Stroke Prediction Dataset Report

Jessica Xia

Vikasini Kuppa

Ashley Lopez

Sarah Armstrong

Department of Statistics, University of Riverside

STAT011: Introduction to Statistical Modeling

Professor Xu Cao

June 6, 2025

Introduction

Our group is analyzing the Stroke Prediction dataset, created by McKinsey & Company and displayed on Kaggle. Originating in Bangladesh, the data gathered in the study was used to predict the likelihood of a person experiencing a stroke. The data set presented contains over 5,110 patients, including their medical records and lifestyle indicators. Some components and variables that are worth investigating are the age of each patient, gender, hypertension status, and average glucose level. By analyzing these components and variables, we are able to improve the predictive accuracy of stroke prediction, allowing us to identify high-risk groups and create public health strategies to intervene and raise awareness.

Methodology

For our project, we set a seed for reproducibility to ensure consistent results throughout the analysis. During exploratory data analysis, we analyzed key variables using histograms, box plots, bar plots, and five-number summaries to understand their distributions and give insight into our research questions. For analysis #1, we chose to conduct a two-sample difference of two means test comparing the presence of strokes with BMI to answer the question of whether or not there is a significant difference in average BMI between individuals who have experienced a stroke and those who have not. For analysis #2, we wanted to know which factors among the available explanatory variables best predict BMI, and how they relate to BMI, so we conducted multiple linear regression with BMI as the response variable and conducted forward selection to select 3 explanatory variables. Lastly, for analysis #3, to know how age, hypertension, heart disease, smoking status, and average glucose levels influence the likelihood of having a stroke we conducted a logistic regression, with stroke as the response variable and age, hypertension, heart disease, smoking status, and average glucose levels as our explanatory variables. To

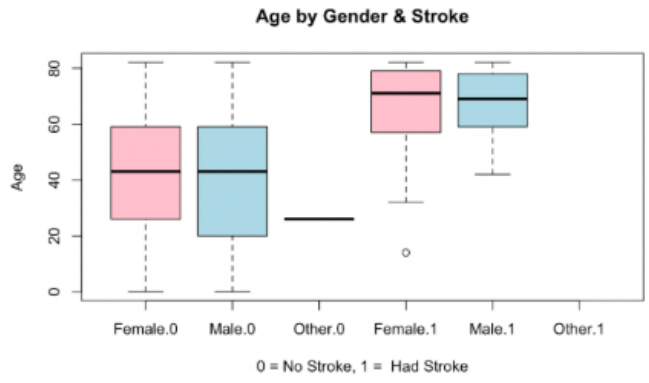
perform all of this, we used the standard R library. Some benefits of the methods we chose were the ability for us to conclude, we chose each one with the goal of trying to better understand relationships and variables. For our analysis choices, we also tried to avoid any repetitive tests to help find more unique relationships.

Exploratory Data Analysis

As part of the exploratory data analysis, our group used the "stroke.csv" file provided on Kaggle. To ensure a cleaner and more accurate data summary and ensure R functions run smoothly, we decided to omit all incomplete or unusable rows from your dataset that contain missing values. By doing so, approximately 201 observations were removed, leaving 4909 complete observations from the original 5110. The main variables we decided to look into intensively are the patient's age, their average glucose level, smoking status, whether they have hypertension, and their BMI.

Age

For the patient's age, we visually displayed it as a histogram, 5-number summary, and box plots in the context of gender and whether they had a stroke, which allowed us to put into context which age group and whether gender has a role in the predictability of having a stroke.



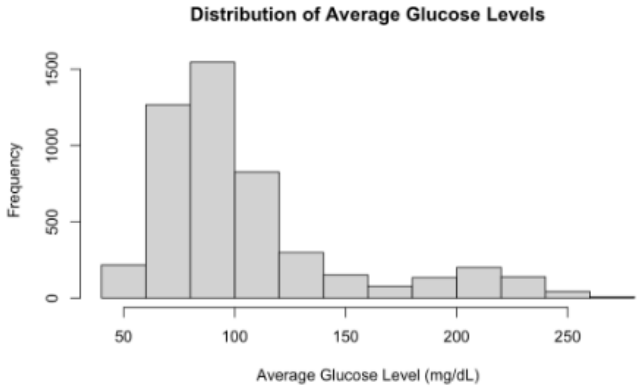

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.08   25.00   44.00   42.87   60.00   82.00
```

By converting the data into a histogram, it shows distribution of all patient ages is visually right-skewed, but statistically left-skewed, as the 5-number summary shows the mean is less than

the median. This can be explained by extremely low data, as seen by the range of age, the youngest patient being 0.08 years old, or approximately 1 month old, to the oldest being 82 years old. According to the 5-number summary, 25% of patients are 25 years or younger. 50% of patients are 44 years or younger. 75% of patients are 60 years or younger, while the top 25% have levels are 60+ years old. As seen in the boxplot shown above of age by gender and whether or not they had a stroke, we can reasonably assume that the risk of having a stroke increases with age, and females have a slightly higher chance of having a stroke compared to males.

Average Glucose Level

For the patients' average glucose level, similarly, we did a histogram and a 5-number summary.

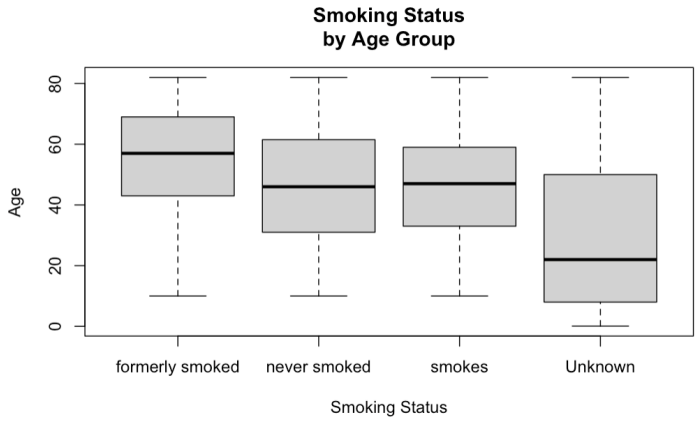



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  55.12   77.07   91.68  105.31  113.57  271.74
```

The distribution of the average glucose level of all patients is right-skewed, with extreme cases on the right suggesting possible pre-diabetic or diabetic conditions that may add as a factor to the predictability of having a stroke. For the 5-number summary, our data revealed that the patients' overall average glucose level ranges from 55.12 mg/dL to 271.74 mg/dL. 25% of patients have an average glucose level of 77.07 mg/dL or lower. 50% of patients have an average glucose level of 91.68 mg/dL or lower. 75% of patients have an average glucose level of 113.57 mg/dL or lower, while the top 25% have levels above 113.57 mg/dL.

Smoking Status

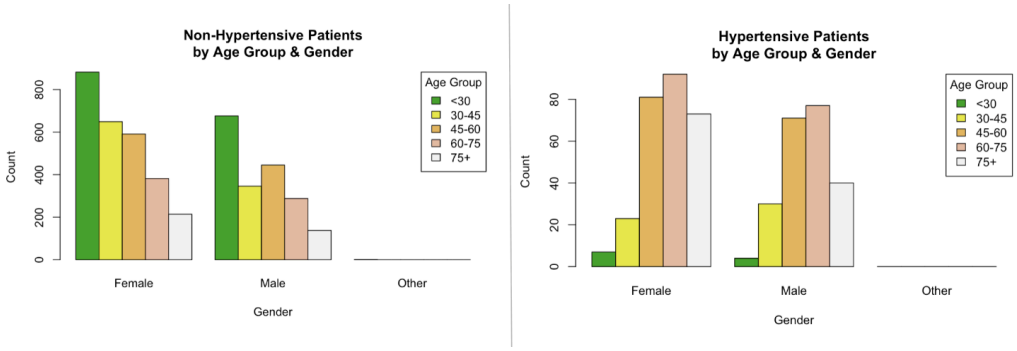
For the patients' smoking status in age relations, we decided to use a box plot and divide smoking status between age groups, starting from < 30, 30 to 45, 45 to 60, 60 to 75, and 75+ years old.



The median age of patients who formerly smoked is around 55 years, being the oldest among all smoking statuses. Patients who never smoked and those who currently smoke show similar age distributions, with a median age of around 45 years. The "Unknown" smoking status group has the widest age range. Analyzing this box plot, we can reasonably assume that patients who formerly smoked are related to old age, meaning it also has a relationship with the risk of having a stroke.

Hypertension

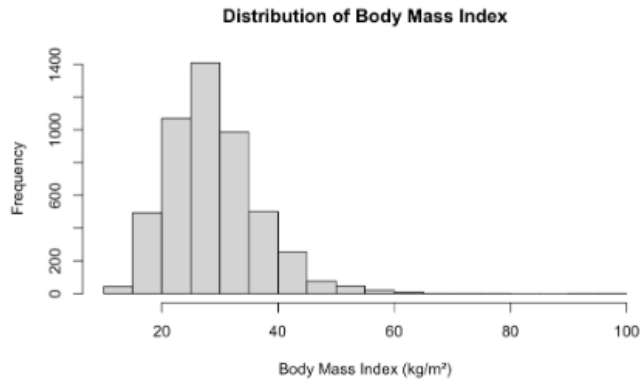
For the patients' hypertension, we decided to do bar graphs using the variable hypertension in relation to age and gender.



Females outnumber males across all age groups for both non-hypertensive patients and hypertensive patients. For non-hypertensive patients, the majority are younger than 30 years old. Contrary to hypertensive patients, the prevalence of hypertension increases with age, being more prominent in the 60 to 75 age group. Similarly with smoking status and age, it is reasonable to assume that hypertension plays a factor in whether in the probability of having a stroke.

Body Mass Index (BMI)

For the patients' BMI, we used a histogram and a 5-number summary.




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.30   23.50   28.10   28.89   33.10   97.60
```

The histogram is right-skewed, showing that the majority of patients' BMIs are between 20 to 40 kg/m², but there are a few extreme outliers with much higher BMI levels and pulling the tail of the distribution to the right. For the 5-number summary, our data revealed that the patients' overall BMI ranges from 10.30 kg/m² to 97.60 kg/m². 25% of patients have an average glucose level of 23.50 kg/m² or lower. 50% of patients have a BMI of 33.10 kg/m² or lower. 75% of patients have a BMI of 33.10 kg/m² or lower, while the top 25% have levels above 97.60 kg/m².

Here is a glimpse of the first 10 rows of the data, after omitting the NA values.

	id <int>	gender <chr>	age <dbl>	hypertension <int>	heart_disease <int>	ever_married <chr>	work_type <chr>	Residence_type <chr>	avg_glucose_level <dbl>
1	9046	Male	67	0	1	Yes	Private	Urban	228.69
3	31112	Male	80	0	1	Yes	Private	Rural	105.92
4	60182	Female	49	0	0	Yes	Private	Urban	171.23
5	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12
6	56669	Male	81	0	0	Yes	Private	Urban	186.21
7	53882	Male	74	1	1	Yes	Private	Rural	70.09
8	10434	Female	69	0	0	No	Private	Urban	94.39
10	60491	Female	78	0	0	Yes	Private	Urban	58.57
11	12109	Female	81	1	0	Yes	Private	Rural	80.43
12	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46

Analysis #1 - Two-Sample Difference Of Two Means

We conducted a Welch t-test to answer the question: Is there a significant difference in the mean BMI between individuals who have experienced a stroke and those who have not? Our null hypothesis (H_0) was that there is no significant difference in the BMI of those who do and do not have a stroke. Our alternative hypothesis (H_a) was that there is a significant difference in BMI of those who do and do not have a stroke. Our test gave us the results below.

```
        Welch Two Sample t-test

data:  bmi by stroke
t = -3.6404, df = 237.83, p-value = 0.000334
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -2.5401580 -0.7562981
sample estimates:
mean in group 0 mean in group 1
       28.82306        30.47129
```

The results indicate that the mean BMI for the group with no stroke was 28.82306 and the mean BMI for the group with stroke was 30.47129. The confidence interval from -2.54 to -0.756 suggests that the average BMI for the stroke group was significantly greater by 0.756 to 2.54 units. The p-value of 0.000334 is less than the alpha level of 0.05 at a 95% confidence interval, thus, we reject the null hypothesis. The t-value of -3.6404, combined with a very small p-value, indicates a statistically significant difference in BMI between the two groups, with Group 0 (no stroke) having a lower mean BMI than Group 1 (stroke). We have support that there is a significant difference in the BMI of those who do and do not have a stroke. Thus, BMI can be used as a significant predictor for stroke.

Analysis #2: Multiple Linear Regression

In this analysis, we explore the relationship between BMI and other variables such as gender, age, hypertension, heart disease, marital status, average glucose level, and smoking

status. Given the significant difference in BMI between people who had or did not have a stroke, it is of interest to see how BMI relates to other variables. In particular, we use multiple linear regression to predict BMI based on explanatory variables selected using forward selection.

Forward Selection Results

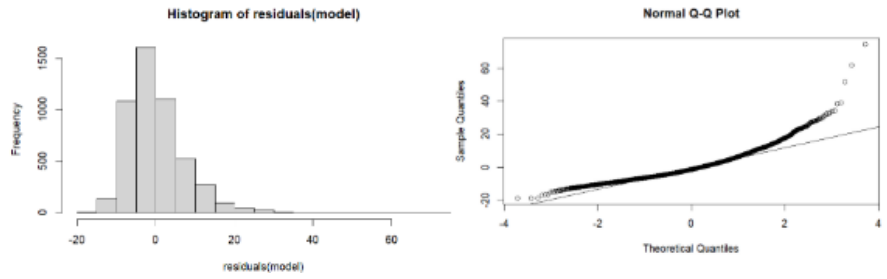
Model	Adjusted R^2
gender	0.0004164
age	0.111
hypertension	0.02796
heart_disease	0.001507
ever_married	0.1166
avg_glucose_level	0.0306
smoking_status	0.07383

Model	Adjusted R^2
gender	0.1164
age	0.1353
hypertension	0.1294
heart_disease	0.1164
avg_glucose_level	0.1321
smoking_status	0.1433

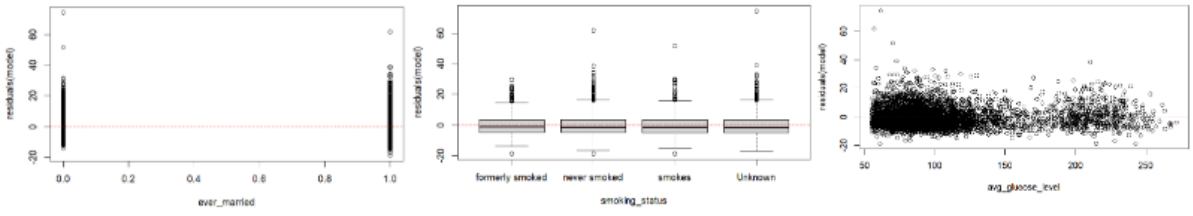
Model	Adjusted R^2
gender	0.1431
age	0.1537
hypertension	0.1529
heart_disease	0.1431
avg_glucose_level	0.1566

Forward selection procures variables for marital status, smoking status, and average glucose level. After fitting the linear regression model with these variables, diagnostics are checked.

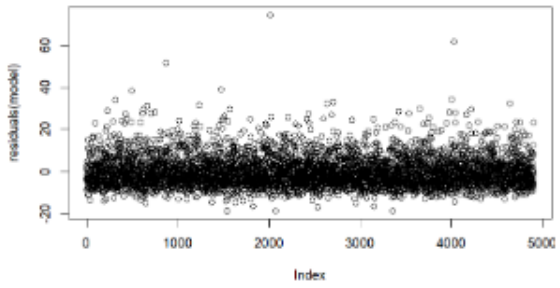
Nearly normal residuals



Constant variability in residuals



Independent residuals



The histogram and normal Q-Q plot show a slight right skew, but overall the distribution is nearly normal. The plots for residuals for each explanatory variable show similar residual distributions near the y = 0 line. The plot for all residuals is absent of fan or curve shapes. Therefore, all conditions are satisfied.

Model Summary

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 24.72202    0.39691  62.286   <2e-16
ever_married                 4.39375    0.23248  18.900   <2e-16
smoking_statusnever smoked -0.12199    0.30184  -0.404    0.686
smoking_statussmokes        0.20542    0.36486   0.563    0.573
smoking_statusUnknown      -2.89673    0.32790  -8.834   <2e-16
avg_glucose_level           0.02083    0.00235   8.866   <2e-16
```

Regression line: BMI = 24.72202 + 4.39375(ever_married) - 0.12199(never_smoked) + 0.20542(smokes) - 2.89673(unknown) + 0.02083(avg_glucose_level)

The intercept of 24.72202 is the predicted BMI for a person who is not married, formerly smoked, and has a glucose level of 0. While keeping other variables constant, the slope for ever_married means a married person's BMI is 4.39375 higher than a single person. Slopes for smoking status show how BMI increases or decreases based on whether a person has never smoked, currently smokes, or if they have an unknown smoking status. The slope for avg_glucose_level means BMI increases by 0.02083 for every one-unit increase in average glucose level.

Significant variables with p-value < 0.05: ever_married, smoking_status(Unknown), and avg_glucose_level. These variables are strongly associated with BMI.

Predictions of 5 New Observations

	ever_married	smoking_status	avg_glucose_level	predicted BMI
1	1	smokes	102	31.44612
2	0	never smoked	73	26.12082
3	0	formerly smoked	95	26.70113
4	1	unknown	60	27.46900
5	1	never smoked	49	30.01458

We see that married individuals generally had a higher predicted BMI than those who were not married. Higher average glucose levels also correspond to higher predicted BMI unless offset by a favorable smoking status, such as never smoked or unknown.

Analysis #3: Logistic Regression

We used Logistic Regression to predict the probability of a stroke considering age, hypertension, heart disease, smoking status, and average glucose levels. This helped us answer our research question of what medical conditions and habits contribute to a greater risk of having a stroke. Looking at the summary, it can be found that age ($p < 2e\text{-}16 < .05$), hypertension ($p = 0.0155 < 0.05(\text{alpha})$), and average glucose level ($0.00039 < 0.05$) all are statistically significant and can be associated with an increased risk of stroke. Whereas, heart disease ($p = 0.112 > 0.05$) and smoking status ($p > 0.05$) seem not to be statistically significant. This tells us that heart disease and smoking status seem to be less involved in increasing someone's risk for a stroke as opposed to age, hypertension, and average glucose levels. When holding all other variables constant, we found that those with hypertension were 48.6% more likely to have a stroke, for each one-year increase in age, the odds of having a stroke increased by 7.2%, and for a one unit increase in glucose levels odds increased by 0.4%. We also evaluated five new observations, and our highest predicted probability was 52.07%, which was 82 year old with all risk factors. Our lowest predicted probability of our five observations was 1.25%, which was a 30-year-old with hypertension, no heart disease, smokes, and slightly elevated glucose. This demonstrates that

there is a statistically significant association between strokes and our three statistically significant variables.

Conclusion

Overall, through our analyses, we were able to determine which variables impacted stroke significantly. From analysis 1, we concluded that there is a significant difference in the BMI of people who do and do not experience a stroke. From analysis 2, we concluded that ever_married, smoking_status(unknown), and avg_glucose_level were the most significant predictors of BMI. From analysis 3, we concluded that every unit increase in hypertension made individuals 48.6% more likely to have a stroke, per unit increase in age made individuals 7.6% more likely to have a stroke, and per unit increase in glucose made individuals 0.4% more likely to have a stroke. These variables provide valuable insight into stroke predictability. This is useful for healthcare workers and drug developers as analyzing relationships and trends can aid in early detection, as well as prevention, of stroke. Oftentimes, stroke can lead to death, making the prevention of stroke even more valuable towards a patient's health.

Though our dataset does encounter some limitations, given that it was taken from Bangladesh, this can constrain our ability to generalize our findings to all populations. In future work, we should consider incorporating more variables into our analyses, as well as incorporating factors such as socioeconomic and environmental status. Overall, our data gave a glimpse into stroke research using statistics and showed us the important applications of finding significant causes of stroke. Through increased research efforts, this data can be applied to real-world issues of stroke and contribute to saving people from stroke.

References

Hassan, A., Ahmad, S. G., Munir, E. U., Khan, I. A., & Ramzan, N. (2024). Predictive modelling and identification of key risk factors for stroke using machine learning. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-61665-4

Stroke Prediction Dataset. (2021, January 26). Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data