Impact of Social, Work and Physical Condition on Stroke

Saurav Banerjee

3/3/2021

Research Scenario

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Apart from the physical condition of the body, there could be other factors, which can impact the chances of stroke. Some of these factors are -
a. Environment in which the person lives - Rural or Urban,
b. Work life - Govt. Work or other work categories, c. Physical metrics - BMI and Avg. Glucose levels

In this project, we are investigating the impacts of social environment, work life and physical metrics (BMI and avg. Glucose Level) on the likelyhood of stokes in human beings.

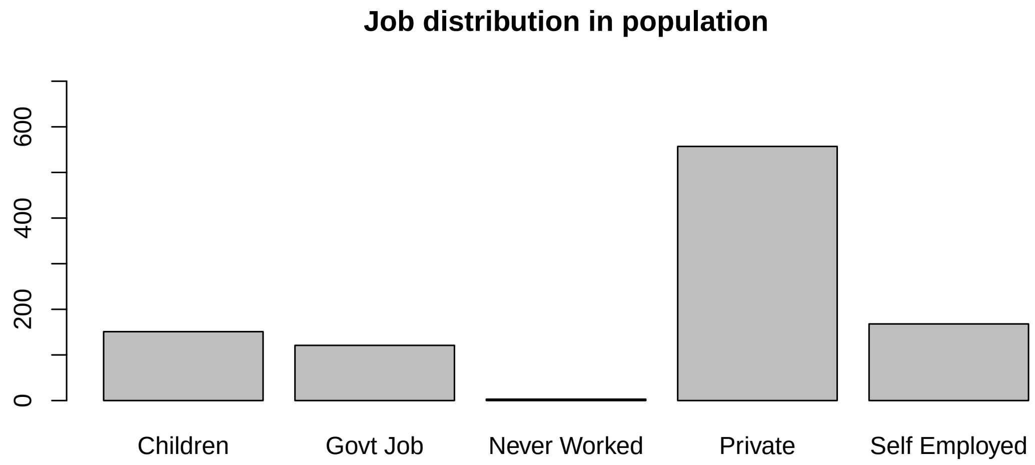
Description of Stroke Data set

The dataset used for this project is taken from kaggle.com (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset). The data set contains 12 variables, out of which we have selected the 4 variables -

a. Residence_type - The environment where the person lives. This variable has 2 categories - Rural and Urban.

Residence	Count
Rural	492
Urban	508

b. Work_type - The work life that the person is involved in. The data set contains 5 categories - "Children", "Govt. job", "Never worked", "Private" and "Self employed". Below is the ditribution of jobs in the sample



-

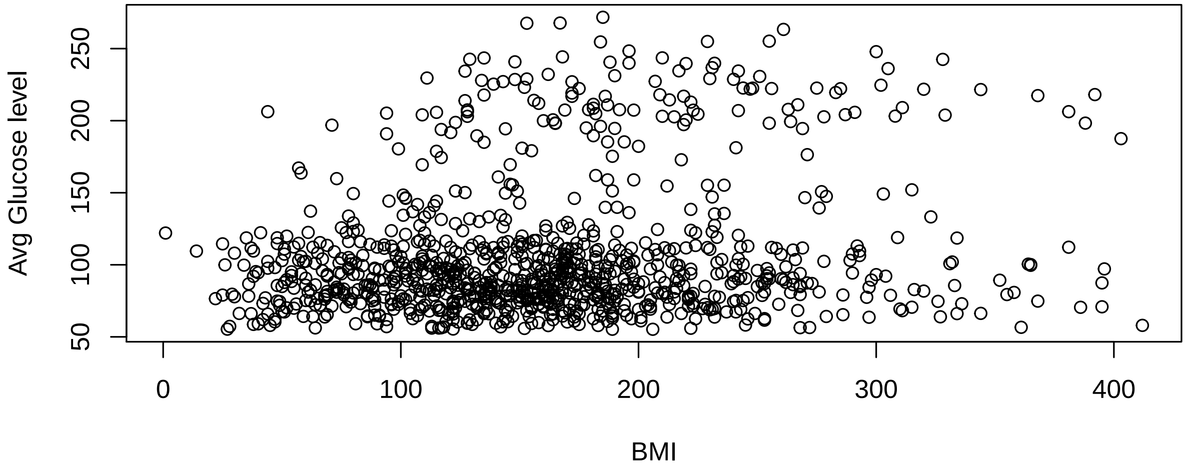
For the simplicity of this project, we have reduced this variable to have 2 categories - Private Job, Non private job.

Residence	Count
Rural	492
Urban	508

c. BMI - The body mass index. This is a continuous, numerical variable.

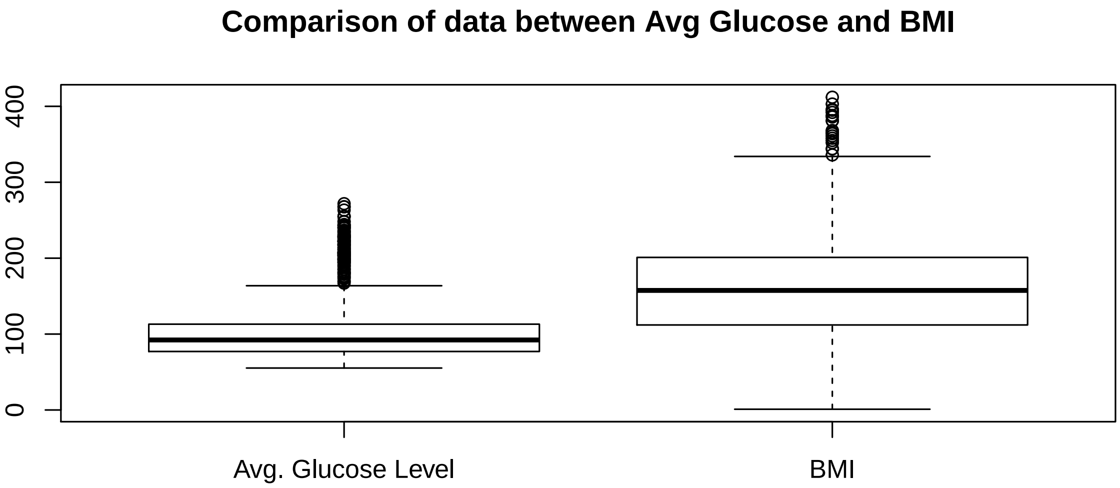
d. Avg. Glucose Level - This is the avg. glucose level in blood. This is a continious numerical variable.

For BMI and Avg. Glucose Level, below is the scatter plot showing their correlation -



There's a low correlation between BMI and Avg. Glucose Level of 0.1914781.

Also, below is the distribution of the data across the median.



The data for Glucose levels and BMI have outliers, as we observe from the box plots above.

Data Cleanup

Data for BMI were having rows with value as 'N/A'. Considering that BMI is a continious numeric variable, as part of data cleanup, data with BMI = 'N/A', were removed from the data set, before taking the sample for below statistical analysis.

Multiple Logistic Regression

To determine the impact of these 4 variables on Stroke, we will model the logistic regression model, and determine if these variables are significant, in determining Stroke.

We would perform a global test, to delermine, if there exists atleast one of these 4 variables, which affects the Stroke outcome significantly.

For this test, we assume the level of significance $\alpha = 0.05$.

From the global test, we find that p-Value = $7.6634747 \times 10^{-5} < 0.05$. We reject the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, and conclude that there is atleast one β_i != 0, where -

β_1 - Regression cofficent for BMI
β_2 - Regression cofficent for Avg Glucose Level
β_3 - Regression cofficent for Work Type
β_4 - Regression cofficent for Residence

As the global test rejects the null hypothesis, we perform tests for each of the 4 variables to check if the regression cofficients of that variable is not equal to 0.

Results

Below are the regression coefficients and p-Values, for this logistic regression model -

	Estimate	Std. Error	z value	Pr(>|z|)
bmi_n	-0.0000193	0.0019984	-0.0096679	0.9922863
avg_glucose_level_numeric	0.0109354	0.0023725	4.6092680	0.0000040
work_type1	0.1905782	0.2872872	0.6633719	0.5070924
Residence_type1	0.1697753	0.2768504	0.6132387	0.5397185

BMI:

The hypothesis tested here are - H_0: $\beta_1 = 0$, or $OR_{BMI} = 1$, (no association between BMI am stroke, after controlling for glucose level, work type and residence)

H_1: β_1 != 0, or OR_{BMI} != 1, (There is association between BMI and stroke, after controlling for glucose level, work type and residence)

We do not have significant evidence at $\alpha = 0.05$ level, that β_1 != 0. (p-Value = sum$coefficients[2,'Pr(>|z|)'])

Avg. Glucose Level:

The hypothesis tested here are - H_0: $\beta_2 = 0$, or $OR_{GlucoseLvl} = 1$, (no association between Glucose Level and stroke, after controlling for BMI, work type and residence)

H_1: β_2 != 0, or $OR_{GlucoseLvl}$!= 1, (There is association between Glucose Level and stroke, after controlling for BMI, work type and residence)

We do not have significant evidence at $\alpha = 0.05$ level, that β_2 != 0. (p-Value = sum$coefficients[3,'Pr(>|z|)'])

Work Type:

The hypothesis tested here are - H_0: $\beta_3 = 0$, or $OR_{WorkType} = 1$, (no association between Work Type and stroke, after controlling for BMI, Avg. Glucose Level and residence)

H_1: β_3 != 0, or $OR_{WorkType}$!= 1, (There is association between Work Type and stroke, after controlling for BMI, Avg. Glucose Level and residence)

We do not have significant evidence at $\alpha = 0.05$ level, that β_3 != 0. (p-Value = sum$coefficients[4,'Pr(>|z|)'])

Residence Type:

The hypothesis tested here are - H_0: $\beta_4 = 0$, or $OR_{ResidenceType} = 1$, (no association between Residence Type and stroke, after controlling for BMI, Avg. Glucose Level and Work Type)

H_1: β_4 != 0, or $OR_{ResidenceType}$!= 1, (There is association between Residence Type and stroke, after controlling for BMI, Avg. Glucose Level and Work Type)

We do not have significant evidence at $\alpha = 0.05$ level, that β_4 != 0. (p-Value = sum$coefficients[4,'Pr(>|z|)'])

Odd's ratio and Confidence Interval

Below is the Odd's Ratio and Confidence Interval for each of the 4 variables -

	OR	2.5 %	97.5 %
bmi_n	0.9999807	0.9960716	1.003905
avg_glucose_level_numeric	1.0109954	1.0063052	1.015707
work_type1	1.2099490	0.6890136	2.124743
Residence_type1	1.1850386	0.6887745	2.038863

BMI:
The odds ratio of chances of Stroke is 0.9999807. For every 1 unit increse in BMI (body mass index), the odds ratio of chances of stroke is 0.9999807, after controlling for avg. glucose level, work type and residence type.

We are 95% confident that the true odds ratio is between 0.9960716 and 1.0039051 after adjusting for avg. glucose level, work type and residence type.

Avg. Glucose Level:
The odds ratio of chances of Stroke is 1.0109954. For every 1 unit increse in Avg. Glucose Level, the odds ratio of chances of stroke is 1.0109954, after controlling for BMI, work type and residence type.

We are 95% confident that the true odds ratio is between 1.0063052 and 1.0157074 after adjusting for BMI, work type and residence type.

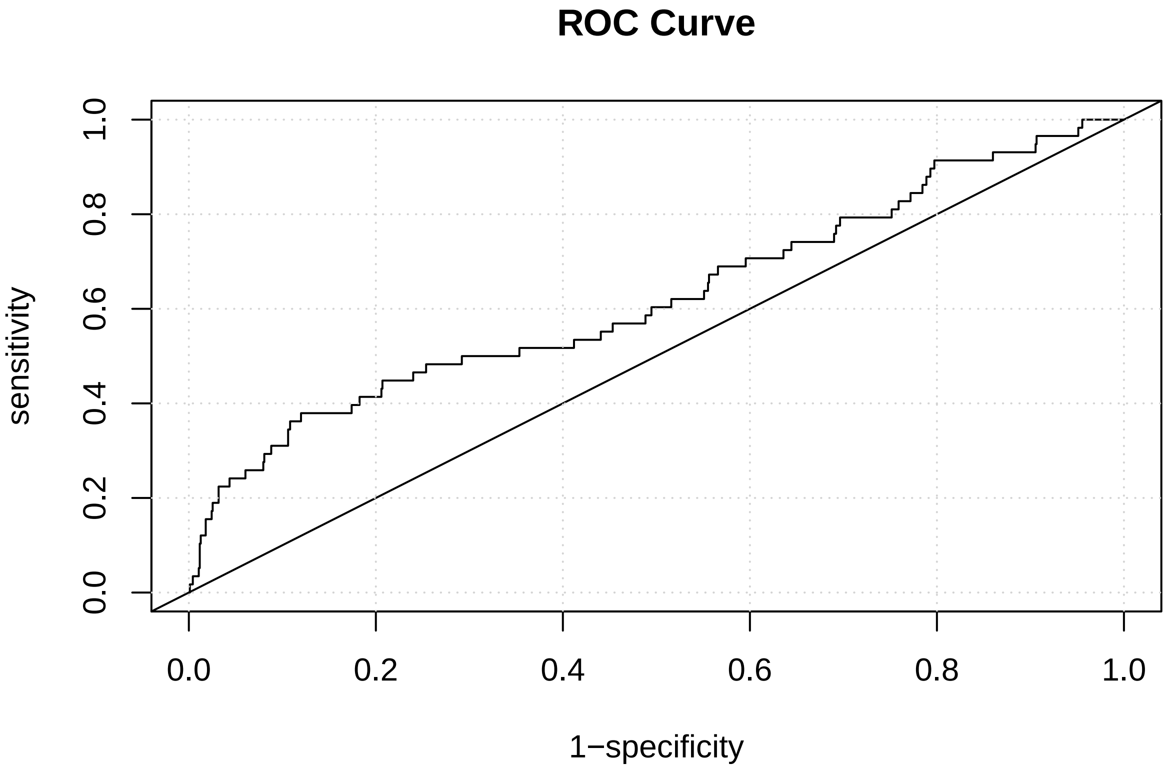
Work Type:
The odds ratio of chances of Stroke is 1.209949. The odds ratio of chances of stroke for people in Private Jobs is 1.209949, as compared to people not in Private jobs, after controlling for BMI, avg. Glucose Level and residence type.

We are 95% confident that the true odds ratio is between 0.6890136 and 2.1247429 after adjusting for BMI, work type and residence type.

Residence Type:
The odds ratio of chances of Stroke is 1.1850386. The odds ratio of chances of stroke for people in Urban residences is 1.1850386, as compared to people in Rural areas, after controlling for BMI, work type and work type.

We are 95% confident that the true odds ratio is between 0.6887745 and 2.0388626 after adjusting for BMI, work type and work type.

ROC Curve



Using the model based on the association between bmi, glucose levels, residence and work type, we see the model predictions to be fairly accurate. The area under the curve was over 50% (roc = 62.2%)

Conclusion

In this exercise, we set out to find the association between Stroke and BMI, Avg. Glucose Level, Residence and Work Type. Based on the Logistic Regression model, and the statistical tests performed, we have the following conclusions -

a. We were *not able to find evidence* of an association between BMI and stroke after adjusting for glucose level, residence and work type.

b. We were **able to find evidence** of an association between Avg. Glucose Level and Stroke, after adjusting for BMI, Residence and work type. Hence, we can say that Avg. Glucose Level impacts chances of stroke.

c. We were *not able to find evidence* of an association between Work Type as Private or Non Private, and stroke after adjusting for glucose level, residence and BMI. This implies that we do not have evidence, that the type of job impact the chances of Stroke.

d. We were *not able to find evidence* of an association between Residence Type as Urban or Rural and stroke after adjusting for glucose level, BMI and work type. This implies that we do not have evidence, that the environment has impact the chances of Stroke.

e. The chances of prediction of stroke from - BMI, Avg. Glucose Level, Residence Type and Work Type is 62.2%

Limitation of this model

The Regression model evaluated based on the data has an accuracy of over 60%. Though this is better than flipping a coin, the accuracy of prediction from this model is not reliable.

Appendix: R Code

```
knitr::opts_chunk$set(echo = FALSE)
# Won’t show the code in the document
library(knitr) # Need knitr for kable and the appendix
library(aod)
library(pROC)

setwd(paste(’/Users/sauravbanerjee/Documents/MSADA/’,
      ’MET CS 555 - Data Analysis and Visualization/Final Project/’, sep = ’’))
# Read Data
data <- read.csv(’data.csv’)
# Data Clueanup - Remove records having N/A
data <- data[data$bmi != ’N/A’,]
# Sample 1000 records from dataset
data <- data[sample(nrow(data), 1000),]
write.csv(data, file = ’sample.csv’, sep = ’,’)
# Residence Frequency
kable(table(data$Residence_type), col.names = c(’Residence’, ’Count’))
par(mfrow=c(1,1))

# Custom lables for display
data$work_type_desc <- ifelse(
  data$work_type == ’children’,
  ’Children’,
  ifelse(data$work_type == ’Govt_job’, ’Govt Job’,
       ifelse(data$work_type == ’Self-employed’,’Self Employed’,
            ifelse(data$work_type == ’Never_worked’, ’Never Worked’, ’Private’))))


# Bar plot
barplot(table(data$work_type_desc)
        , main = ’Job distribution in population’
        , ylim = c(0, 700))

# Simplify variable to Private/ non private
data$work_type_desc <-  ifelse(data$work_type_desc != ’Private’, ’Non-Private’, ’Private’)
# Frequency
kable(table(data$Residence_type), col.names = c(’Residence’, ’Count’))
par(mfrow=c(1,1))
data$bmi_n <- as.numeric(data$bmi)
data$avg_glucose_level_numeric <- as.numeric(data$avg_glucose_level)
# Scatter plot
plot(data$avg_glucose_level_numeric~data$bmi_n,
     xlab = ’BMI’, ylab = ’Avg Glucose level’)

par(mfrow=c(1,1))
# Scatter plot
boxplot(data$avg_glucose_level_numeric,
        data$bmi_n, names = c(’Avg. Glucose Level’, ’BMI’),
        main = ’Comparison of data between Avg Glucose and BMI’)

lrDataModel <- data[, c(’bmi_n’, ’avg_glucose_level_numeric’, ’stroke’)]
lrDataModel$work_type <- as.factor(ifelse(
  data$work_type_desc == 'Private',1, 0))
lrDataModel$Residence_type <- as.factor(ifelse(
  data$Residence_type == 'Urban',1, 0))

# Logistic Regression
m <- glm(
  data = lrDataModel,
  stroke ~ bmi_n + avg_glucose_level_numeric + work_type + Residence_type, family = binomial)

# Global Test
globalTest <- wald.test(b=coef(m), Sigma = vcov(m), Terms = 2:5)

# p-values
sum <- summary(m)
# Odd's Ratio and CI
op <- exp(cbind(OR = coef(m), confint.default(m)))
kable(sum$coefficients[-1,])
tab <- exp(cbind(OR = coef(m), confint.default(m)))[-1,]
kable(tab)
# ROC
lrDataModel$predicted <- predict(m, type= c("response"))
g <- roc(lrDataModel$stroke ~ lrDataModel$predicted)
plot(1-g$specificities, g$sensitivities,
     type="l", xlab = "1-specificity",
     ylab = "sensitivity", main = "ROC Curve")
abline(a=0, b = 1)
grid()
```