Stroke Prediction

Table of Contents

Abstract

This report centers around the prediction of the probability of suffering a stroke for an individual based on several characteristics, such as gender, age, whether the person has hypertension or heart disease, average glucose level after a meal, and body mass index (BMI), among others. The data set was extracted from Kaggle and consists of 12 variables for a total of 5,110 patients.

The report is organized as follows. Section I contains general information about the dataset and the research question. Section II includes steps taken to clean and/or transform the data set. Section III describes the initial variable selection methods for numerical and categorical predictors. Section IV and V deal with the identification of multicollinearity and interaction terms, respectively, for the logistic model to be constructed. Section VII contains the logistic regression model including final variable selection, model results, and classification report. Section VII concludes this report with a summary of the steps taken and any issues encountered in the previous sections.

I. General Information

Research Question

Strokes are a leading cause for adult deaths and disabilities. As a team, we wanted to do research on stroke probabilities for adult people, to see if we can predict the occurrence of stroke or not, given a set of characteristics or attributes.

The dataset itself was completely complying with the project requirements to implement Logistic Regression to calculate the prediction of an outcome, and several other techniques mentioned in the body of this report. The dataset has a number of key variables that are a combination of categorical and numerical, both of which are critical to make a prediction.

Additionally, the response variable, in our case stroke, is categorical (0 for patients who did not suffer a stroke, and 1 for patients who did) and logistic regression is best suited for doing analysis for such scenarios. The logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1.

Our population of interest is the patients of the hospital the dataset was collected from. However, it is specified in Kaggle that this is confidential data, so we don't know exactly which hospital this is. We are trying to predict whether or not a patient will suffer a stroke, based on several different attributes, using the prediction techniques learned in this course.

Dataset Description

The data set was obtained from Kaggle. It contains a total of 5110 observations with 12 variables. The table below summarizes a brief description of variable types.

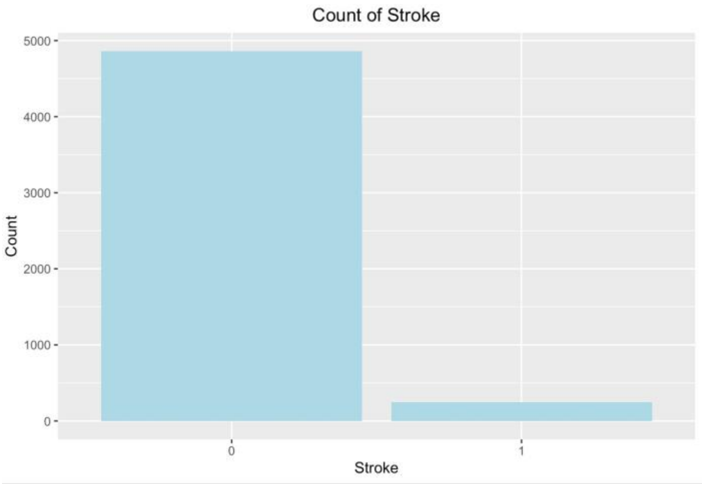
Variable	Description	Type
id	Unique Identifier for Patient	Numerical
gender	Gender of patient	Categorical
age	Age of patient	Numerical
hypertension	Hypertension status (0/1)	Categorical
heart_disease	Heart Disease status (0/1)	Categorical
ever_married	Marital Status (Yes/No)	Categorical
work_type	Type of work	Categorical
Residence_type	Type of residence	Categorical
avg_glucose_level	Glucose levels	Numerical
bmi	Body mass Index	Numerical
smoking_status	Smoking	Categorical

stroke	Suffered stroke(0/1)	Categorical

II. Data Screening

To start with the cleaning of the data, we first formatted all categorical variables as factors, so that R recognizes them as categorical, and not character. After this, we performed a summary (see Appendix for R output) for each of the variables, which led us to find that the gender variable had one observation labeled as “Other”. Since this was only one out of the 5,110 observations, we decided to remove the level “Other” from the gender variable. Additionally, we removed the ID variable because it is not going to be useful for our analysis. We also checked for duplicates and the data set did not have any.

The response variable “stroke” has 249 observations for patients who suffered a stroke (encoded as 1); the rest of the observations 4,860 belong to patients who did not suffer a stroke. The graph below helps visualize how imbalanced the data set is.



However, this may not represent an issue for our purpose of predicting stroke probability given the Central Limit Theorem, because we have more than 10 outcomes equal to 1 for every explanatory variable. We have 10 explanatory variables, and 249 outcomes equal to 1. Therefore, we have 24.9 outcomes equal to 1 for every explanatory variable.

We also noticed that the bmi variable had 201 missing values. These observations were not removed because out of them, 40 observations are those where the patient suffered a stroke. This information is valuable considering the fact that only 249 patients suffered a stroke in the dataset. Hence, records with empty value in BMI were replaced using the mean according to each gender. After the changes, we were left with a dataset containing 5,109 rows and 11 columns.

III. Initial Variable Selection

For the initial variable selection, we needed to identify which of the possible predictors have an effect or are associated with "stroke".

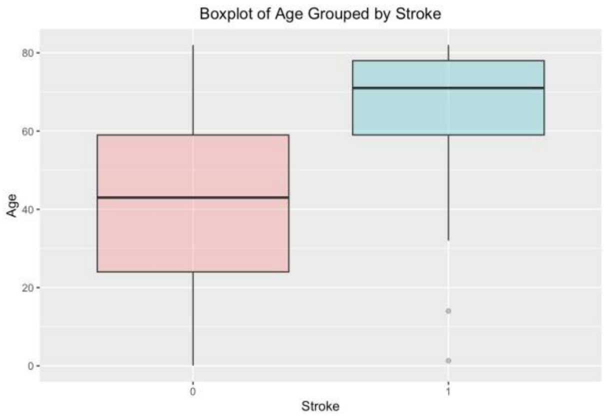
To identify numerical predictors, we first visualized the distribution in boxplots for each outcome of stroke and then performed Wilcoxon Rank Test, a non-parametric alternative to the t-test used for data that is not normally distributed. The difference between the tests is that t-test takes the null hypothesis as equal means, whereas wilcoxon rank test takes the null hypothesis as equal medians. To identify categorical predictors, we performed chi-square tests for all variables vs stroke.

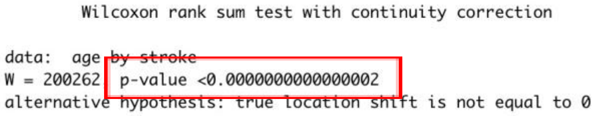
The results shown below are for the variables that were found to have an effect or to be associated with "stroke" (See Appendix for the rest of the variables).

Numerical Variables

1. **Age**

From the graph below, we can see that there's a clear difference in age median for each of the stroke outcomes, which could mean that age plays a significant role on whether or not an individual will have a stroke. To confirm this, let's also check the Wilcoxon Test results.





```
        Wilcoxon rank sum test with continuity correction

data:  age by stroke
W = 200262, p-value <0.0000000000000002
alternative hypothesis: true location shift is not equal to 0
```

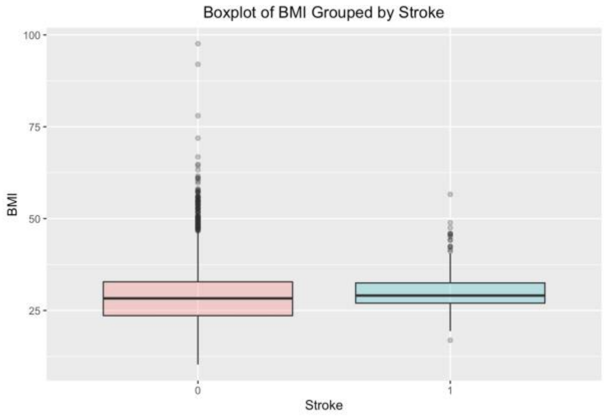
H0: The medians of the two groups are equal.

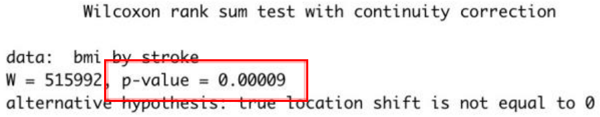
Ha: The medians of the two groups are different.

The p-value is less than 0.05, which allows us to reject the null hypothesis and confirms that age played a role on whether or not an individual had a stroke.

2. BMI

Even though the graph below does not show a clear difference in the median BMI for each stroke outcome, let’s check the Wilcoxon Test results.





```
        Wilcoxon rank sum test with continuity correction

data:  bmi by stroke
W = 515992, p-value = 0.00009
alternative hypothesis: true location shift is not equal to 0
```

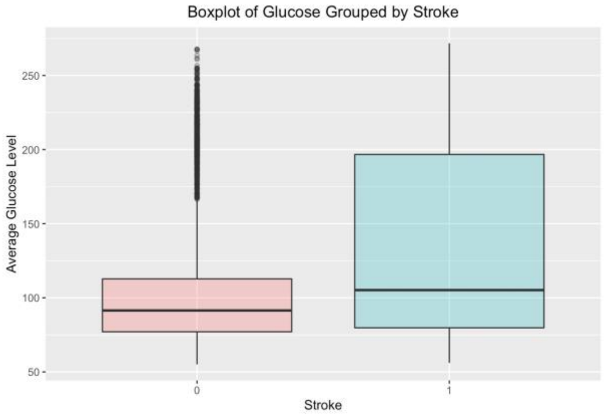
H0: The medians of the two groups are equal.

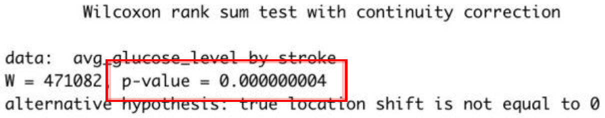
Ha: The medians of the two groups are different.

The p-value is less than 0.05, which allows us to reject the null hypothesis and confirms that BMI played a role on whether or not an individual had a stroke.

3. Average Glucose Level

The graph below shows a slight difference in the median of average glucose level for each stroke outcome. However, to confirm that the medians are statistically different, let's check the test results.





```
        Wilcoxon rank sum test with continuity correction

data:  avg_glucose_level by stroke
W = 471082, p-value = 0.000000004
alternative hypothesis: true location shift is not equal to 0
```

H0: The medians of the two groups are equal.

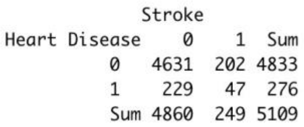
Ha: The medians of the two groups are different.

The p-value is less than 0.05, which allows us to reject the null hypothesis and confirms that Average Glucose Level played a role on whether or not an individual had a stroke.

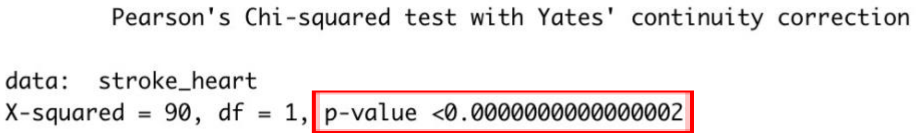
Categorical Variables

1. Heart Disease

To be able to perform a chi-square test between stroke and heart disease, we created a contingency table. All observed frequencies are greater than 5.



```
             Stroke
Heart Disease    0    1  Sum
          0   4631  202 4833
          1    229   47  276
          Sum 4860  249 5109
```

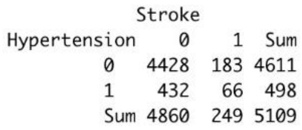


```
        Pearson's Chi-squared test with Yates' continuity correction

data:  stroke_heart
X-squared = 90, df = 1, p-value <0.0000000000000002
```

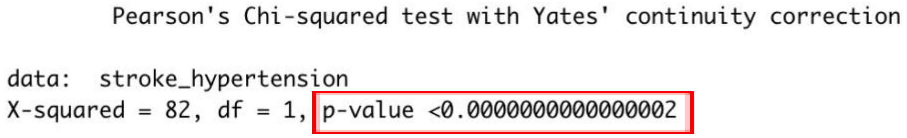
H0: Stroke is independent of Heart Disease Ha: Stroke and Heart Disease are not significantly independent

The p-value is less than 0.05, so we have enough statistical evidence to reject the null hypothesis, which confirms that Stroke and Heart Disease are not significantly independent. Whether or not an individual has heart disease played a role on whether or not they had a stroke.

2. Hypertension



```
               Stroke
Hypertension     0    1  Sum
         0    4428  183 4611
         1     432   66  498
         Sum  4860  249 5109
```

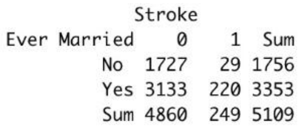


```
        Pearson's Chi-squared test with Yates' continuity correction

data:  stroke_hypertension
X-squared = 82, df = 1, p-value <0.0000000000000002
```

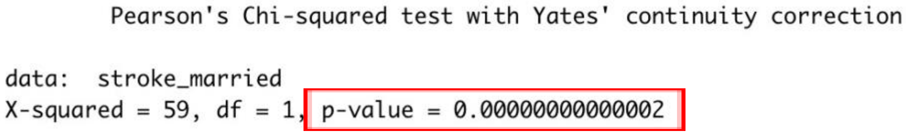
H0: Stroke is independent of Hypertension Ha: Stroke and Hypertension are not significantly independent

As before, all observed frequencies are greater than 5, so the chi-square test results can be relied on. The p-value is less than 0.05, so we have enough statistical evidence to reject the null hypothesis, which confirms that Stroke and Hypertension are not significantly independent. Whether or not an individual has hypertension played a role on whether or not they had a stroke.

3. Ever Married



```
              Stroke
Ever Married     0    1  Sum
         No   1727   29 1756
         Yes  3133  220 3353
         Sum  4860  249 5109
```



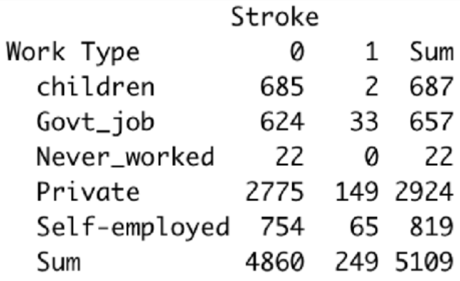
```
        Pearson's Chi-squared test with Yates' continuity correction

data:  stroke_married
X-squared = 59, df = 1, p-value = 0.00000000000002
```

H0: Stroke is independent of Ever Married Ha: Stroke and Ever Married are not significantly independent

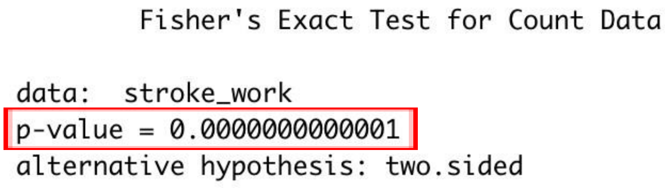
The p-value is less than 0.05, so we have enough statistical evidence to reject the null hypothesis, which confirms that Stroke and Ever Married are not significantly independent. Whether or not an individual has ever been married played a role on whether or not they had a stroke.

4. Stroke and Work Type

Since the observed frequencies of several cells are less than 5, we can't use the chi-square test. We'll use Fisher's Exact Test, an alternative test used when the sample is small.



```
               Stroke
Work Type         0    1  Sum
  children      685    2  687
  Govt_job      624   33  657
  Never_worked   22    0   22
  Private      2775  149 2924
  Self-employed 754   65  819
  Sum          4860  249 5109
```

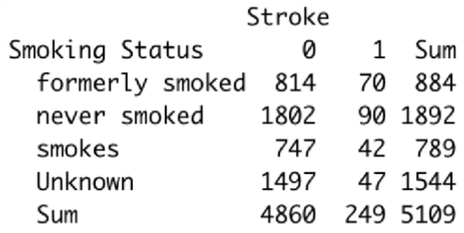


```
        Fisher's Exact Test for Count Data

data:  stroke_work
p-value = 0.0000000000001
alternative hypothesis: two.sided
```

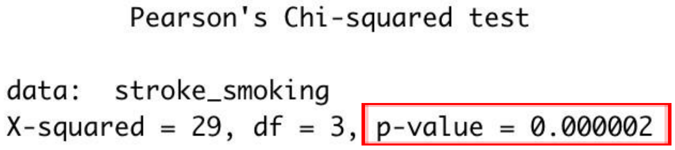
H0: Stroke is independent of Work Type Ha: Stroke and Work Type are not significantly independent

The p-value is less than 0.05, so we have enough statistical evidence to reject the null hypothesis, which confirms that Stroke and Work Type are not significantly independent. The type of work an individual had, played a role on whether or not they had a stroke.

5. Stroke and Smoking Status



```
                   Stroke
Smoking Status        0    1  Sum
  formerly smoked   814   70  884
  never smoked     1802   90 1892
  smokes            747   42  789
  Unknown          1497   47 1544
  Sum              4860  249 5109
```



```
        Pearson's Chi-squared test

data:  stroke_smoking
X-squared = 29, df = 3, p-value = 0.000002
```

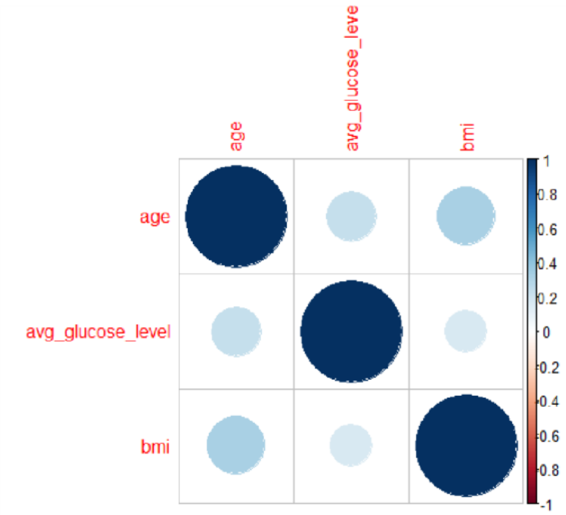
H0: Stroke is independent of Smoking Status Ha: Stroke and Smoking Status are not significantly independent

The p-value is less than 0.05, so we have enough statistical evidence to reject the null hypothesis, which confirms that Stroke and Smoking Status are not significantly independent. The smoking status of an individual played a role on whether or not they had a stroke.

IV. Multicollinearity Detection

Multicollinearity is defined as the presence of high correlation between two or more independent variables in a regression model. It can lead to skewed or misleading results when we try to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model. Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. In other words, it results when you have factors that are a bit redundant. In general multicollinearity should not be present in the model.

For initial testing we used the correlation coefficient to check if any of the predictors are closely related to each other. The graph and table below, show the correlation plot and coefficient for numerical predictors. Initial assessment shows there may not be multicollinearity present in our dataset.



	Age	Glucose Level	BMI
Age	1	0.2383228	0.325823
Glucose Level	0.238323	1	0.168785
BMI	0.325823	0.1687847	1

To dig deeper, we ran the logistic model to predict stroke including all the explanatory variables initially selected from the previous section. To this model, we calculated the Generalized Variance Inflation Factor (GVIF) for each predictor. GVIFs greater than 5 indicate the possible presence of multicollinearity in a model. The table below shows that our initial model is free from the presence of multicollinearity.

Predictor	GVIF^(1/(2*Df))
Age	1.20
Hypertension	1.03
Heart Disease	1.04

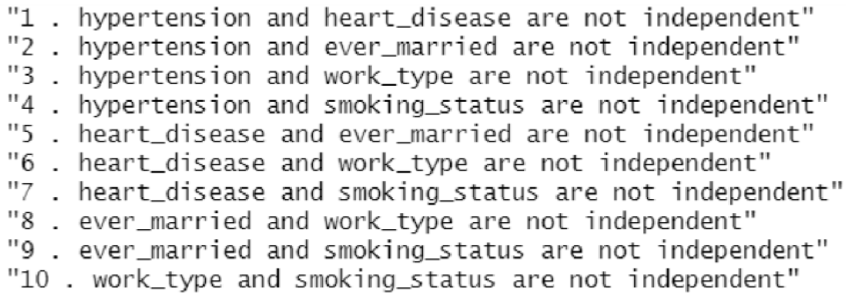
Average Glucose Level	1.05
BMI	1.05
Ever Married	1.05
Work Type	1.04
Smoking Status	1.01

V. Interaction Terms Detection

In a regression model, interaction occurs when two predictors have a joint effect on the response variable. Interactions can occur between two categorical predictors, between two numerical predictors and between one numerical and one categorical predictors. Based on the variables selected in section IV, we performed the following interaction detection processes.

Interaction Between Two Categorical Variables

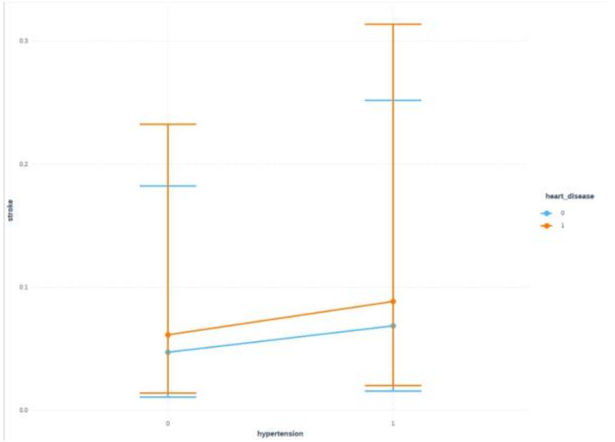
First, we performed a chi-squared test between all possible pairs of categorical predictors to identify whether or not there exists association between them. The 10 pairs of predictors below show potential interaction between them.



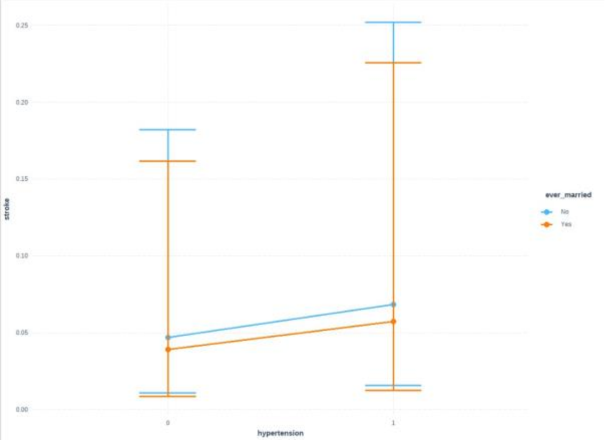
```
"1 . hypertension and heart_disease are not independent"
"2 . hypertension and ever_married are not independent"
"3 . hypertension and work_type are not independent"
"4 . hypertension and smoking_status are not independent"
"5 . heart_disease and ever_married are not independent"
"6 . heart_disease and work_type are not independent"
"7 . heart_disease and smoking_status are not independent"
"8 . ever_married and work_type are not independent"
"9 . ever_married and smoking_status are not independent"
"10 . work_type and smoking_status are not independent"
```

Let's visualize this potential interaction with an interaction plot for two pairs of categorical variables (cat_plot function in R, See Appendix for the graphs of all variable pairs).

Hypertension and Heart Disease



Hypertension and Ever Married

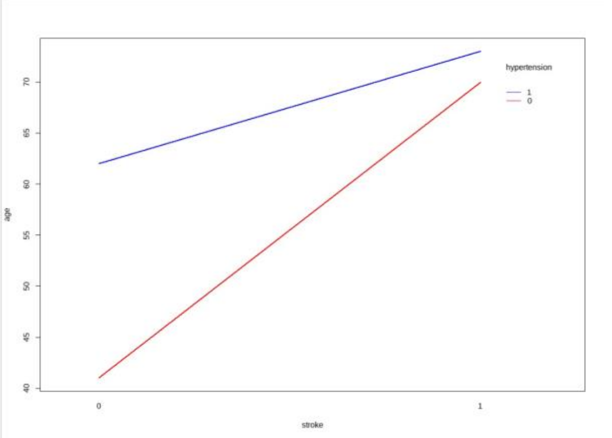


The graphs above confirm the presence of potential interaction between the respective variables. We can see that the lines are not completely parallel, which means that eventually they will cross. This is an indication that hypertension and heart disease or hypertension and ever being married may have a joint effect on our response variable, whether or not a patient will suffer a stroke.

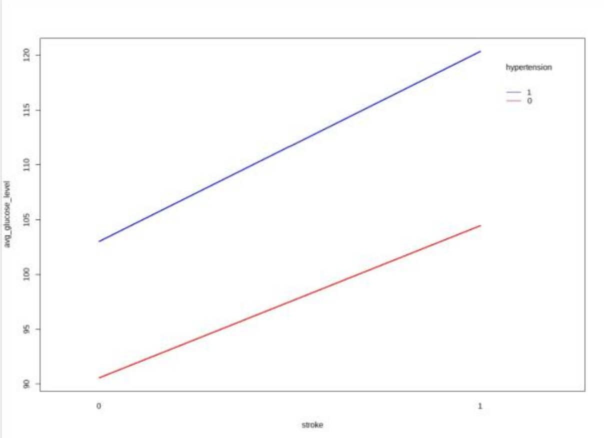
Interaction Between One Categorical and One Numerical Variables

To identify potential interactions between one categorical and one numerical variable, we created an interaction plot for every possible pair of numerical-categorical variables (interaction_plot function in R). Visualize the graphs below.

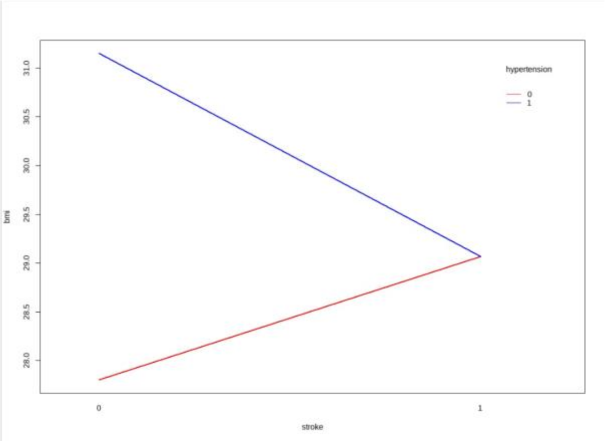
Hypertension and Age



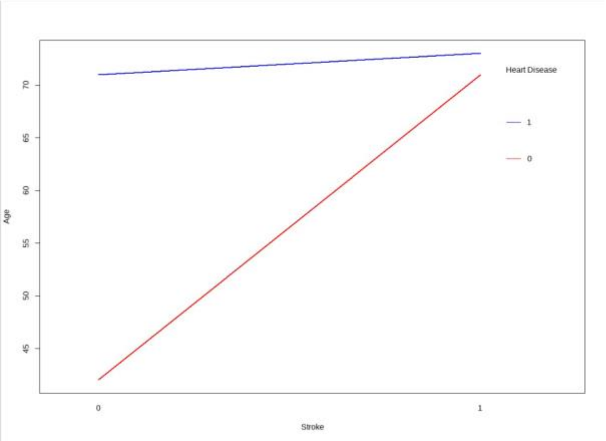
Hypertension and Average Glucose Level



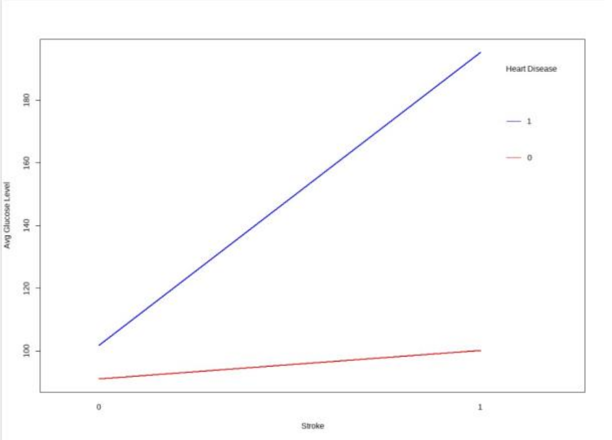
Hypertension and BMI



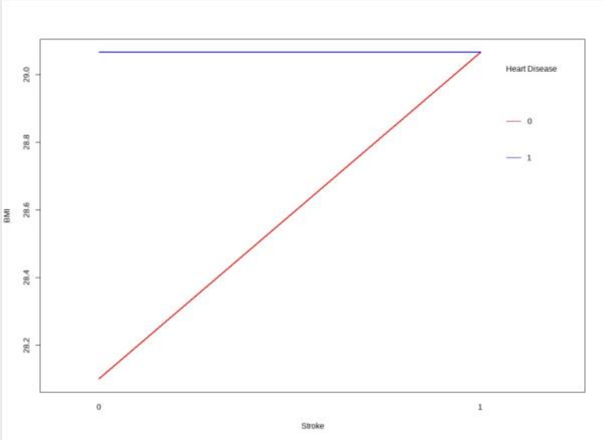
Heart Disease and Age



Heart Disease and Average Glucose Level



Heart Disease and BMI

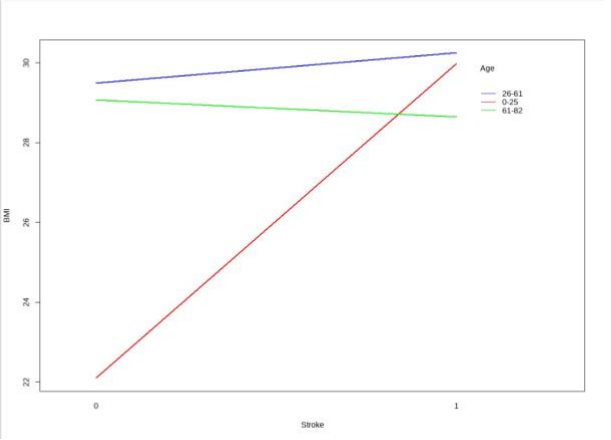


The six pairs of variables shown in the graphs above are the pairs that show the strongest interaction (See Appendix for the rest of the graphs).

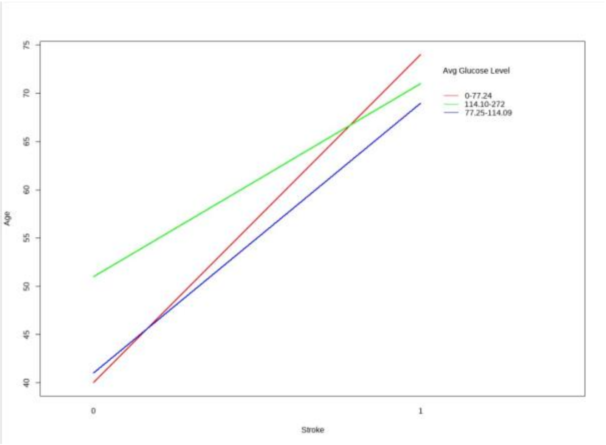
Interaction Between Two Numerical Variables

To identify interactions between two numerical variables, one of the numerical variables was binned into categories using the quartiles as reference. After that, the variables were plotted using the interaction plot for numerical vs categorical interactions. See the graphs below.

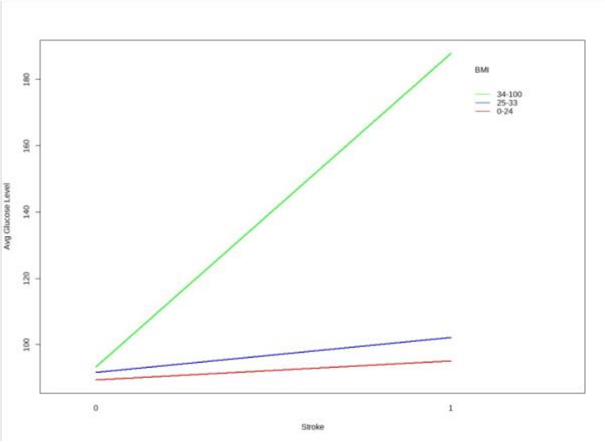
Age and BMI



Age and Average Glucose Level



BMI and Average Glucose Level



Again, the graphs shown above represent the strongest interactions found between two numerical variables (See Appendix for the rest of the graphs).

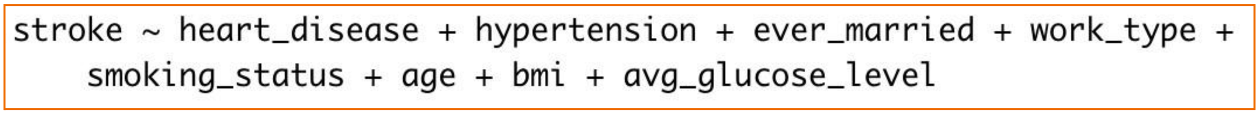
VI. Logistic Regression Model

Before building the Logistic Regression Model, the stroke dataset was split into training and testing sets at a ratio of 70:30, respectively. The distribution of the target variable among the training and the testing dataset is as follows:

Dataset	No Stroke (0)	Stroke (1)
Training	3,402	175
Testing	1,458	74

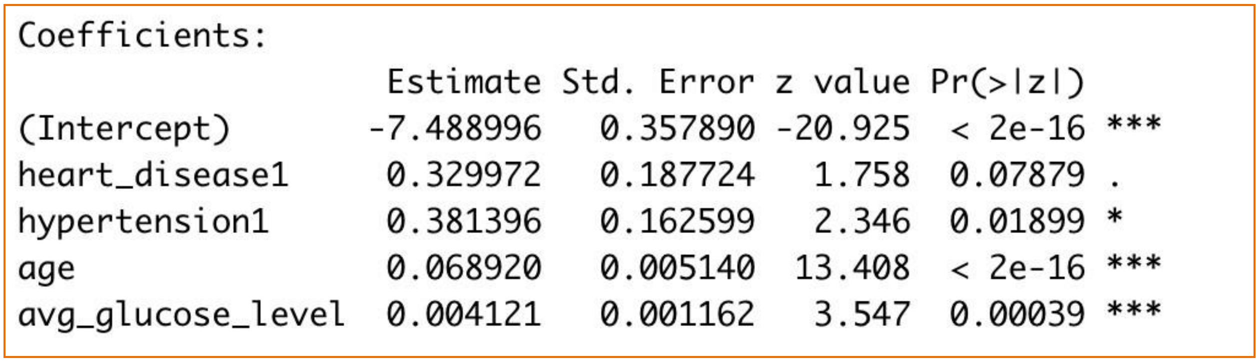
Final Variable Selection

To determine the variables to be used for building a logistic model, the backward elimination process has been implemented on R. Since the backward elimination process requires a model as an input, we created a model with the variables selected in section IV. Below is the model specification from R.



```
stroke ~ heart_disease + hypertension + ever_married + work_type +
    smoking_status + age + bmi + avg_glucose_level
```

On executing the code, the final step of the backward elimination process presented the following set of variables:



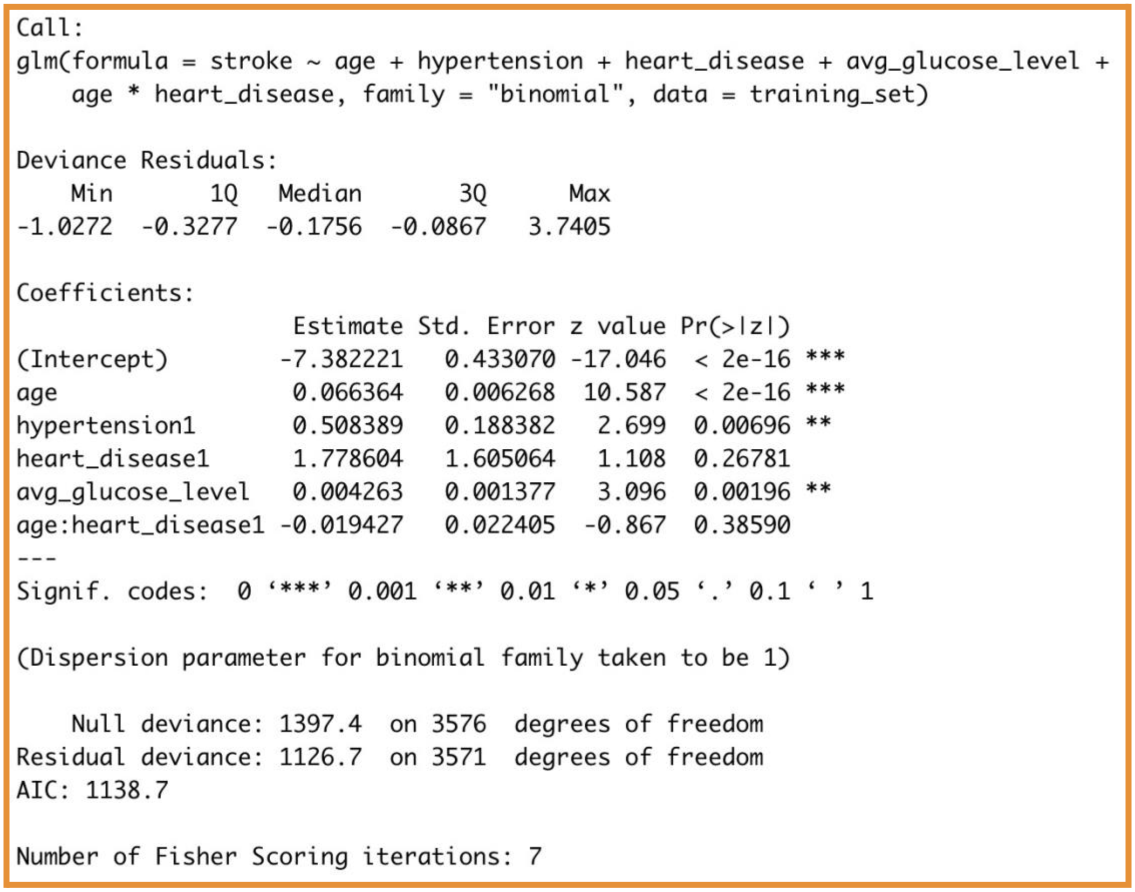
```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -7.488996   0.357890 -20.925  < 2e-16 ***
heart_disease1     0.329972   0.187724   1.758  0.07879 .  
hypertension1      0.381396   0.162599   2.346  0.01899 *  
age                0.068920   0.005140  13.408  < 2e-16 ***
avg_glucose_level  0.004121   0.001162   3.547  0.00039 ***
```

From the above results, it can be observed that four variables are selected of the total eight variables initially provided. Although ‘heart_disease’ can be seen having a low significance (p-value > 0.05), it has been retained out of domain knowledge.

In Addition to the model with the four variables selected, we will build a logistic model including an interaction between age and heart disease, which we found to have a joint effect on stroke from section V.

Logistic Model with Interaction

With the results from the backward elimination, a logistic model was built using the training dataset. To the variables selected, we added an interaction between age and heart disease. The summary of the model built is shown below.



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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0272  -0.3277  -0.1756  -0.0867   3.7405

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.382221   0.433070 -17.046  < 2e-16 ***
age                0.066364   0.006268  10.587  < 2e-16 ***
hypertension1      0.508389   0.188382   2.699  0.00696 **
heart_disease1     1.778604   1.605064   1.108  0.26781
avg_glucose_level  0.004263   0.001377   3.096  0.00196 **
age:heart_disease1 -0.019427   0.022405  -0.867  0.38590
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1397.4  on 3576  degrees of freedom
Residual deviance: 1126.7  on 3571  degrees of freedom
AIC: 1138.7

Number of Fisher Scoring iterations: 7
```

This result can also be expressed as the following equation:

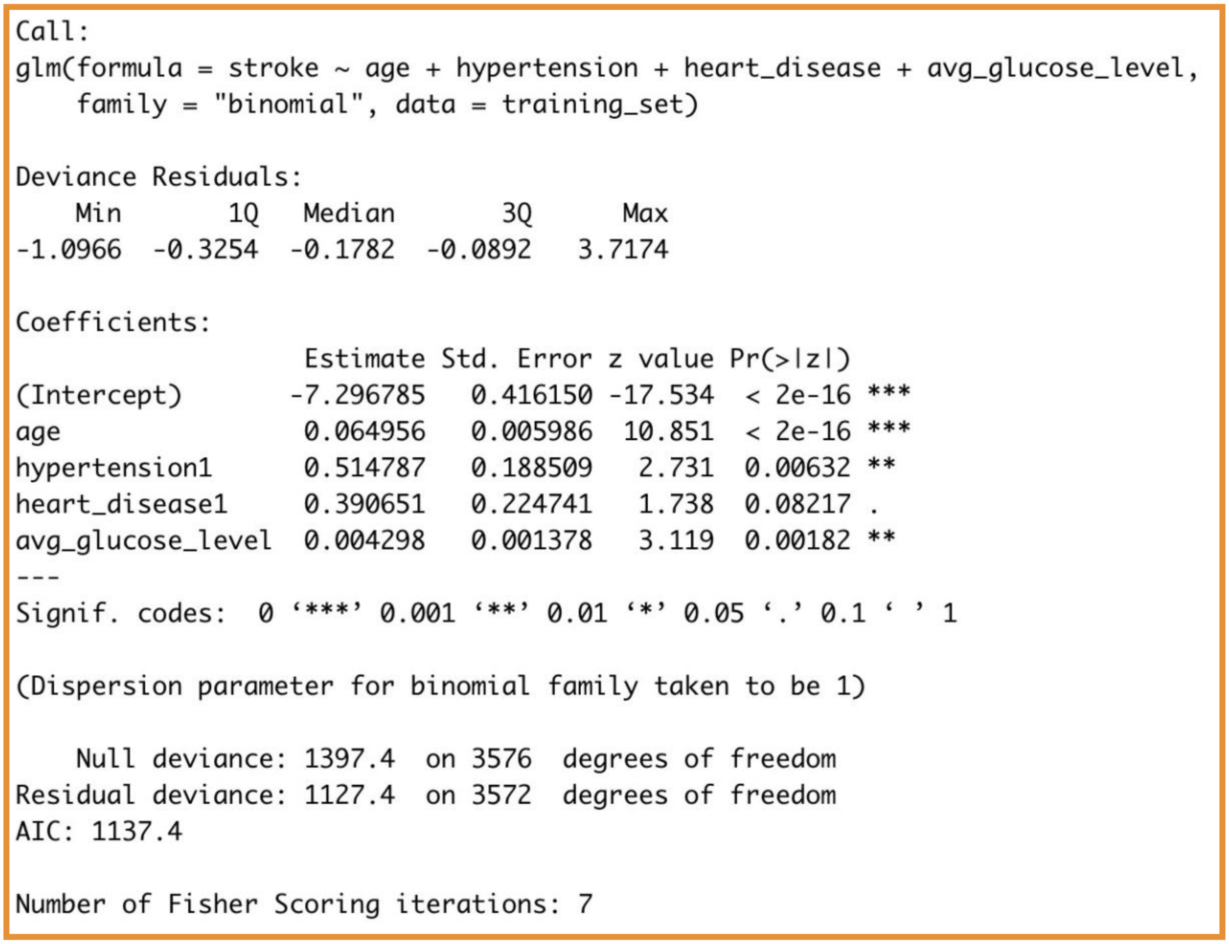
$$log\left(\frac{\pi_{stroke}}{1-\pi_{stroke}}\right) =- 7.38 + 0.06 * age + 0.51 * hypertension + 1.78 * heartdisease + 0.004 * avgglucoselevel - 0.02 * (age * heartdisease)$$

From the above summary, we can see that the sign of the intercept is negative, which implies that the probability of having a stroke is less than 0.5 for all patients. Additionally, the sign of the coefficients for age, hypertension, heart disease and avg glucose level is positive, implying that an increase on either of these attributes will result in an increase of the probability of suffering a stroke. On the other hand, the interaction term has a negative coefficient which means that there is an inverse relationship between stroke and ‘age*heart_disease’.

Now, from the p-values, it can be noted that the variables age, hypertension and average glucose level are statistically significant (as p-value < 0.05) in predicting stroke. At the same time, heart disease and age*heart_disease are not significant (as p-value > 0.05) in predicting stroke.

Logistic Model without Interaction

With the results from the backward elimination, a logistic model was built using the training dataset. The summary of the generated model is as follows:



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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0966  -0.3254  -0.1782  -0.0892   3.7174

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.296785   0.416150 -17.534  < 2e-16 ***
age                0.064956   0.005986  10.851  < 2e-16 ***
hypertension1      0.514787   0.188509   2.731  0.00632 **
heart_disease1     0.390651   0.224741   1.738  0.08217 .
avg_glucose_level  0.004298   0.001378   3.119  0.00182 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1397.4  on 3576  degrees of freedom
Residual deviance: 1127.4  on 3572  degrees of freedom
AIC: 1137.4

Number of Fisher Scoring iterations: 7
```

This result can also be expressed as the following equation:

$$log(\frac{\pi^{stroke}}{1-\pi_{stroke}}) =- 7.30 + 0.06 * age + 0.51 * hypertension + 0.39 * heartdisease + 0.004 * avgglucoselevel$$

From the above summary, we can also see a negative intercept, which implies that the probability of having a stroke is less than 0.5. Also, the sign of the coefficients for age, hypertension, heart disease and avg glucose level is positive, implying that an increase on either of these attributes will result in an increase of the probability of suffering a stroke.

Looking at the p-values, it can be noted that the variables age, hypertension and average glucose level are statistically significant (as p-value < 0.05) in predicting stroke. At the same time, heart disease is not significant (as p-value > 0.05) in predicting stroke.

Model Comparison

As part of assessing the better model of the two models generated above, AIC and LRT test have been performed. The results are shown below.

AIC

From the model summaries presented above, it can be noted that the AIC's for the model with interaction and the model without interaction are 1138.7 and 1137.4, respectively.

```
AIC: 1138.7    AIC: 1137.4
```

As low AIC indicates a better model, the model without interaction is the better of the two. Since the difference between the two AIC's is very small, further the LRT test has been conducted to identify the better model.

Log Likelihood Ratio Test

A log likelihood ratio test has been performed between the models, with interaction and without interaction. The hypotheses for the test are written below.

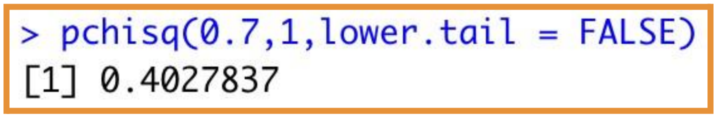
Null Hypothesis: Reduced Model is Appropriate (no interaction)

Alternate Hypothesis: Full Model is Appropriate (with interaction)

To get the LRT statistic, we subtracted the residual deviance of the full model, from the residual deviance of the reduced model. This statistic approximates to the chi-square statistic with degrees of freedom from the same subtraction. The p-value is calculated below using R.

$$LRT = Residual\ Deviance_{reduced\ model} - Residual\ Deviance_{full\ model}$$

$$LRT = 1127.4 - 1126.7 = 0.7; DF = 3572 - 3571 = 1$$
$$LRT = 0.7\ ; DF = 1$$

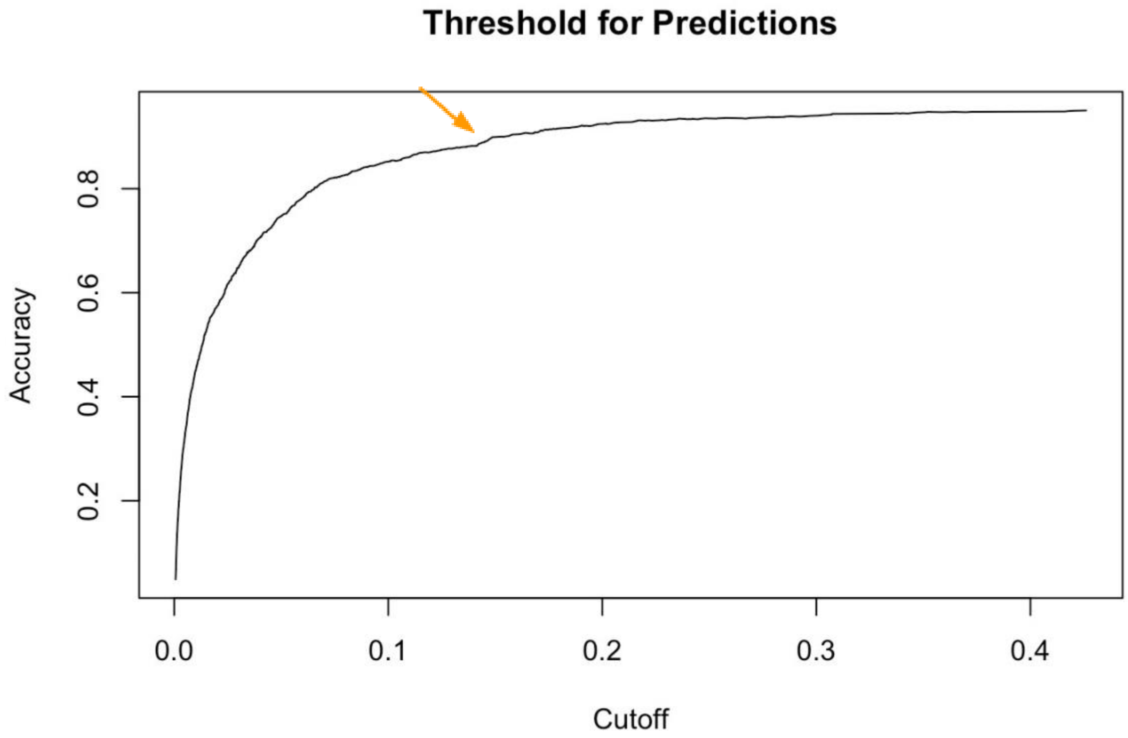


```
> pchisq(0.7,1,lower.tail = FALSE)
[1] 0.4027837
```

Since the p-value is greater than 0.05, we accept the null hypothesis and conclude that the reduced model or the model without interaction is appropriate.

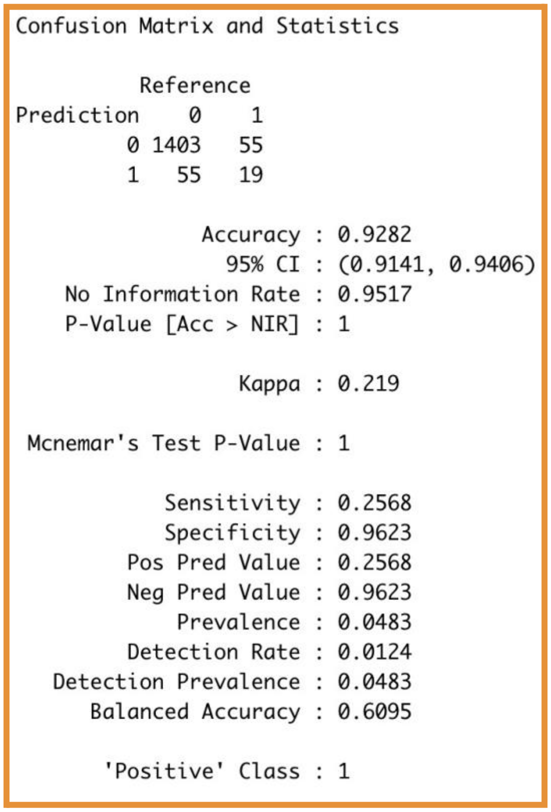
Classification Report

The classification reports were generated on R using the confusionMatrix function for the models with interaction and without interaction. The predictions are made on the test data using the models generated over the trained data. The cut-off point for the predictions was chosen by referring to the 'Accuracy vs cutoff graph'. From the below graph it can be noted that a threshold close to 0.2 is the most ideal point and hence 0.2 is used as the cut-off for the predictions.



From the classification reports shown below, the positive class is defined as '1', which is having suffered a stroke. The resulting reports are as follows:

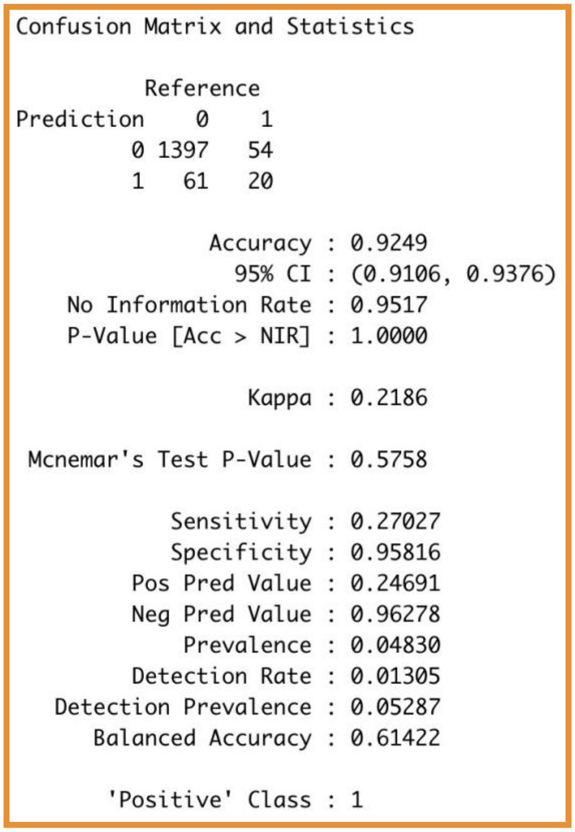
Model with Interaction



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1403   55
         1   55   19

               Accuracy : 0.9282
                 95% CI : (0.9141, 0.9406)
    No Information Rate : 0.9517
    P-Value [Acc > NIR] : 1

                  Kappa : 0.219

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.2568
            Specificity : 0.9623
         Pos Pred Value : 0.2568
         Neg Pred Value : 0.9623
             Prevalence : 0.0483
         Detection Rate : 0.0124
   Detection Prevalence : 0.0483
      Balanced Accuracy : 0.6095

       'Positive' Class : 1
```

From the above results, it can be noticed that the model’s accuracy is 92.82%. But looking at specificity and sensitivity, it can be observed that while specificity is 96.23%, the sensitivity is very low at 25.68%. This may be due to the imbalance in our data set. The model is doing a good job at predicting not suffering a stroke, but it is lacking when predicting suffering a stroke, which is the interesting outcome in our case.

Model without Interaction

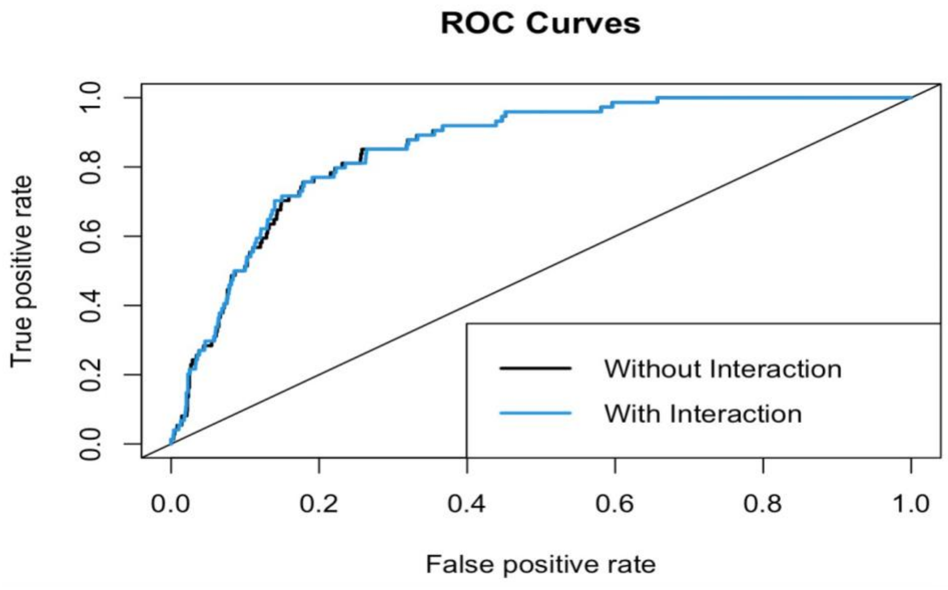


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1397   54
         1   61   20

               Accuracy : 0.9249
                 95% CI : (0.9106, 0.9376)
    No Information Rate : 0.9517
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.2186

 Mcnemar's Test P-Value : 0.5758

            Sensitivity : 0.27027
            Specificity : 0.95816
         Pos Pred Value : 0.24691
         Neg Pred Value : 0.96278
             Prevalence : 0.04830
         Detection Rate : 0.01305
   Detection Prevalence : 0.05287
      Balanced Accuracy : 0.61422

       'Positive' Class : 1
```

Similar results can be observed in the model without interaction where the model's accuracy is 92.49%. While the specificity is 95.81%, the sensitivity is 27.02%. Again, the model is doing a good job at predicting not suffering a stroke, but it is lacking when predicting suffering a stroke, which is the interesting outcome in our case.

ROC Curve

The below ROC curve has been generated on R. The graph compares the ROC curve for both models, with interaction and without interaction. Since the diagonal of an ROC curve graph represents the threshold points where the true positive rate is equal to the false positive rate, the ideal threshold for predictions would be a point far from the diagonal. Looking at the graph, it can be observed that 0.2 would be the ideal point as it is the farthest. Hence, 0.2 has been used as the threshold point for all the predictions made above. It should also be noted that the curve is highly similar for values with and without interaction.



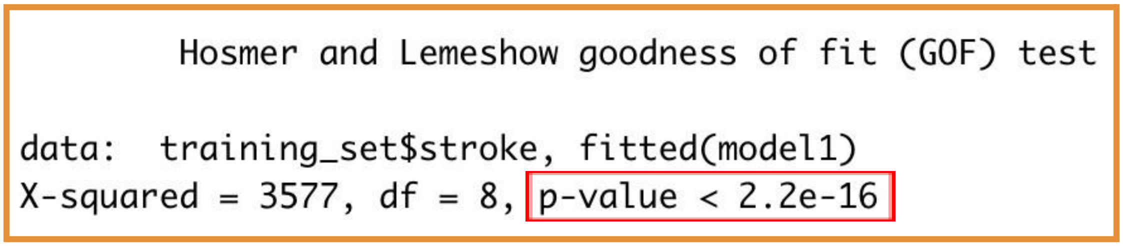
Lack of fit test

A lack of fit test has been conducted for the two models (with and without interaction) to assess if either of them is a good fit. Since the data is ungrouped, the Hosmer-Lemshow test has been performed. The Hypothesis for the test are as follows:

Null Hypothesis: Reduced model is appropriate (our model)

Alternate Hypothesis: Full model is appropriate (saturated model)

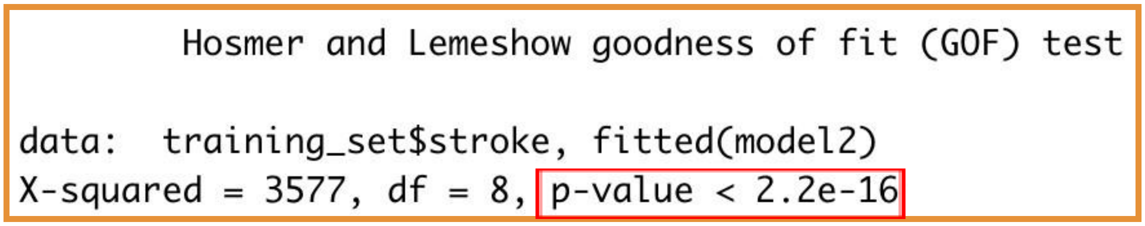
Model with Interaction



```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  training_set$stroke, fitted(model1)
X-squared = 3577, df = 8, p-value < 2.2e-16
```

From the above results, we can see that the p-value is less than 0.05. Hence, we reject the null hypothesis and conclude that the full model or the saturated model is appropriate.

Model without Interaction



```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  training_set$stroke, fitted(model2)
X-squared = 3577, df = 8, p-value < 2.2e-16
```

Similar results as the model with interaction can be noted here. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that the full model or the saturated model is appropriate.

VII. Summary

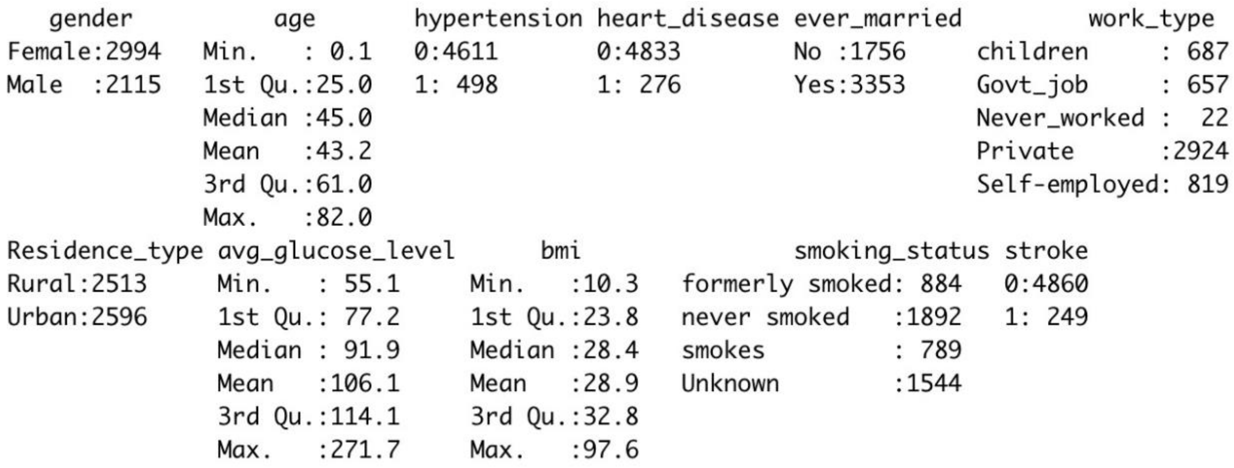
To perform the analysis to predict whether or not a patient will suffer a stroke, we built two logistic regression models using attributes of the patient such as age, whether or not the patient suffers from hypertension, heart disease and their average glucose level, along with an interaction term between age and heart disease. The two models consisted of one with the interaction, and one without it.

The performance of both models is almost the same. However, the AIC is lower for the model without interaction and the Log Likelihood Ratio test indicates that the interaction term is not significant for the prediction of stroke, showing that the model without interaction is better. In the same sense, if the performance of the models is similar, then we should always go with the less complex model. Therefore, our final model is the model without interaction term.

The accuracy of our final model is 92.49%, but the sensitivity is approximately 27% which indicates that our model misclassified the patients that had a stroke (positive cases), which is the outcome we’re interested in. This could be due to the unbalanced data. In our dataset, the number of patients who suffered stroke is only 249; the rest of the observations are 4,860. If we feed more data to the model with more positive cases; performance of our model may increase.

VIII. Appendix

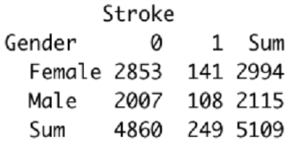
Dataset Summary



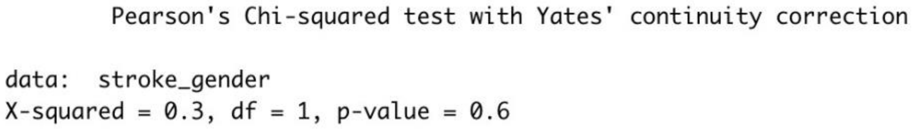
```
    gender          age        hypertension heart_disease ever_married        work_type
 Female:2994   Min.   : 0.1   0:4611       0:4833        No :1756     children     : 687
 Male  :2115   1st Qu.:25.0   1: 498       1: 276        Yes:3353     Govt_job     : 657
               Median :45.0                                           Never_worked :  22
               Mean   :43.2                                           Private      :2924
               3rd Qu.:61.0                                           Self-employed: 819
               Max.   :82.0
 Residence_type avg_glucose_level      bmi                  smoking_status stroke
 Rural:2513     Min.   : 55.1     Min.   :10.3   formerly smoked: 884   0:4860
 Urban:2596     1st Qu.: 77.2     1st Qu.:23.8   never smoked   :1892   1: 249
                Median : 91.9     Median :28.4   smokes         : 789
                Mean   :106.1     Mean   :28.9   Unknown        :1544
                3rd Qu.:114.1     3rd Qu.:32.8
                Max.   :271.7     Max.   :97.6
```

Initial Variable Selection

1. Stroke vs Gender

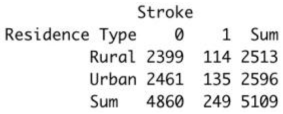


```
        Stroke
Gender      0    1  Sum
  Female 2853  141 2994
  Male   2007  108 2115
  Sum    4860  249 5109
```

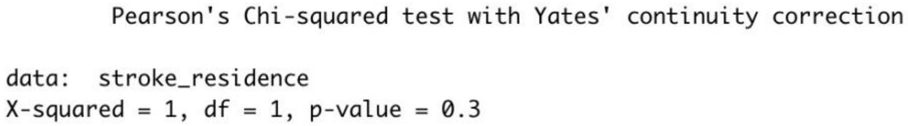


```
	Pearson's Chi-squared test with Yates' continuity correction

data:  stroke_gender
X-squared = 0.3, df = 1, p-value = 0.6
```

2. Stroke vs Residence Type



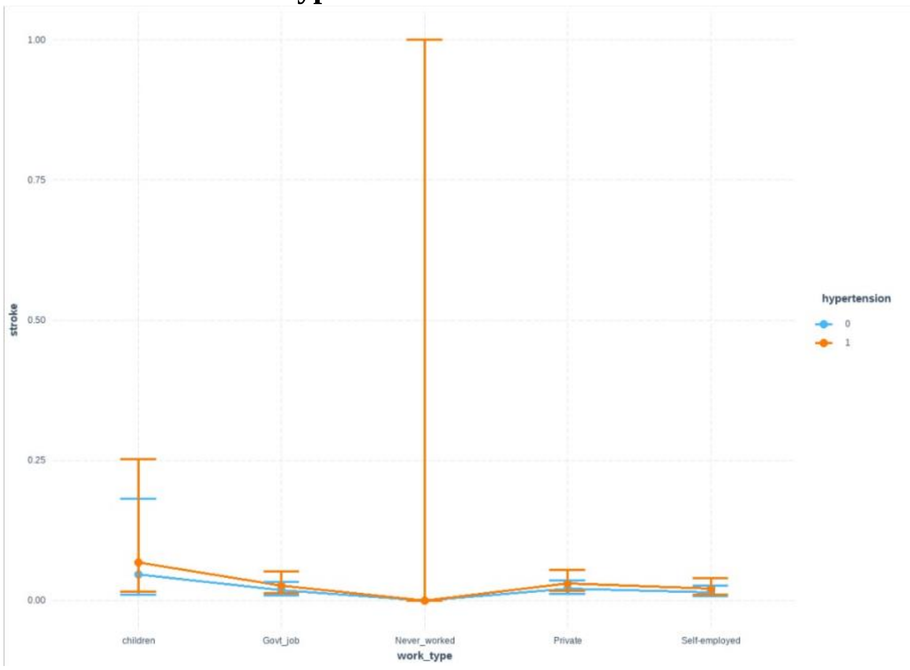
```
              Stroke
Residence Type    0    1  Sum
         Rural 2399  114 2513
         Urban 2461  135 2596
         Sum   4860  249 5109
```



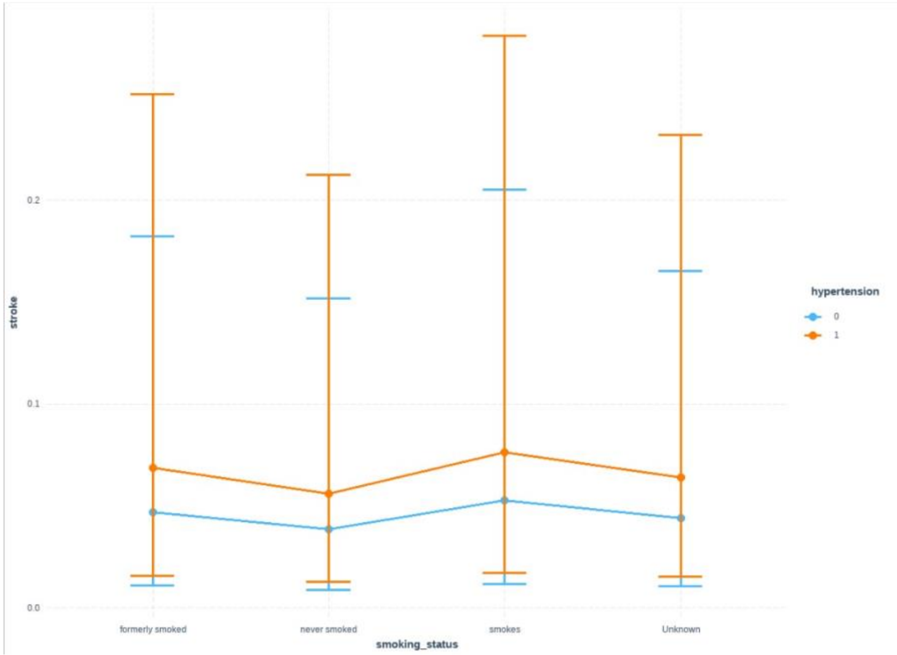
```
	Pearson's Chi-squared test with Yates' continuity correction

data:  stroke_residence
X-squared = 1, df = 1, p-value = 0.3
```

Interaction Plots

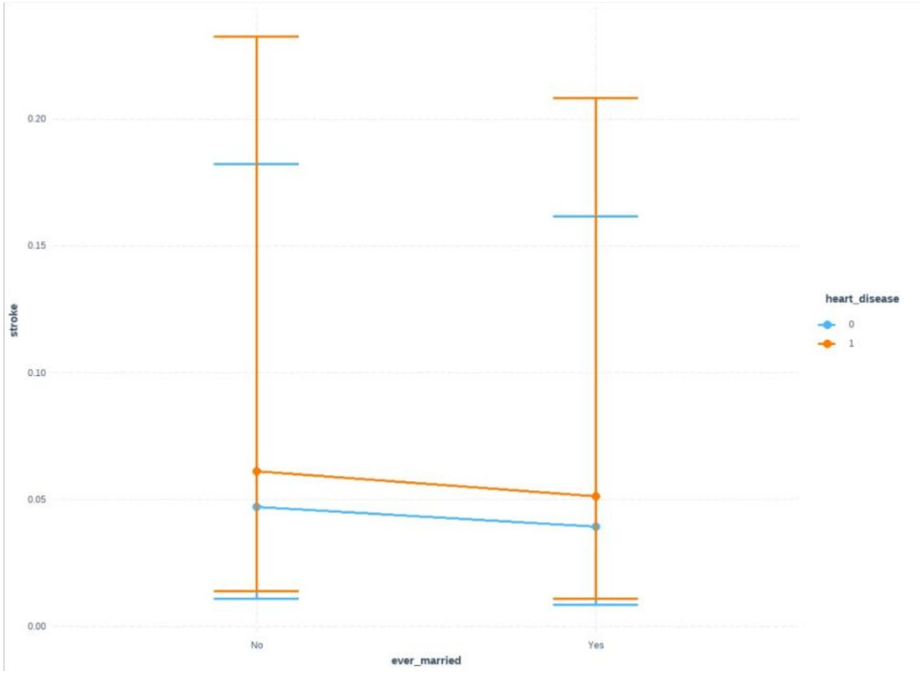
1. Hypertension and Work Type



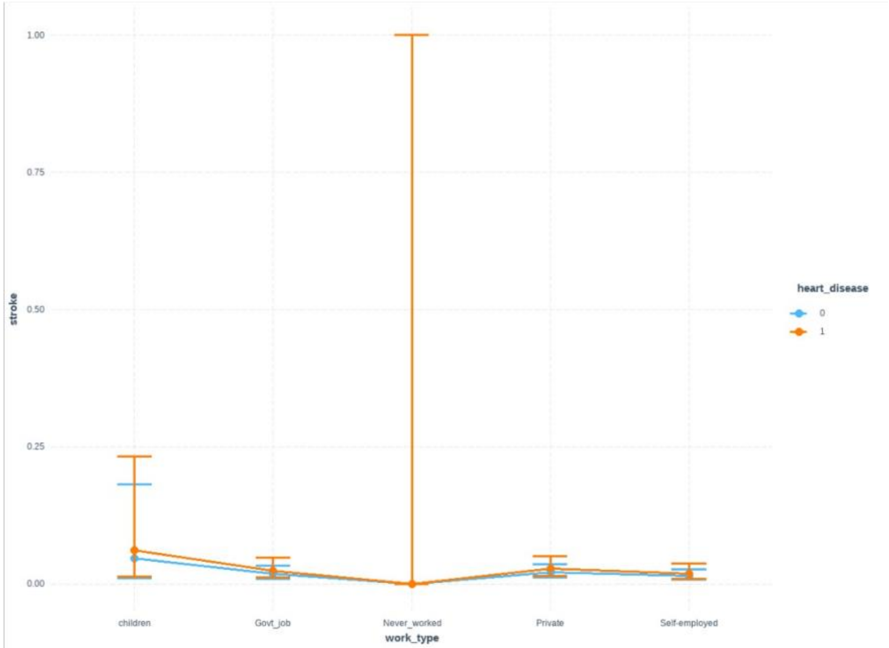
2. Hypertension and Smoking Status



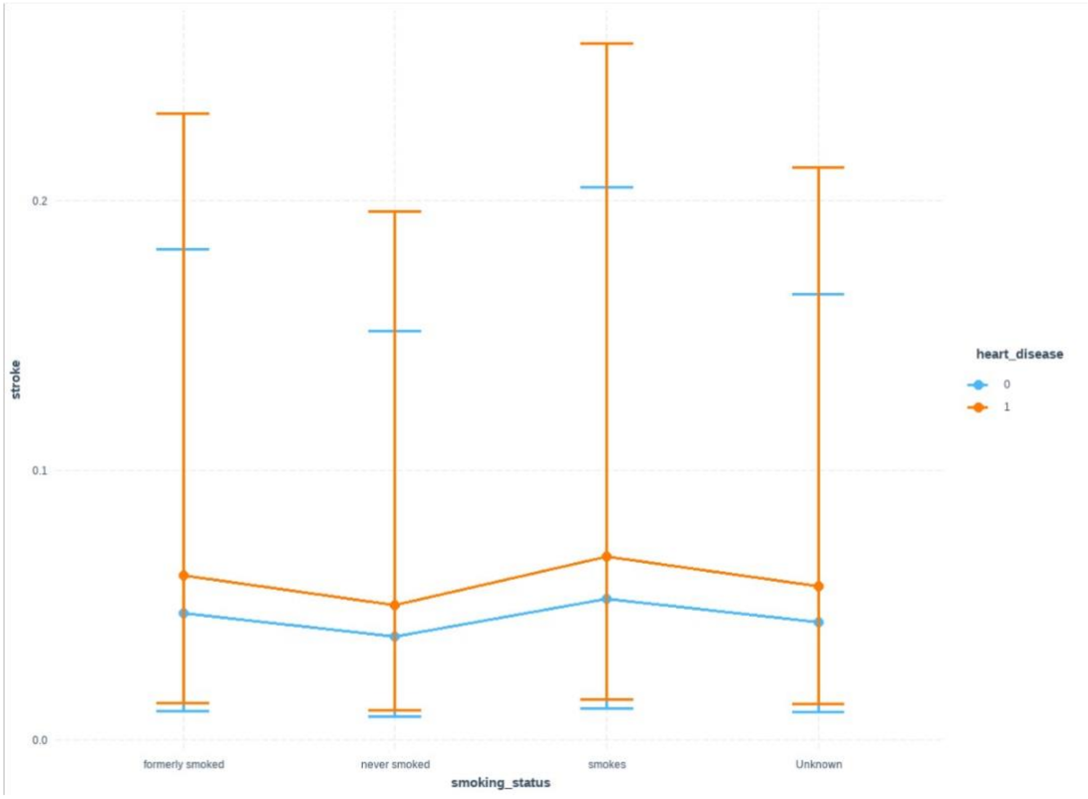
3. Heart Disease and Ever Married



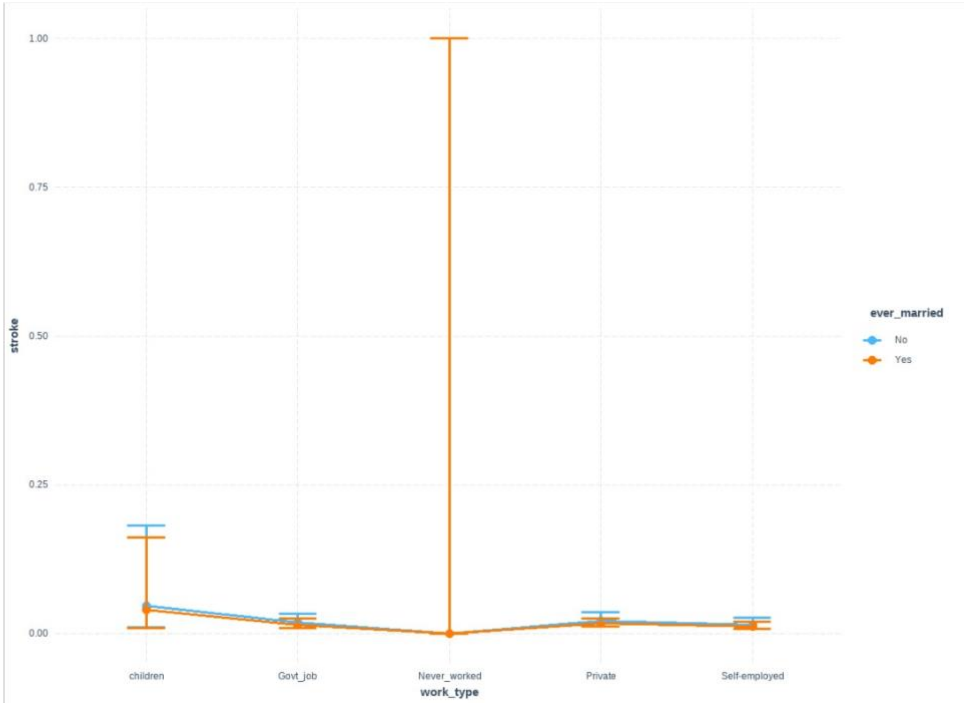
4. Heart Disease and Work Type



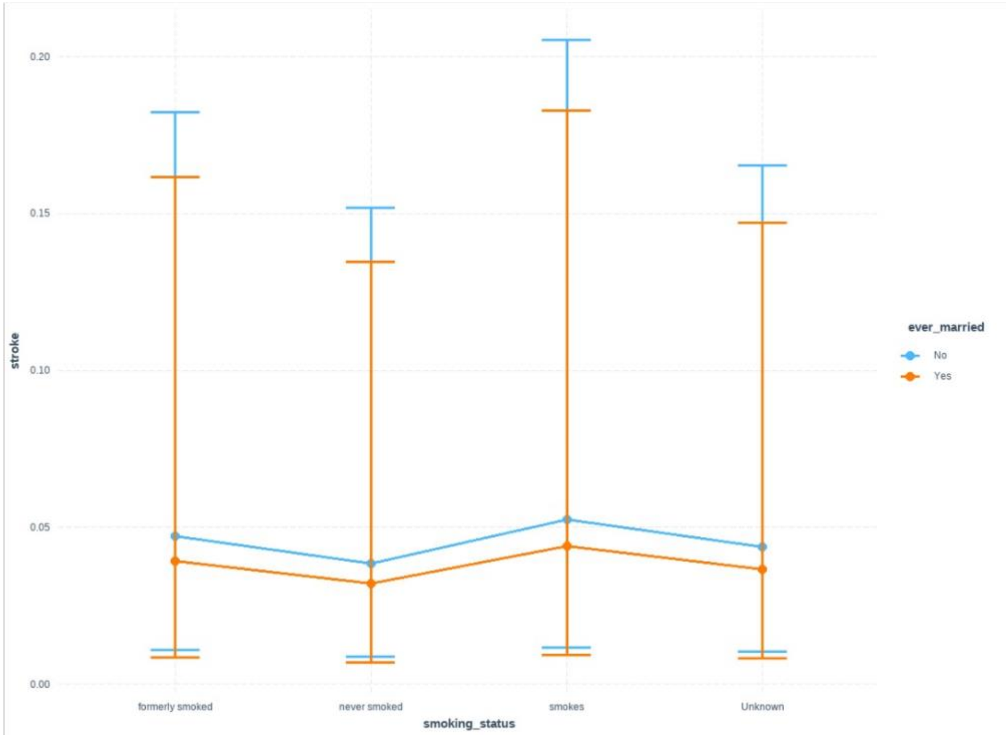
5. Heart Disease and Smoking Status



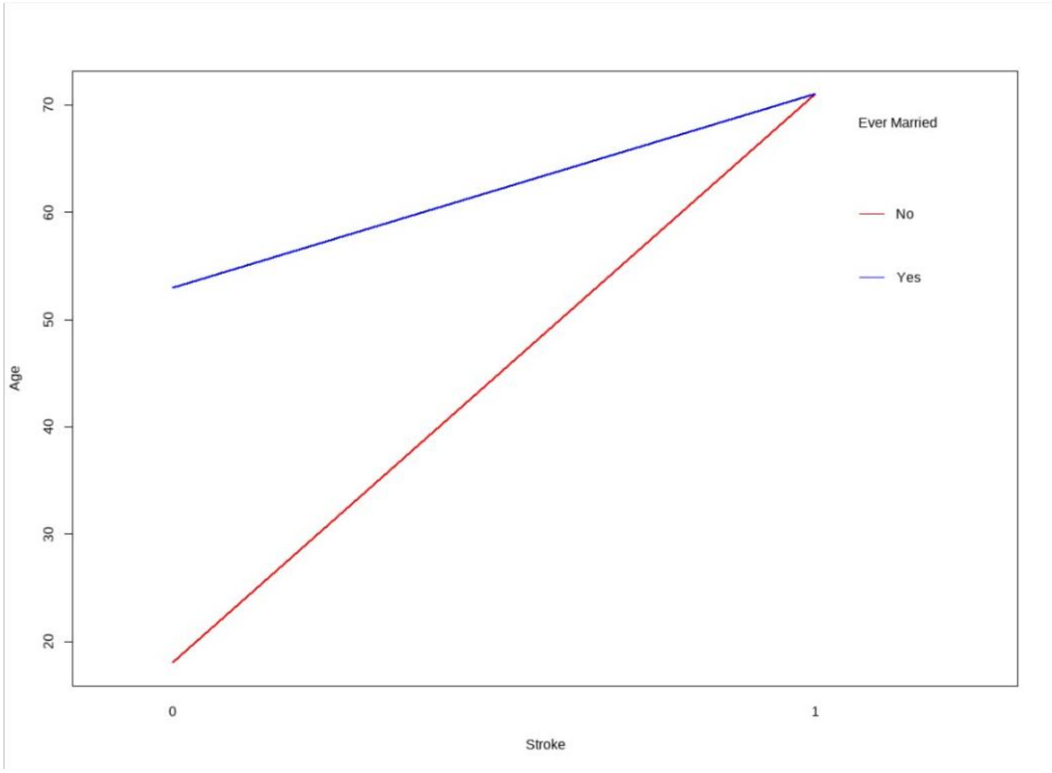
6. Ever Married and Work Type



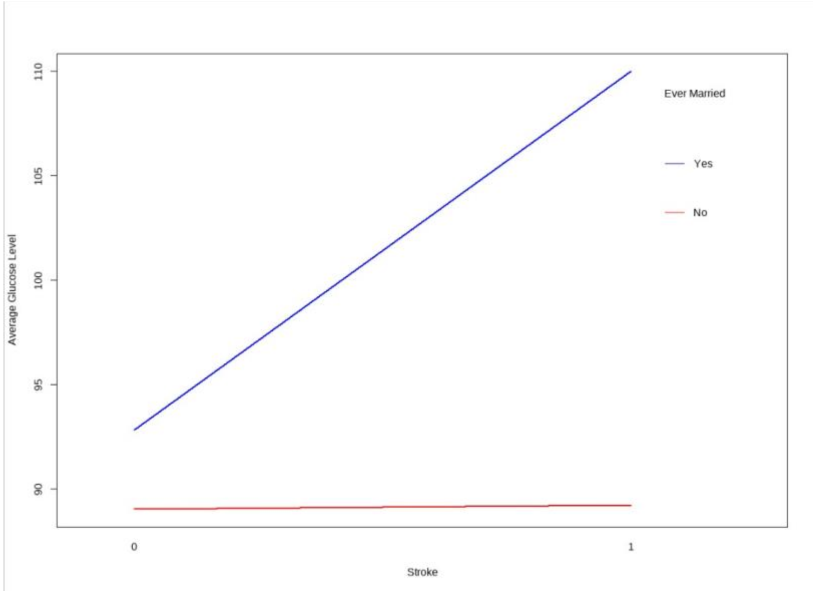
7. Ever Married and Smoking Status



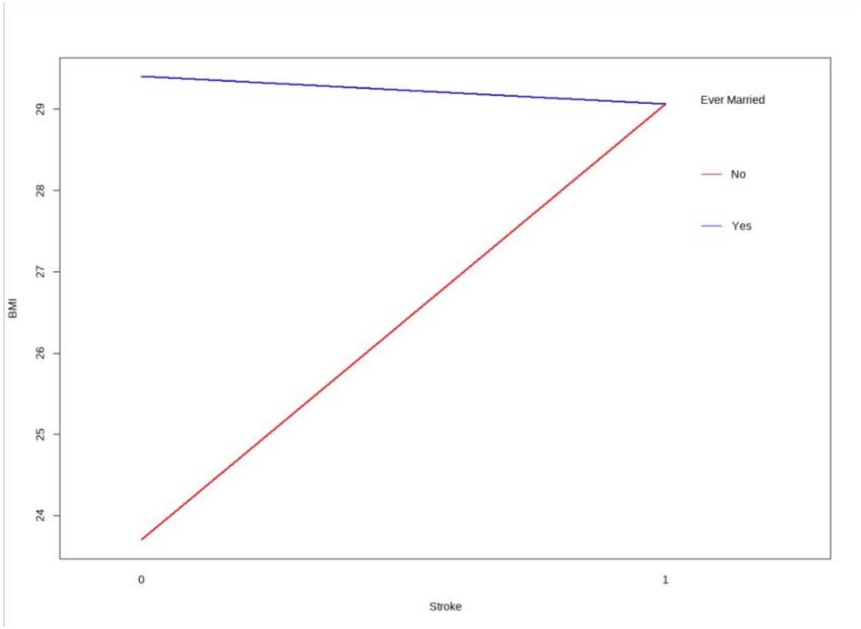
8. Ever Married and Age



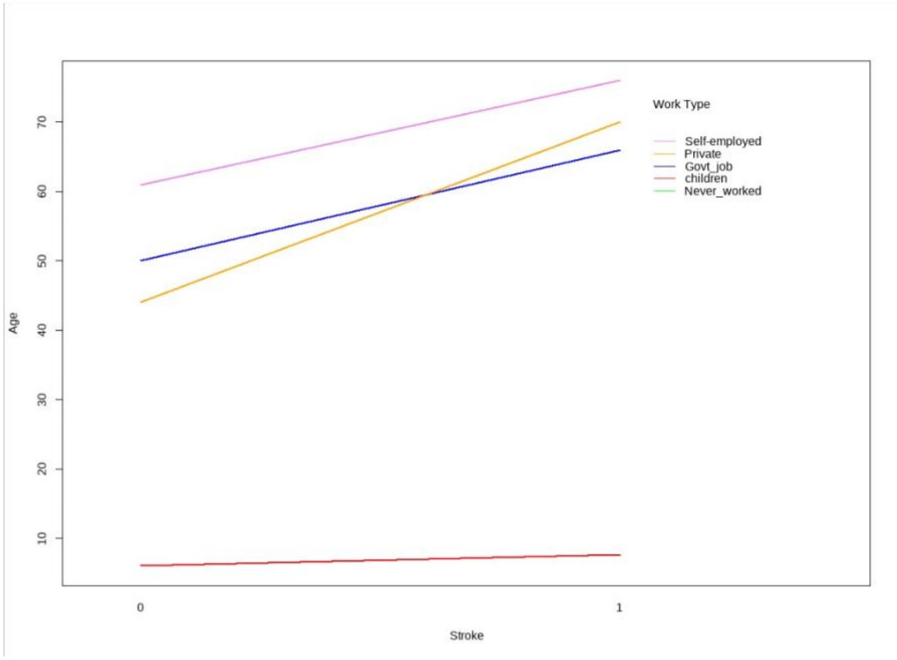
9. Ever Married and Avg Glucose Level



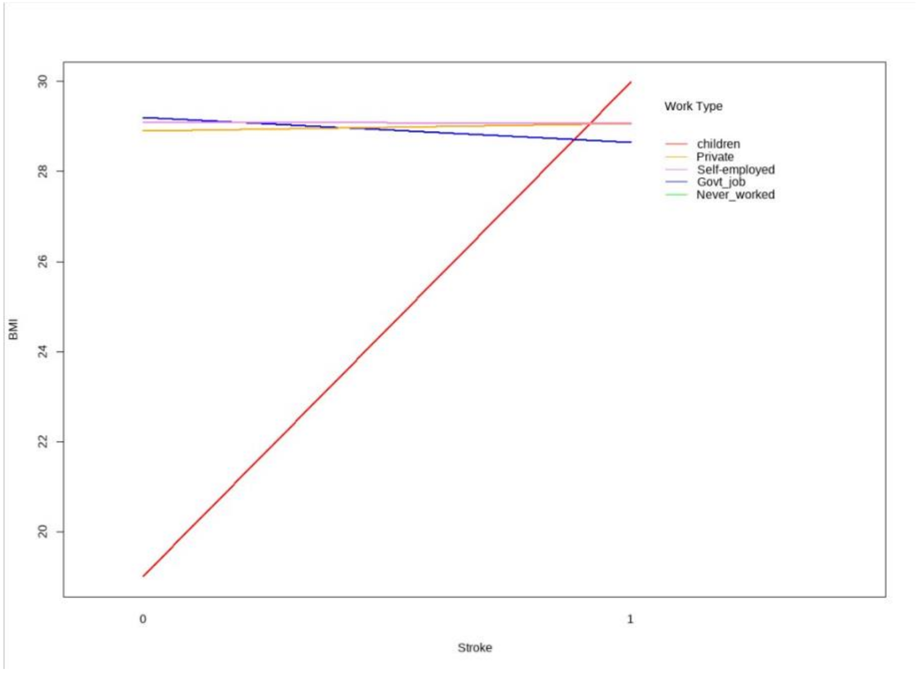
10. Ever Married and BMI



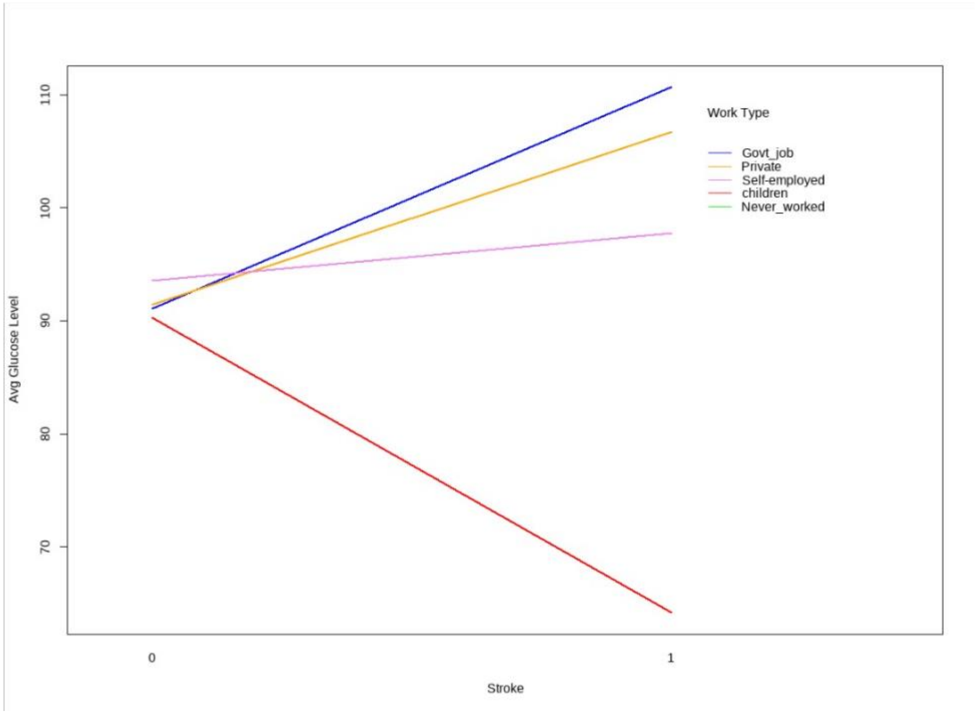
11. Work Type and Age



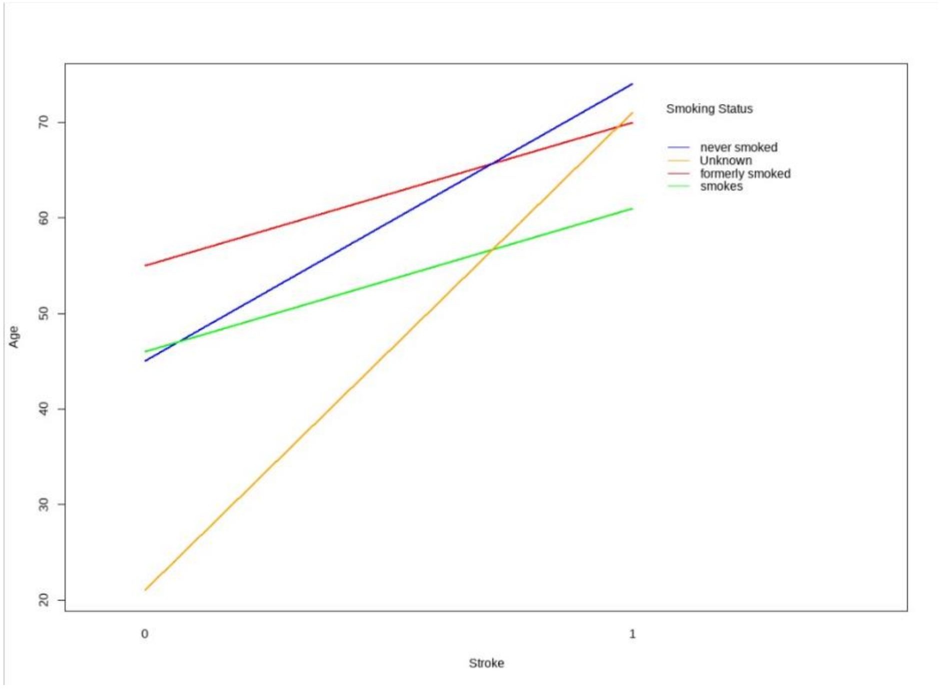
12. Work Type and BMI



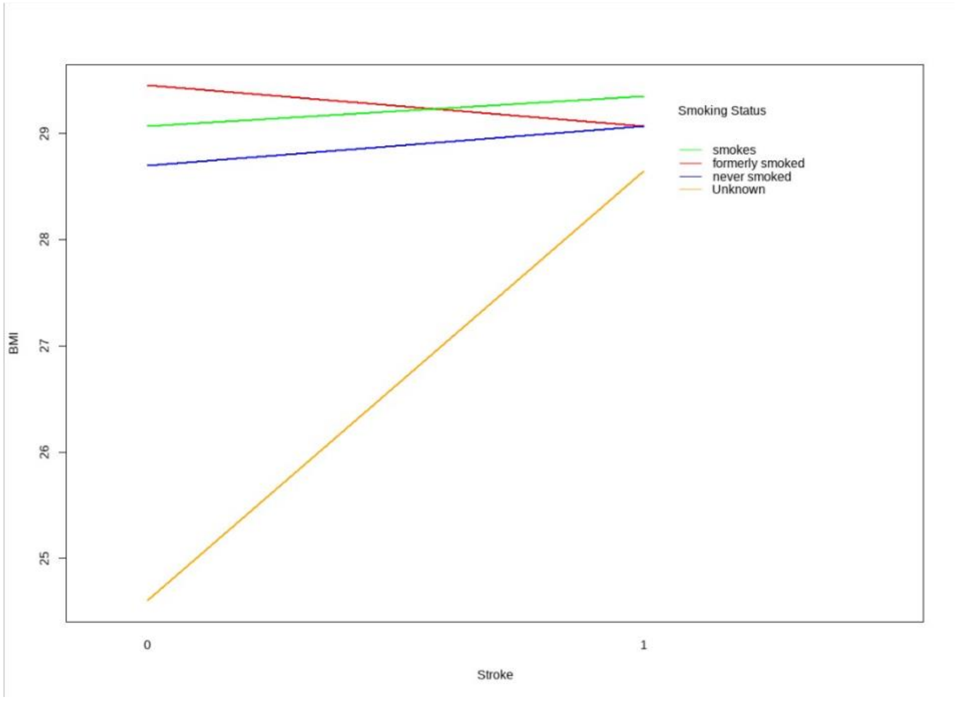
13. Work Type and Avg Glucose Level:



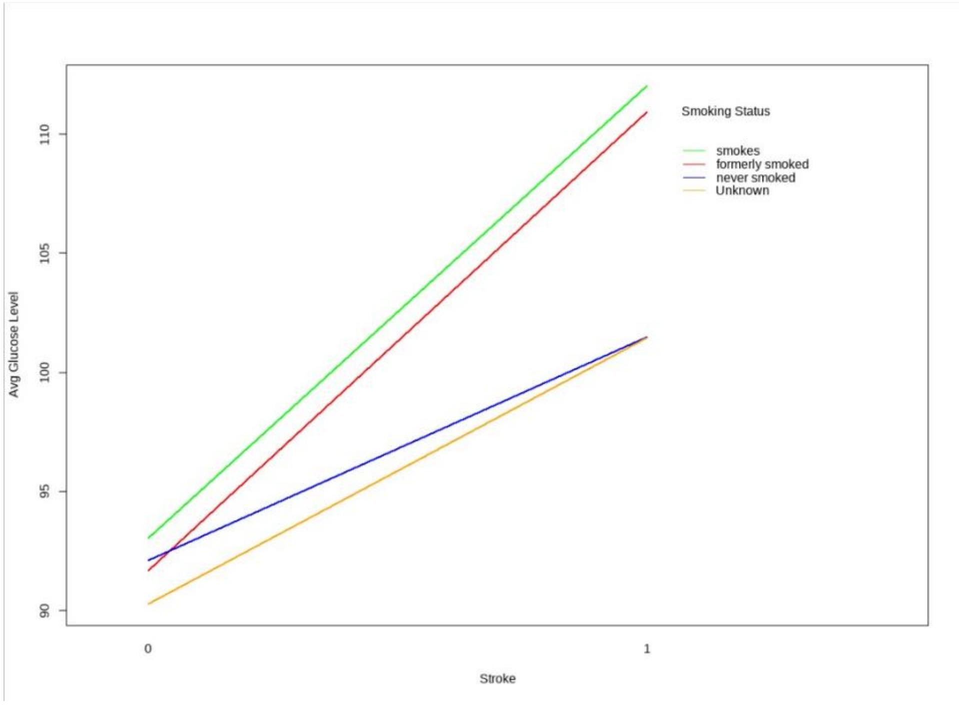
14. Smoking Status and Age



15. Smoking Status and BMI



16. Smoking Status and Avg Glucose Level


Smoking Status
smokes
formerly smoked
never smoked
Unknown
Avg Glucose Level
90
95
100
105
110
0
1
Stroke