AN ANALYSIS OF STROKE AND AFFILIATED FACTORS

A FINAL PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE
STAT 364 – LINEAR MODELS II
DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

Miray Çınar 2290609

Ecenaz Tunç 2290971

Levent Sarı 2290930

İpek Aydın 2290526

Deniz Bezer 2290559

June 2021

ABSTRACT

Health has always been a concern of people, and there are many factors that affect our health. Some of these factors are attributes that people can control, such as smoking or exercising; others are disabilities or life status are ones that are not controllable, but still affect human health. This project focuses on one specific health concern of humans, which is having a stroke. Stroke is ranked as the second leading cause of death worldwide with an annual mortality rate of about 5.5 million (Donkor, 2018). This study investigates how the possibility of having a stroke is explained by external, internal, and social attributes. Firstly, the physical conditions of the individuals such as BMI and age were tested to find a relationship. As a result, it was found that the median BMIs of the people who have had strokes are higher. Moreover, it was also found that their ages, independent from the first result, are also higher. Secondly, external factors such as work type, residence type, smoking and marital status were examined. A weak positive relationship between work type and having a stroke was found. It can be deduced that self-employed people are more likely to have a stroke. Again, a weak negative relationship was found between smoking status and stroke. However, it is negligible; therefore, it can be said that smoking status has no relationship with having a stroke. Another factor that was examined was marriage status, a strong relationship was found. It can be concluded that married people are 4.5 times more likely to have a stroke. Then the past and current diseases and health conditions such as heart disease, hypertension, and glucose level were examined to find a relationship. It was found that people with heart diseases are 5.24 times more likely to get a stroke and people with hypertension are 4.43 times more likely to get a stroke. There was a positive relationship between stroke and glucose level, it may be confirmed that people who had a stroke and who did not have a stroke do not have the same median glucose levels. To sum up, it can be concluded that present or past diseases and / or disorders, such as heart disease, hypertension, sub-optimal glucose level increase the likelihood of having a stroke. Finally, selected covariates were gathered to conduct a model to predict stroke status. The conducted model has a binomial response for stroke and has the covariates age, hypertension, heart disease status, and log of average glucose level. According to the model; when all the covariates were held constant; the odds of having a stroke increases by 39.4 % with each additional age of the patients. The odds of having a stroke increase by 63.1 % if the patient has hypertension. Similarly, the odds of having a stroke increase by 55.1 % if the patient has heart disease. Lastly, the odds of having a stroke increases by 37 % with each additional average level of glucose of the patients.

1. Introduction

According to the World Health Organization, stroke is the second most common cause of death worldwide. It is responsible for approximately 11% of the deaths that occur.
In this study, the factors that have a relationship with the patient's probability of having a stroke were examined. Various physical, health, and environmental factors have been used to estimate the probability of a patient having a stroke.
The dataset includes records from both ischemic and hemorrhagic strokes. According to American Stroke Association, an ischemic stroke happens when a blood vessel that carries oxygen and nutrients to the brain is blocked by a clot. A hemorrhagic stroke occurs when a blood vessel that carries oxygen and nutrients to the brain bursts.
Since it is health data, it is kept confidential from where and from whom the data was collected.

1.1. Data Description

The data set that was used in this study is collected to record the attributes of individuals who have already had strokes and those who have not. The data set consists of 4908 observations with 11 attributes of which 8 of them are categorical and 3 numerical which contains records from both ischemic and hemorrhagic strokes. The names of the attributes are; “gender”, “age”, “hypertension”, “heart_disease”, “ever_married”, “work_type”, ”Residence_type”, “bmi”, “avg_glucose_level”, “smoking_status”, ”stroke”.

The description for individual variables in the study are:

1) “gender”: gender of the individuals, categorical: female & male,
2) “age”: age of the individuals, numerical,
3) “hypertension”: hypertension status of individuals, categorical: 0 for no hypertension & 1 for having hypertension,
4) “heart_disease”: heart disease status of individuals, categorical: 0 for no heart disease & 1 for having heart disease,
5) “ever_married”: marriage status of individuals, categorical: No for never married & Yes for married,
6) “work_type”: work type of individuals, categorical: children for not working age individuals, Govt_job for government officers, Never_worked for never worked

individuals, Private for individuals who work for private companies, Self-employed for self-employed individuals,

7) "Residence_type": type of residence, categorical: Rural & Urban residence types
8) "avg_glucose_level": average glucose level of individuals, numerical
9) "bmi": body mass index of the individuals, numerical,
10) "smoking_status": smoking status of the individuals, categorical: formerly smoked, never smoked, smokes, unknown,
11) "stroke": stroke status of the individuals, categorical: 0 for not having stroke, 1 for having stroke

1.2. Research Questions

In this research, factors that cause a stroke are investigated. 4 main research questions were examined to figure out the relationships between having a stroke and depending factors such as physical conditions, external conditions, diseases, and disorders. The four main research questions are;

1. Do people who have had strokes have worse physical conditions, such as suboptimal BMI or older age?
2. Do external conditions, such as work type, residence type, smoking, and marriage status of people change their likelihood to have a stroke?
3. Do present or past diseases / disorders, such as heart disease, hypertension, suboptimal glucose levels increase the likelihood of having a stroke?
4. Can an accurate prediction be made for the possibility of having a stroke by using any or all of the independent variables?

These research questions were examined, to find which factors of his study related to having a stroke.

1.3.Aim of the Study

According to the World Health Organization, stroke speedily presents medical signs of global disturbance of cerebral function, it has symptoms that can last for 24 hours or longer, and can also cause death (Warlow, 1998). Strokes are also the second most common cause of death in the world (Donkor, 2018). Moreover, according to WHO Eastern Mediterranean Office, 15 million people in the world experience a stroke every year, a third of which are met

with death and the other 1/3 are left disabled, additionally, strokes are common in people who are older than 40. These facts imply stroke as a fatal condition, and there are multiple factors that cause strokes. Hence, in this study, the intention is to examine the factors that have a relationship with having a stroke and by investigating these factors, raising awareness for the medical condition.

2. Methodology

In this study, in order to analyze the four research questions proposed as the main interest, various statistical analyses were applied to nine different sub-questions and plots were drawn where relevant. To apply these analyses on, the dataset "Stroke Prediction" that tries to explain the stroke events with eleven other categories was used. The main tool of research was chosen as the R language. At the beginning, each category in the data was examined closely and assigned the correct data type. There were no problematic subjects found and further cleaning of the data was considered unnecessary.

To conduct the analysis for testing the first research question that examines the association between physical features and past stroke status, two different sub-questions were assigned to inspect the age and BMI of the subjects separately. For each of these sub-questions, the data was first divided into two parts separated by their past stroke status. Subsequently, their boxplots were drawn to find a significant difference between the means of the groups. Then, for further confirmation, two-sample Wilcoxon Rank Sum Test was applied to both questions. This method of testing was chosen as the data in both cases were found to be non-normally distributed via Shapiro-Wilk Normality Test and the box-cox normality transformation failed. From the Wilcoxon Test, significant relationships were found both for BMI and past stroke experience, and also for age and past stroke experience. By applying polyserial correlation for each sub-question, an extremely small positive correlation and a mediocre positive correlation were found for BMI and age respectively, showing that people who have had past strokes generally have higher age and BMI.

In the analysis of the second research question which was conducted to look for a significant relationship between strokes and external conditions, four different hypotheses were conducted to closely study the link between strokes and each external condition, namely, work type, residence type, smoking status and marriage status one by one. As all of the aforementioned

variables were categorical with varying levels, appropriate contingency tables were used. Upon closer examination of the first hypothesis, as work type has five different categories, a Chi-Squared Test was first tried on the table. However, upon findings of an expected value lower than 5 in the table, the application of Fisher's Exact Test was seen as more appropriate. The result of the test has provided a significant relationship between work type and strokes, and the following polychoric correlation classified a weak positive relationship. For the examination of the affiliation between residence type and strokes, as both variables have only two categories, an odds ratio calculation that resulted in no significant relation was done. The third sub-question that was conducted for smoking status was inspected following a similar procedure with the first sub-question examining the work type. However, as there were no results smaller than 5 on the expected values table for the Chi-Squared Test, its result showing a significant relationship was considered acceptable, and a weak negative polychoric correlation was found. For the last sub-question a method similar to the second one was followed and from the odds ratio, a very strong relationship between marriage status and strokes was found. After the statistical analysis for each sub-question, a plot to strengthen the argument was also drawn and significant relationships between all external factors except residence type and stroke occurrences were found.

For the third research question which aims to inspect the relationship between past diseases and occurrences of strokes, three different sub-questions were asked to independently look for conclusions in heart diseases, hypertension, and suboptimal glucose levels. As both the stroke status and past heart diseases variables are categorical with two categories, a very high odds ratio was calculated that provided a significant relationship between these two variables. A similar method was applied in the second sub-question which aimed to find the relationship between hypertension and stroke occurrences to find an odds ratio that also presented a significant relationship between the variables. Lastly, to inspect the relationship between strokes and the average glucose level, a two-sample Wilcoxon Rank Sum Test was applied as the stroke variable failed both the Shapiro-Wilk Normality Test and the Box-Cox normality transformation. Upon conduction of the test, a difference in medians was observed, providing sufficient evidence to conclude a significant relationship between the variables. In the end, all past diseases were found to be related to stroke occurrences.

The aim of the fourth research question was to build a model that could explain the outcome of a stroke by any or all the other variables given. Since the response variable was chosen as the binary stroke status, a logistic regression model was seen as appropriate. First, the data was

separated for the train-test method with a %90 and %10 observation rate, respectively. Subsequently, a logistic regression model containing all variables was fit into the train data. Three different methods of AIC selection were applied in the order of forward selection, backward elimination, and stepwise regression, resulting in a model containing four independent variables against the response of stroke. Afterwards, the continuous variables of the model were plotted against the logit results and the variable corresponding to the average glucose level was found to be a bad fit. As the graph showed exponential growth for that specific variable, a method of log transformation was then tried on the variable, which ultimately ended in failure. The best possible threshold was chosen with the ROC Curve method. The necessary performance tests were then applied to the model with the original four variables via the usage of a confusion matrix, which provided a good accuracy, satisfying sensitivity and specificity, a strong negative prediction value however an unacceptable positive prediction value. In the end, it was decided that the train-test method was not effective enough to be the main application. A model was then fit into the data as a whole with all variables and then another one was fit including the second-degree interactions of the variables alongside them. Looking at the VIF values, it was confirmed that adding the interaction terms resulted in multicollinearity, and the model including interaction terms was removed. To the full model applied to the full data, the three AIC procedures were again applied, resulting in the same four independent variables being used in the final model. When the logit versus variables graph was again drawn, the exponentially increasing average glucose level was once more transformed using a log transformation, resulting in an unimproved version of the final model used in the explanation of the research question, which was then crossed out. In the end, the model provided sufficient accuracy, satisfactory sensitivity and specificity values, an excellent negative prediction rate, however, an unacceptable positive prediction rate. The extremely low positive prediction rate and detection rate were considered as the result of the uneven distribution of responses and very low positive observations in the response variable. In the end, the model was considered acceptable, while keeping in mind that further enhancements can be made by training with a more evenly distributed and larger dataset to obtain a better positive prediction rate in further studies.

3. Results and Findings

Research Question 1: Do people who have had strokes have worse physical conditions, such as suboptimal BMI or older age?

BMI

Firstly, the normality of BMI of the individuals are checked with the Shapiro Wilk test;

H_0: They are coming from an identical population.

H_1: They are not coming from an identical population.

Since p-value < $2.2*10^{-16}$ is smaller than $\alpha = 0.05$, null hypothesis is rejected, BMI is not normally distributed.

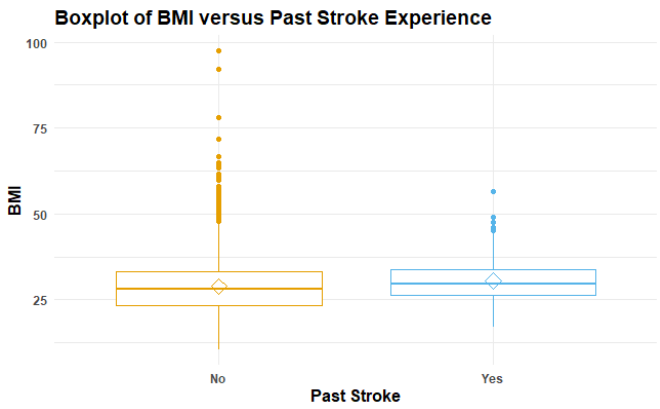
To obtain normality, Box-Cox Transformation is considered. Since λ is close to 0, Y'=log(Y) is applied on BMI, however; normality could not be obtained when Shapiro-Wilk test is applied again. Hence, Wilcoxon Rank Sum Test with Continuity Correction is applied between BMIs of two subgroups, as people who have had strokes, and who have not had strokes.

H_0: They have equal medians.

H_1: They do not have equal medians.

Since p-value = 0.0001042, is smaller than $\alpha = 0.05$, null hypothesis is rejected, medians differ, hence, it can be said that there is a relationship between BMI and having a stroke.

Figure 1: Boxplot of BMI vs. Past Stroke Experience



From Figure 1, it can be seen that people who have had strokes seem to have relatively higher BMI compared to people who have not had stroke before. To see the relationship more clearly, polyserial correlation is examined, which is found as 0.09354, hence, it can be concluded that there is a weak positive correlation between having a stroke and BMIs of the individuals.

Age

For examining the age, firstly, Shapiro Wilk test is applied, where hypotheses are:

H_0: They are coming from an identical population.

H_1: They are not coming from an identical population.

Since p-value < $2.2*10^{-16}$ is smaller than $\alpha = 0.05$, null hypothesis is rejected, age is not normally distributed. To obtain normality, Box-Cox Transformation is considered. Since λ is close to 0.85, $Y'=Y^{0.85}$ is applied on age, however; normality could not be achieved when the normality test is applied again.

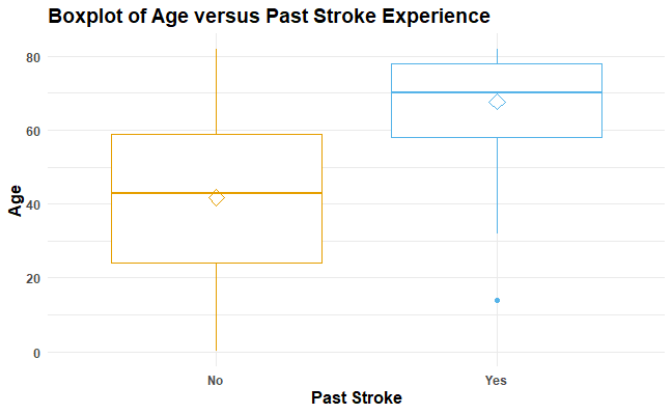
Hence, Wilcoxon Rank Sum Test with Continuity Correction is applied between ages of two subgroups, as people who have had strokes, and who have not had a stroke.

H_0: They have equal medians.

H_1: They do not have equal medians.

Since p-value < $2.2*10^{-16}$ is smaller than $\alpha = 0.05$, null hypothesis is rejected, medians differ, hence, it can be said that there is a relationship between age and having a stroke.

Figure 2: Boxplot of Age vs. Past Stroke Experience



From Figure 2, it can be seen that people who have had strokes seem to have higher age compared to people who have not had a stroke before. To see the relationship more clearly, polyserial correlation is examined, which is found as 0.6143.

Hence, it can be said that there is a positive correlation between age and having a stroke.

Research Question 2: Do external conditions, such as work type, residence type, smoking, and marriage status, of people change their likelihood to have a stroke?

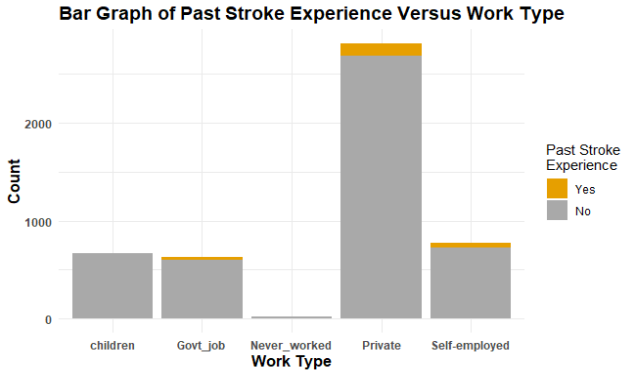
Work Type

	Children	Gov. Job	Never worked	Private	Self-employed
Had a stroke	670	602	22	2683	722
Haven't had a stroke	1	28	0	127	53

Table 1: Contingency Table of Work Status versus Past Strokes

First, a contingency table created with the table() function is set up to cross-classifies the number of rows that are in the categories specified by the two categorical variables. Hence, it can be said that there is a relationship between age and having strokes.

Figure 3: Bar Graph of Work Type vs. Past Stroke Experience



According to Figure 3, there seems to be a relationship between some covariates. A Chi-Square Test was created to examine it in more detail.

H_0: There is no relationship between the two categorical variables.

H_1: There is a relationship between the two categorical variables.

	Children	Gov. Job	Never worked	Private	Self-employed
Had a stroke	642.43	603.17	21.06	2690.4	741.99
Haven't had a stroke	28.58	26.83	0.94	119.66	33.01

Table 2: Expectations of Chi-Square Test Results

Looking at Table 2, where expected values of the Chi-Square test are examined, the existence of an expected value smaller than five was determined.

Hence, Fisher-test was applied. The null hypothesis is rejected because p-value = 0.0004998 < $\alpha = 0.05$. Hence, there is a significant relationship between the two categorical variables.

The polychoric correlation was measured to locate the direction of the relationship and a value of 0.2128 was reached. The output shows a weak positive correlation between stroke and work type. It can be deduced that people who are self-employed have more strokes.

Residence Type

	Rural	Urban
Had a stroke	2318	2381
Have not had a stroke	100	109

Table 3: Contingency Table of Residence Type versus Past Strokes

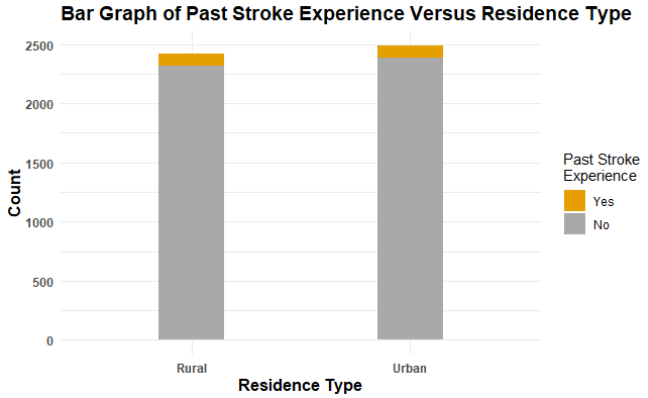
Again, analysis was started by setting up a contingency table, Table 3, for the residence type. The Chi-square Test was performed to measure whether there was a relationship between them.

	Rural	Urban
Had a stroke	2315.033	2383.967
Have not had a stroke	102.967	106.033

Table 4: Expectations of Chi-Square Test Results

Since, by looking at Table 4, no expected value of the Chi-squared test was less than five, the Chi-squared test could be used. Since the p-value of the Pearson's Chi-squared test is 0.7272, it can be said that there is no relation between stroke and residence type.

Figure 4: Bar Graph of Residence Type vs. Past Stroke Experience



The odds ratio of the test also almost 1 (1.061159), again, it can be said that there is no significant relation. The odds ratio Wald Test also includes 0, confirming that it is not significant. Looking at Figure 4, there is no significant relationship.

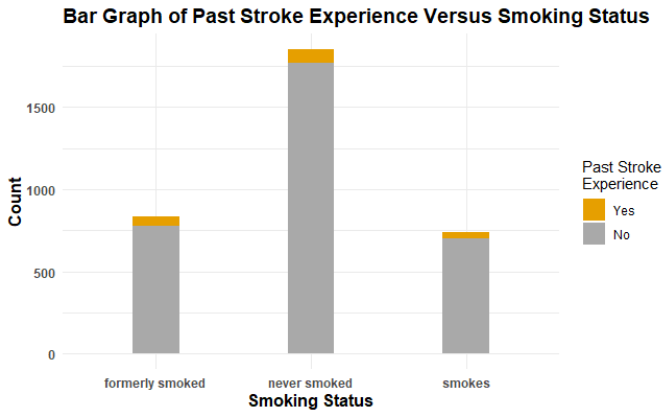
Smoking Status

Since there are too many unknowns in the smoking status variable, the analysis is started by extracting them. A contingency table as below is obtained.

	Formerly smoked	Never smoked	Smokes
Had a stroke	779	1768	698
Have not had a stroke	57	84	39

Table 5: Contingency Table of Smoking Status versus Past Strokes

Figure 5: Bar Graph of Smoking Status vs. Past Stroke Experience



Looking at Figure 5, it can be said that there is no relationship between smoking status and stroke, but a Chi-square test should be done to examine it in more detail.

	Formerly smoked	Never smoked	Smokes
Had a stroke	792.06423	1754.66861	698.26715
Have not had a stroke	43.93577	97.33139	38.73285

Table 6: Expectations of Chi-Square Test Results

Since none of the expected values of the Chi-Square Test was less than five, the Chi-Square Test could be used. According to Chi-Square Test, the p-value of 0.04906, is almost equal to 0.05. It can be inferred that there is almost no significant relationship. Polychoric correlation also confirms this with a -0.0566 correlation value.

Marital Status

	Not ever married	Ever married
Had a stroke	1681	03018
Have not had a stroke	23	186

Table 7: Contingency Table of Marital Status versus Past Strokes

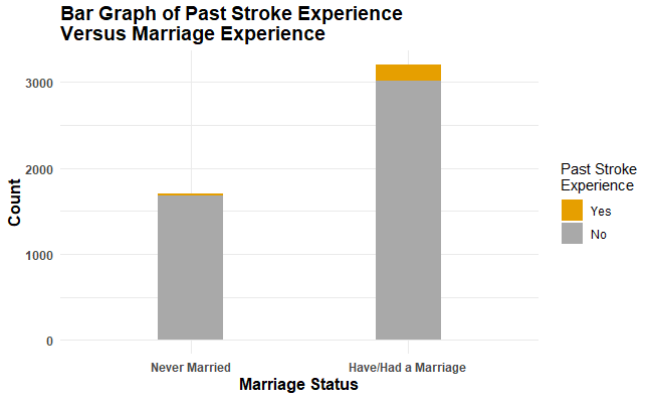
First, the odds ratio was found to determine which group was more prone to stroke.

H_0: The odds ratio is equal to 1.

H_1: The odds ratio is not equal to 1.

When the odds ratio measure test is also examined, it is seen that there is a significant relationship since it does not include 1.

Figure 6: Bar Graph of Residence Type vs. Past Stroke Experience



As a result of the calculations, the odds ratio was found to be 4.504365. That means married people are 4.5 times more likely to get a stroke. Figure 6 can also confirm this conclusion.

Overall, it is found that married people and people that work in self-employed jobs are more likely to have a stroke, while smoking status and type of residence are not related to the probability of stroke.

Research Question 3: Do present or past diseases and / or disorders, such as heart disease, hypertension, suboptimal glucose levels increase the likelihood of having a stroke?

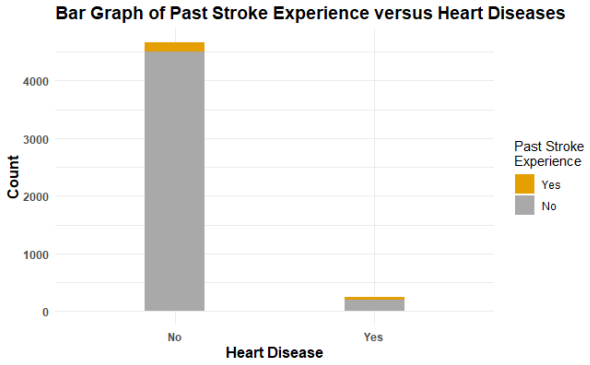
Heart Disease

First, a contingency table was constructed between heart disease history and having a stroke.

	Has / Had a Heart Disease	Did not Have a Heart Disease
Had a stroke	4496	203
Have not had a stroke	169	40

Table 8: Contingency Table of Marital Status versus Past Strokes

Figure 7: Bar Graph of Heart Disease vs. Past Stroke Experience



According to Figure 7, people with heart disease are more likely to have a stroke. The odds ratio was used to analyze this. The odds ratio is 5.24, which means people with heart diseases are 5.24 times more likely to have strokes. Confidence interval for the odds ratio was calculated to see whether it is significant or not. Confidence interval is found as 3.61 and 7.61, it does not include 1, hence, it is not significant. Hence, there is no association between having a stroke and heart disease.

Hypertension

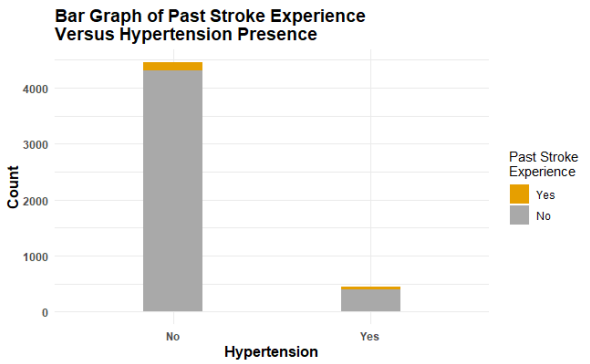
For hypertension, again a contingency table was constructed to see the association between hypertension and having a stroke.

	Has / Had Hypertension	Did not Have Hypertension
Had a stroke	4308	391
Have not had a stroke	149	60

Table 9: Contingency Table of Hypertension Status versus Past Strokes

When the odds ratio was examined, a value of 4.436739 was obtained. This means that a person with hypertension is 4.43 times more likely to have strokes.

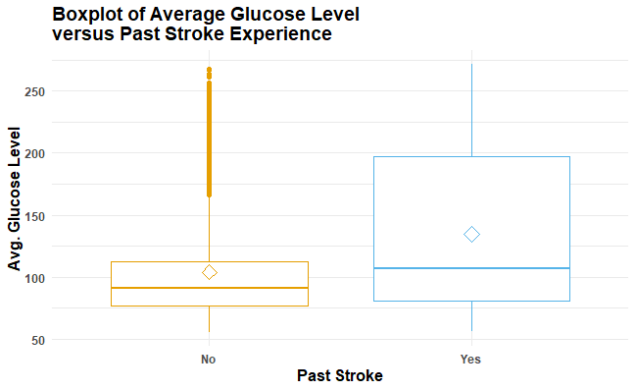
Figure 8: Bar Graph of Hypertension vs. Past Stroke Experience



Since the confidence interval (3.23, 6.09) of the odds ratio does not include 1, it can be deduced that hypertension has a significant effect. This result can be confirmed by looking at Figure 8.

Glucose Level

Figure 9: Boxplot of Average Glucose Level vs. Past Stroke Experience



Looking at Figure 9, it can be seen that the medians for the two categories are different and the data contains outliers.

To reach more precise results, Shapiro-Wilk normality test was performed.

H_0: They are coming from an identical population.

H_1: They are not coming from an identical population.

Since p-value $< 2.2*10^{-16}$ is smaller than $\alpha = 0.05$, the null hypothesis is rejected, the average glucose level is not normally distributed.

Box-cox transformation method was used to eliminate this situation, λ is found as -1. Shapiro-Walk test was performed again by taking $Y'=Y^{-1}$ is applied on average glucose level, but still, a normal distribution could not be obtained.

The Wilcoxon rank sum test was used to decide whether there was a relationship between glucose level and stroke.

H_0: They have equal medians.

H_1: They do not have equal medians.

It was determined that there was a relationship since p-value $< 8.038*10^{-10}$ is smaller than $\alpha = 0.05$, hence the null hypothesis is rejected. The results of the Polyserial correlation test showed that there is a weak positive correlation, 0.2274, between glucose level and stroke.

To sum up, in this research question, it can be concluded that present or past diseases / disorders, such as heart disease, hypertension, suboptimal glucose levels increase the likelihood of having a stroke.

Research Question 4: Can an accurate prediction be made for the possibility of having a stroke by using any or all of the independent variables?

Output 1: Full Model

```
Call:
glm(formula = stroke ~ ., family = "binomial", data = train.stroked)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1573  -0.2881  -0.1418  -0.0684   3.4829

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.373264   1.074452  -6.862 6.77e-12 ***
genderMale               0.030562   0.163739   0.187 0.851933
age                      0.079146   0.007001  11.305  < 2e-16 ***
hypertension1            0.544343   0.184280   2.954 0.003138 **
heart_disease1           0.208461   0.225332   0.925 0.354900
ever_marriedYes         -0.002259   0.278888  -0.008 0.993538
work_typeGovt_job       -1.264859   1.140972  -1.109 0.267611
work_typeNever_worked  -11.025735 533.302051  -0.021 0.983505
work_typePrivate        -1.098034   1.126342  -0.975 0.329626
work_typeSelf-employed  -1.526433   1.146664  -1.331 0.183125
Residence_typeUrban     -0.015727   0.159451  -0.099 0.921429
avg_glucose_level        0.004792   0.001373   3.489 0.000484 ***
bmi                      0.007794   0.012550   0.621 0.534601
smoking_statusnever smoked -0.061539   0.198853  -0.309 0.756963
smoking_statussmokes     0.270661   0.246405   1.098 0.272012
smoking_statusUnknown   -0.316336   0.264816  -1.195 0.232262
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1548.6  on 4416  degrees of freedom
Residual deviance: 1201.5  on 4401  degrees of freedom
AIC: 1233.5

Number of Fisher Scoring iterations: 15
```

In order to build a model, the train-test method was first tried. In the process, data was separated in a 9:1 ratio to train and test sets, respectively. Then, the following model, Output 1, was fit into the train model that includes all the independent variables and the binary stroke status as the response.

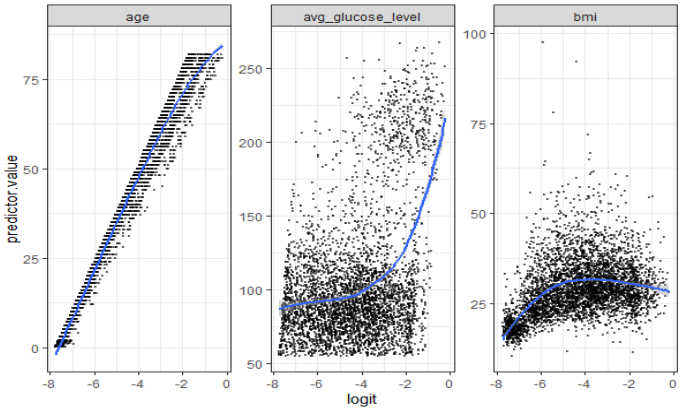
Output 2: Reduced Model

```
Call:
glm(formula = stroke ~ age + hypertension + avg_glucose_level,
    family = "binomial", data = train.stroked)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0657  -0.2891  -0.1515  -0.0689   3.6689

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.061359   0.424671 -18.983  < 2e-16 ***
age               0.073583   0.006013  12.238  < 2e-16 ***
hypertension1     0.574264   0.182071   3.154  0.00161 **
avg_glucose_level 0.005213   0.001322   3.942 8.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1548.6  on 4416  degrees of freedom
Residual deviance: 1213.4  on 4413  degrees of freedom
AIC: 1221.4

Number of Fisher Scoring iterations: 7
```

By looking at the coefficient table, the removal of insignificant variables from the model was considered necessary. To achieve that, forward selection, backward elimination, and stepwise regression methods were applied to the model, resulting in the following model, Output 2, with four independent variables.

Upon drawing the following logit versus variables plots, Figure 10, for the model, a log transformation on the average glucose level variable was seen as applicable since the graph showed exponential increase.

Figure 10: Logit versus Numerical Variables Plot

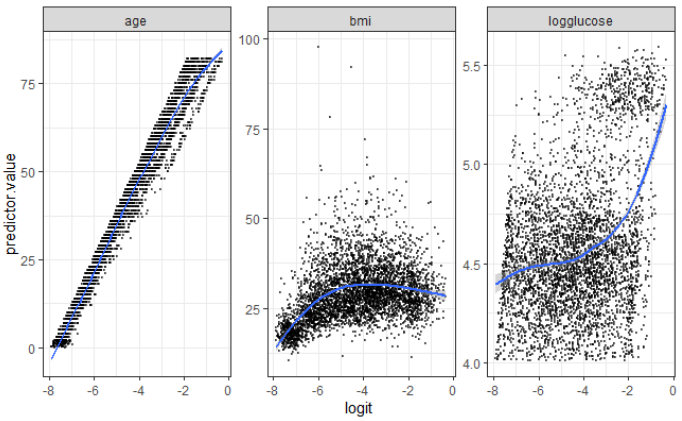


After the transformation of the variable was done, the output provided was as in Output 3:

Output 3: Transformed Model

```
Call:
glm(formula = stroke ~ age + hypertension + log(avg_glucose_level),
    family = "binomial", data = train.stroked)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0466  -0.2897  -0.1514  -0.0690   3.7032

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -10.651995   0.912783 -11.670  < 2e-16 ***
age                   0.073967   0.006004  12.320  < 2e-16 ***
hypertension1         0.580774   0.181904   3.193 0.001409 **
log(avg_glucose_level)   0.680078   0.180070   3.777 0.000159 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1548.6  on 4416  degrees of freedom
Residual deviance: 1214.2  on 4413  degrees of freedom
AIC: 1222.2

Number of Fisher Scoring iterations: 7
```

Figure 11: Logit vs. Transformed Num. Vars. Plot



And the logit versus variable plots were drawn, Figure 11, as a version with no significant improvement. Thus, the model with the original variable was chosen. The best threshold for the

model was considered to be 0.03641819 via the ROC curve, and the following confusion matrices were obtained from the train and test sets respectively.

Output 4: Confusion Matrices of Train and Test Performances

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3039   28
         1 1191  159

               Accuracy : 0.724
                 95% CI : (0.7106, 0.7372)
    No Information Rate : 0.9577
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1432

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.85027
            Specificity : 0.71844
         Pos Pred Value : 0.11778
         Neg Pred Value : 0.99087
             Prevalence : 0.04234
         Detection Rate : 0.03600
   Detection Prevalence : 0.30564
      Balanced Accuracy : 0.78435

       'Positive' Class : 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 279   3
         1 190  19

               Accuracy : 0.6069
                 95% CI : (0.5622, 0.6504)
    No Information Rate : 0.9552
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0908

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.86364
            Specificity : 0.59488
         Pos Pred Value : 0.09091
         Neg Pred Value : 0.98936
             Prevalence : 0.04481
         Detection Rate : 0.03870
   Detection Prevalence : 0.42566
      Balanced Accuracy : 0.72926

       'Positive' Class : 1
```

From the calculations in Output 4, it was seen that the model does not perform well enough on the test data, resulting in the decision of the train-test method being removed and a new model being fit to the whole data.

After the full model was conducted with all variables, the three AIC methods were again used on the model and the following resulting model with four independent variables was found.

Output 5: Final Reduced Model Fitted on Whole Data

```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
    family = "binomial", data = stroked)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0995  -0.2941  -0.1600  -0.0778   3.5884

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -7.660318   0.387164 -19.786  < 2e-16 ***
age               0.067540   0.005572  12.122  < 2e-16 ***
hypertension1     0.539597   0.173054   3.118 0.001820 **
heart_disease1    0.404308   0.203446   1.987 0.046889 *
avg_glucose_level 0.004802   0.001255   3.828 0.000129 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.3  on 4907  degrees of freedom
Residual deviance: 1374.6  on 4903  degrees of freedom
AIC: 1384.6

Number of Fisher Scoring iterations: 7
```

After similar logit versus variables plots with the predecessor models were obtained, another log transformation on the average glucose level variable was applied for the whole data.

However, as the improvement was not significant, the log transformation was discarded in order to have a model coefficient with easier interpretation. Hence, the original model with four independent coefficients was used, as seen in Output 5.

The model was then tested with a threshold value of 0.03626418 via ROC Curve, which provides an AUC value of 0.85. Afterwards, the resulting confusion matrix was examined closely.

Output 6: Confusion Matrix of the Final Model

```
Confusion Matrix and Statistics

          Reference                              Mcnemar's Test P-Value : <2e-16
Prediction    0    1                                      Sensitivity : 0.84211
         0 3322   33                                      Specificity : 0.70696
         1 1377  176                                   Pos Pred Value : 0.11333
                                                       Neg Pred Value : 0.99016
               Accuracy : 0.7127                           Prevalence : 0.04258
                 95% CI : (0.6998, 0.7253)             Detection Rate : 0.03586
    No Information Rate : 0.9574                   Detection Prevalence : 0.31642
    P-Value [Acc > NIR] : 1                           Balanced Accuracy : 0.77453

                  Kappa : 0.1348                     'Positive' Class : 1
```

In the end, it was seen on Output 6 that the model provides satisfactory accuracy, sensitivity, and specificity. While it also provided a good negative prediction value, it failed on yielding a decent positive prediction value. With all the results gathered it was decided that the model is an applicable fit on the data, however, it can still be enhanced by being trained on a better dataset that has a better positive and negative distribution ratio, as well as more observations.

4. Discussion

Health is one of the main necessities of human life. There are many diseases and medical conditions that can change human life dramatically. Having a stroke is one of those conditions. Although strokes occur mostly at the age of 65 and over with a high rate of 75%, the incidence increases in the 20-64 age range when people are at their most productive time of life (Purvis et al., 2021). This study examined the factors that have a relationship with having a stroke with various tests and analyses with the main goal of conducting a model to predict stroke status. Many relationships were found between the stroke status and the provided data. However, the stroke status, which was selected as the response for this research, was not evenly distributed. Moreover, additional information needed to be provided to achieve a better estimation. There

are 10.3 million new stroke cases in lower and middle-income countries (Katan & Luft, 2018). Therefore, income, country, and life status data could be used for further investigation and model improvement. Additionally, information about lifestyle and other health conditions such as diet, blood pressure, alcohol consumption, regular exercise, etc. could be used for the improvement of the research. According to Better Health Channel funded by the State Government of Victoria (Australia), factors such as high blood pressure, high blood cholesterol levels, heavy drinking, a diet high in fat (particularly saturated) and salt, lack of regular exercise, and obesity increase the likelihood of having a stroke. Further investigations can be done by adding such important covariates to the data.

References

Donkor, E. S. (2018, November 27). *Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life*. Stroke research and treatment. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6288566/.

Hemorrhagic Strokes (Bleeds). www.stroke.org. (n.d.). https://www.stroke.org/en/about-stroke/types-of-stroke/hemorrhagic-strokes-bleeds.

Ischemic Strokes (Clots). www.stroke.org. (n.d.). https://www.stroke.org/en/about-stroke/types-of-stroke/ischemic-stroke-clots.

Katan, M., & Luft, A. (2018). Global Burden of Stroke. *Seminars in Neurology*, *38*(02), 208–211. https://doi.org/10.1055/s-0038-1649503

Purvis, T., Hubbard, I. J., Cadilhac, D. A., Hill, K., Watkins, J., Lannin, N. A., Faux, S. G., & Kilkenny, M. F. (2021, March 16). *Age-Related Disparities in the Quality of Stroke Care and Outcomes in Rehabilitation Hospitals: The Australian National Audit*. Journal of Stroke and Cerebrovascular Diseases. https://www.sciencedirect.com/science/article/pii/S1052305721001105?via%3Dihub.

Stroke Risk Factors and Prevention. (n.d.). https://www.betterhealth.vic.gov.au/health/ConditionsAndTreatments/stroke-risk-factors-and-prevention.

Thapa, L., Shrestha, S., Kandu, R., Ghimire, M. R., Ghimire, S., Chaudhary, N. K., Pahari, B., Bhattarai, S., Kharel, G., Paudel, R., Jalan, P., Chandra, A., Phuyal, S., Adhikari, B., Aryal, N., & Kurmi, O. P. (2021). Prevalence of Stroke and Stroke Risk Factors in a South-Western Community of Nepal. *Journal of Stroke and Cerebrovascular Diseases*, *30*(5), 105716. https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.105716

Warlow, C. P. (1998). Epidemiology of Stroke. *The Lancet*, *352*. https://doi.org/10.1016/s0140-6736(98)90086-1

World Health Organization Eastern Mediterranean Region Office. (n.d.). Stroke, Cerebrovascular accident. http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html.