ST 3008
Applied Statistical Models
Group Project Report
Group Number 04

S. Raweesha Nethmini S14499
Shameena Wickramathilake S14518
M.R.N. Fernando S14429
U.A.B.M. Gunasekara S14330
T.D. Agrani Nimshani S14487
7-15-2021

Introduction

Stroke Dataset is a portion of a 10-year study conducted by the American Heart Association which provides data on how age, blood pressure and smoking habits relate to the risk of strokes. In this report we are modeling the data to find the best fitted model to check which independent variables affect the risk of stroke and to interpret the final results and conclusions of regression analysis.

Objectives

- To check whether the patient's age, level of the blood pressure and whether the patient is a smoker or not a smoker have a significant impact on the patient's risk of having a stroke.
- To find the best fitted model to predict a patient's risk of having a stroke during the next 10-year period when the corresponding values for the predictors are known.

Methodology & Results

According to our "Stroke" dataset there is one response variable (Probability of risk) and three predictor variables (Age, Pressure & Smoker). We decided that "Multiple linear regression model" would be the best fitted model to analyze the given dataset because it basically describes how a single response variable depends linearly on a number of predictor variables. Here we fitted our model by using "**R software package**". The code is attached in Appendix 02 at the end of the project report. Then we briefly discuss the method & results of developing an estimated equation that relates risk of a stroke by going through each step in multiple linear regression method.

Step 01: Estimation of regression coefficient & interpretation.

Firstly, we imported the given dataset (X4_Stroke) into R studio. Then we identified that there is a qualitative variable (Smoker) which have two levels as "Yes"&"No". So, we defined a dummy variable using "factor" function with 1 indicating a nonsmoker and 0 indicating a smoker for that smoking variable. Then by using "lm" function we fitted the linear model & then estimated the regression coefficients. Following values are obtained as the values of parameters,

lm (formula = Risk ~ Age + Pressure + Smoker)
Coefficients:

(Intercept)	Age	Pressure	Smoker 1
-83.0196	1.0767	0.2518	-8.7399

According to this, $\widehat{Risk}$**= -83.0196+1.0767(Age) +0.2518(Pressure)-8.7399(Smoker)** equation could be taken as the regression equation.

- *When age increases by one year the probability of risk is expected to increase by 0.01077, when all other independent variables are held constant.*
- *When pressure increases by one unit the probability of risk is expected to increase by 0.002518, when all other independent variables are held constant.*
- *The expected risk for a person who is a smoker is 0.087399 more than a person who is not a smoker, when all other independent variables are held constant.*

Step 02: Significance testing in multiple linear regressions.

In multiple regression we use the F test (test for overall significance) in ANOVA to test where there is a regression or not. Hence, we used "anova" function in R to test the significance

F-statistic: 36.82 on 3 and 16 DF, p-value: 2.064e-07
p-value: 2.064e-07 < 0.05 (α)

Full model is significant

And we also wanted to test the individual significance (Using t test) of independent variables by running the "summary" function. Output is the coefficient table which is used to find individual significance of the model.

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	-83.019631	6.74115	-4.959	0.000142***< 0.05
Age	1.07674	0.16596	6.488	7.49e-06 ***< 0.05
Pressure	0.25181	0.04523	5.568	4.24e-05 ***< 0.05
Smoker	-8.73987	3.00082	-2.912	0.010174*< 0.05

All coefficients are significant

After realizing that both overall & individual tests were significant for further clarification, we run the function "stepAIC" under "mass library" to check for the variable selection methods.

Step 03: Goodness of fit testing

By using **Adjusted R-squared** value under "summary" function we tried to find out how much of variability on our response variable (probability of risk) is can be explained by the regression relationship. In MLR, it is better to go with adjusted R^2 instead of using usual R^2.

Multiple R-squared: 0.8735, Adjusted R-squared: 0.8498

Step 04: Residual Analysis and other Diagnostics.

There are few assumptions should be satisfied in MLR model. For regression models, the errors were assumed to be independent normal random variables with mean 0 & constant variance σ^2. Standardized residuals Vs fitted values plot can be used to check the linearity, independence & constant variance assumptions. Also, by using Normal probability plot we can check the normality assumptions. Using "plot" function in R, we constructed those graphs & checked the validity of those assumptions.

Another assumption we assumed in regression model is that there are no any unusual observations. They are outliers & influential observations. If standardized residuals < |2|, then we can say that there are no any outliers. By using "scale" function calculated the standardized residuals. Influential point is a point that pulls the regression line towards it. We can measure them using "**cooks.distance**" function in R. We use the cut off vale "4/(n-p)", for influential points the cook's statistic would exceed this value.

- *From standardized residual values in table 1 we can see that there is only one value past 2 (17^{th} observation). That value can be considered as an outlier.*
- *All the cooks distance values are smaller than the cut off value of 4/(n-p) = 0.2353, according to the table 2. So, there are no influential observations.*

In MLR, we should examine the multicollinearity. That means sometimes among independent variables high correlation could be exist. If multicollinearity exists, we have to remove particular independent variable and re-fit the model with remaining variables. Using "**vif**" function under "car library" we can find these VIF values. If VIF value is closer to 1, it indicates multicollinearity is not present in the fitted model.

Age	Pressure	Smoker
1.46035	1.24862	1.35869

No Multicollinearity

This is the methodology we followed to achieve our goal.

Conclusions

First and foremost, we fitted the model testing of all three predictor variables, “Age”,” Pressure” and “Smoker” in order to obtain the initial model. And the initial model was,

$$\widehat{Risk} = -83.0196 + 1.0767(Age) + 0.2518(Pressure) - 8.7399(Smoker).$$

Subsequently we performed an F test to determine the model’s overall significance which demonstrated that the entire original model was significant and provided a superior fit to the data. Since the initial model was overall significant then we checked for individual significance of predictors. And it was evident that all three predictor variables were significant at 5% level of significance and in consequence all predictors have a significant impact on the response variable “Risk”. With the intention of improving the model, we followed variable selection procedures such as forward selection, backward elimination and stepwise selection. Consequently, all three methods verified that the initial model is the best fit. As claimed by the R^2_a value, adjusting for the number of predictor factors, 84.98% of variability of the response variable is explained via the regression fit. Thus it was a good fit. When investigating the unusual observations, a single outlier was detected which was slightly deviated from the rest of the data points. None of the influential observations were exposed when the cook’s distance values were taken into account. Hence to sum up, the data set hasn’t been influenced by any unusual observations that could make a considerable effect on the model. Furthermore, the VIF values associated with the predictors revealed that multicollinearity doesn’t exist amongst predictors which contribute for an improved model. Finally, we checked the validity of the fit by plotting appropriate graphs. In standardized residuals Vs fitted values plot (Graph 01), the data points were randomly scattered around zero within a horizontal band except the single outlier detected. Moreover, the normal probability plot (Graph 02) was approximately linear. There we could see that most of points were plotted near to the straight line. In conclusion, the assumptions of the regression haven’t been violated. Ultimately by taking into consideration of all the facts discovered through the modeling we can conclude that the initial model is the best fitted line for the data set “Stroke”. Therefore, the best fitted line is,

$$\widehat{Risk} = -83.0196 + 1.0767(Age) + 0.2518(Pressure) - 8.7399(Smoker).$$

According to the data set “Stroke”, the patient’s age, level of the blood pressure and whether the patient smokes or not have a substantial impact on the patient’s risk of having a stroke following 10-year period. Furthermore, this model may also be used to predict the likelihood of a person having a stroke during the next 10-year period when the corresponding values for the predictors are known.

Appendix 01

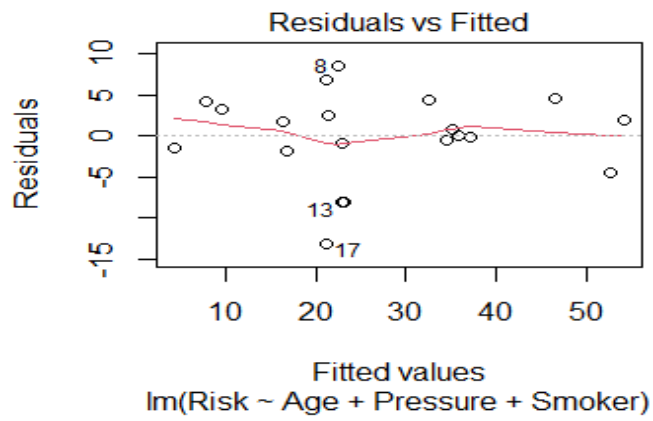

Figure 1: Graph1

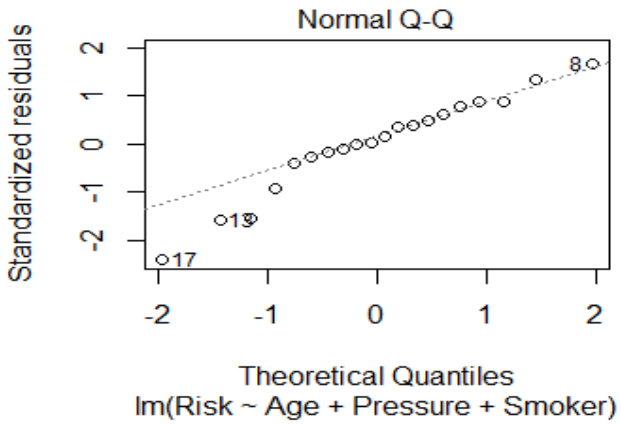

Figure 2: Graph 2

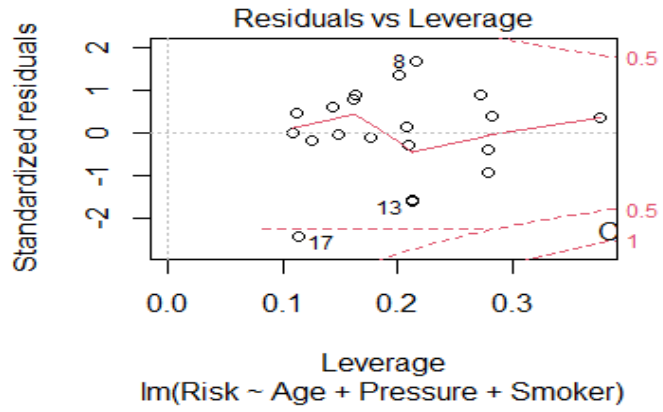

Figure 3: Graph 3

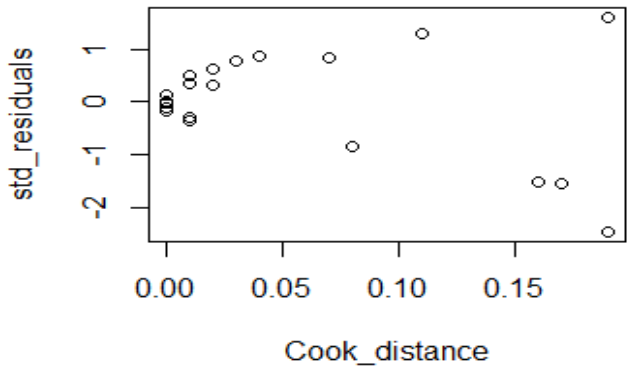

Figure 4: Graph 4

Index	Std_residuals	Std_residuals<|2|
01	0.78	TRUE
02	0.49	TRUE
03	0.62	TRUE
04	0.35	TRUE
05	1.30	TRUE
06	0.87	TRUE
07	0.32	TRUE
08	1.62	TRUE
09	-0.02	TRUE
10	-0.36	TRUE
11	-0.18	TRUE
12	0.02	TRUE
13	-1.54	TRUE
14	-0.86	TRUE
15	-1.51	TRUE
16	0.14	TRUE
17	-2.48	FALSE
18	-0.11	TRUE
19	-0.28	TRUE
20	0.83	TRUE

Table 2: Standardized Residuals

Index	Cook_distance	Cook_distance<0.24
01	0.03	TRUE
02	0.01	TRUE
03	0.02	TRUE
04	0.01	TRUE
05	0.11	TRUE
06	0.04	TRUE
07	0.02	TRUE
08	0.19	TRUE
09	0.00	TRUE
10	0.01	TRUE
11	0.00	TRUE
12	0.00	TRUE
13	0.17	TRUE
14	0.08	TRUE
15	0.16	TRUE
16	0.00	TRUE
17	0.19	TRUE
18	0.00	TRUE
19	0.01	TRUE
20	0.07	TRUE

Table 1: Cooks distance values

Appendix 02

```
X4_Stroke
attach(X4_Stroke)
fit = lm(Risk~Age+Pressure+ factor(Smoker,c("Yes","No"),labels = c(0,1)))
fit
anova(fit)
summary(fit)
library(MASS)
stepAIC(fit,direction = "forward")
stepAIC(fit,direction = "backward")
stepAIC(fit,direction = "both")
plot(fit)
std_residuals = round(scale(fit$residuals),2)
std_residuals
outliers = c((std_residuals)<2 & (std_residuals)>-2)
outliers
Cook_statistics = round(4/(20-3),2)
Cook_statistics
Cook_distance = round(cooks.distance(fit),2)
Cook_distance
influential = c((cooks.distance(fit))<Cook_statistics)
influential
plot(Cook_distance,std_residuals)
library(car)
vif(fit)
```