 1 0 33 11 1 2 3

We now fit the three models to test data and see their performance. We notice again that the model-4 with age and sector varibale performs slightly better than the other two though all of them have comparable efficiency in predicting disease.

Model Assumptions

All inferences are conducted using $\alpha = 0.05$ unless stated otherwise. The model assumes that the resultant probability lies between 0 and 1 which is true as we have checked. Moreoevr, the error should be approximately normal and the variance is constant which is checked easily.

Results

The final model for prediction of disease is Model-4 which is based on logistic regression model with variable Age and Sector. The accuracy of the three models considered above on the test dataset are as follows:

1] Model4 (with two variables)- 75.51% 2] Model5 (with three variables)- 73.47% 3] Model6 (with all the variabels)- 73.47%

Table 1: Results Using Xtable

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.51	0.70	-5.00	0.00
ageyrs	0.03	0.01	2.76	0.01
sector	1.25	0.39	3.19	0.00

So model-4 is our best model with the current analysis. If we look at the coefficients of age and sector, they are 0.0293 and 1.2468 respectively. That means with the increase of each unit in age, the probability of getting the disease increases by 0.0293 unit. Similarly, with sector being 2 instead of 1, the probability changes by 1.2468 unit. So, the probability of incidence of Dengue is higher in Sector-2 of the city.

Discussion and Conclusion

Age and sector of the city are the two most important indicators for the incidence of the disease. While age is not a surpise indicator considering the factor that with increase in age after certian level the immunity decreases, sector or part of city is also an important indicator with available sanitation facility. Interesting thing is socioeconomic factor is not a decider and same is true for patients having savings account. The efficiency I would say is still low at 75%. With allocation of more interaction terms, we can build a better model.

Appendix: R-code

```
library(psych)
library(readxl)
disease <- read_excel("~/Downloads/disease.xlsx")
disease<- disease[,-1]
#View(disease)
pairs.panels(disease, density = FALSE, ellipses = FALSE, main = "Figure-1")

library(leaps)
mod0 <- regsubsets( disease~ ., data = disease)
cp_min <- which.min(summary(mod0)$cp)
par(mfrow = c(1, 2))
plot(1:4, summary(mod0)$cp, type = "b", ylab = expression(C[p]), ylim = c(2, 17), lwd = 3, mai
points(cp_min, summary(mod0)$cp[cp_min], col = "red", cex = 2, pch = 20)
plot(mod0, scale = "Cp", main = "Figure-2(b)")

library(ggplot2)
ggplot(disease, aes(x=ageyrs, y=disease)) + geom_point(aes(color= factor(sector)))
ggplot(disease, aes(x=ageyrs, y=disease)) + geom_line(aes(color= factor(sector)))

ggplot(disease, aes(x=ageyrs, y=disease)) + geom_point(aes(color= factor(ses)))
ggplot(disease, aes(x=ageyrs, y=disease)) + geom_line(aes(color= factor(ses)))


library(car)
mod1<- lm(disease~ageyrs+ sector, disease)
mod2<- lm (disease~ ageyrs+ sector+ ses, disease)
mod3<- lm(disease~., disease)
linearHypothesis(mod1, c("ageyrs=0", "sector=0"))[2,6]
linearHypothesis(mod2, c("ageyrs=0", "sector=0", "ses=0"))[2,6]
linearHypothesis(mod3, c("ageyrs=0", "sector=0", "ses=0", "savings=0"))[2,6]

require(caTools)
library(ROCR)
set.seed(101)
sample= sample.split(disease$disease, SplitRatio = 0.75)
train= subset(disease, sample == TRUE)
test = subset(disease, sample== FALSE)

mod4<- glm(disease~ageyrs+ sector, data= train, family = binomial)
#check error rate
glm.probs4=predict(mod4,type="response")
glm.pred4=rep(0,147)
glm.pred4[glm.probs4 >.5]=1
pred4<- prediction(train$disease, glm.pred4)
table(glm.pred4, train$disease)
```

```
mod5<- glm(disease~ageyrs+ sector+ ses, data= train, family = binomial)
glm.probs5=predict(mod5,type="response")
glm.pred5=rep(0,147)
glm.pred5[glm.probs5 >.5]=1
pred5<- prediction(train$disease, glm.pred5)
table(glm.pred5, train$disease)

mod6<- glm(disease~., data= train, family = binomial)
glm.probs6=predict(mod6,type="response")
glm.pred6=rep(0,147)
glm.pred6[glm.probs6 >.5]=1
pred6<- prediction(train$disease, glm.pred6)
table(glm.pred6, train$disease)

test_mod4<- mod4$coefficients[1] + mod4$coefficients[2]* test$ageyrs+ mod4$coefficients[3]* tes
test.pred4=rep(0,49)
test.pred4[test_mod4 >.5]=1
test4<- prediction(test$disease, test.pred4)
table(test.pred4, test$disease)

#mod5
test_mod5<- mod5$coefficients[1] + mod5$coefficients[2]* test$ageyrs+ mod5$coefficients[3]* tes
test.pred5=rep(0,49)
test.pred5[test_mod5 >.5]=1
test5<- prediction(test$disease, test.pred5)
table(test.pred5, test$disease)

#mod6
test_mod6<- mod6$coefficients[1] + mod6$coefficients[2]* test$ageyrs+ mod6$coefficients[3]* tes
test.pred6=rep(0,49)
test.pred6[test_mod6 >.5]=1
test6<- prediction(test$disease, test.pred6)
table(test.pred6, test$disease)
```