Stroke Data Analysis

Nathan Gin

67117388

Stats 170A HW 3

Abstract

Being able to determine whether or not someone had a stroke before they are brought to a hospital allows the doctors to set up and provide the adequate care that the subjects need. There are multiple methods used to help determine whether someone had a stroke or not such as RACE score or Electroencephalography changes. Through analysis of the stroke dataset, we were able to conclude that there is statistically significant evidence that a higher RACE score is associated with individuals having a stroke given the clinical variables of gender, age, and LKW. Additionally, we have evidence that suggests that the association between RACE and stroke does not vary by age.

Introduction

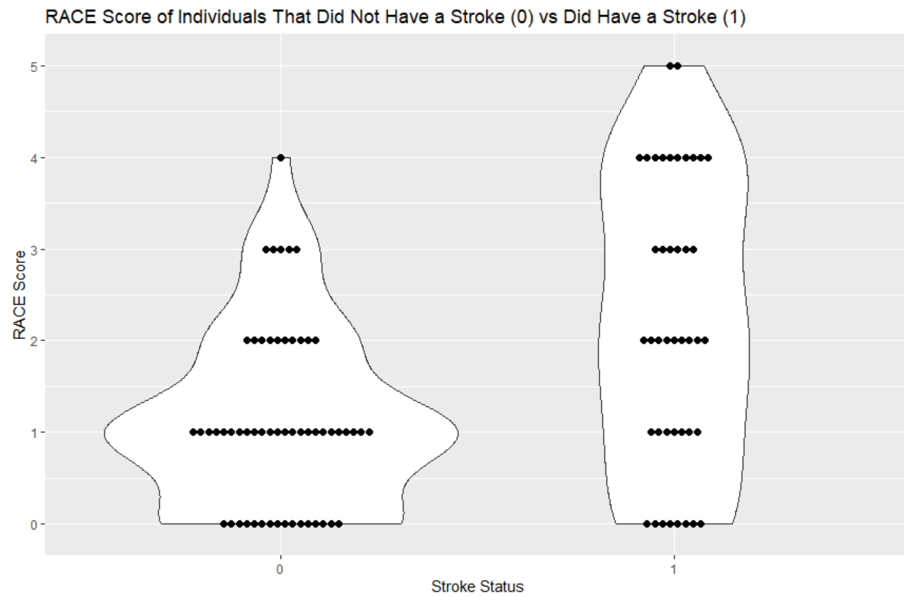
The motivation for this study is to help with a pre-hospital diagnosis of strokes in patients so they can receive adequate and appropriate care upon arrival. The data I am looking at consists of 100 patients that were taken into the hospital for care and diagnosis of if they had a stroke or not. I am looking at the association between RACE (Rapid Arterial Occlusion Evaluation), a tool used to help predict if a patient had a stroke before being admitted to the hospital on a score from 1 to 5, and stroke status. Additional clinical variables age (in years), LKW (Last Known Well in hours), and gender (male or female) are given. Additionally, I am testing the hypothesis that there is an association between RACE score and Stroke status given age while summarizing and visualizing the data given.

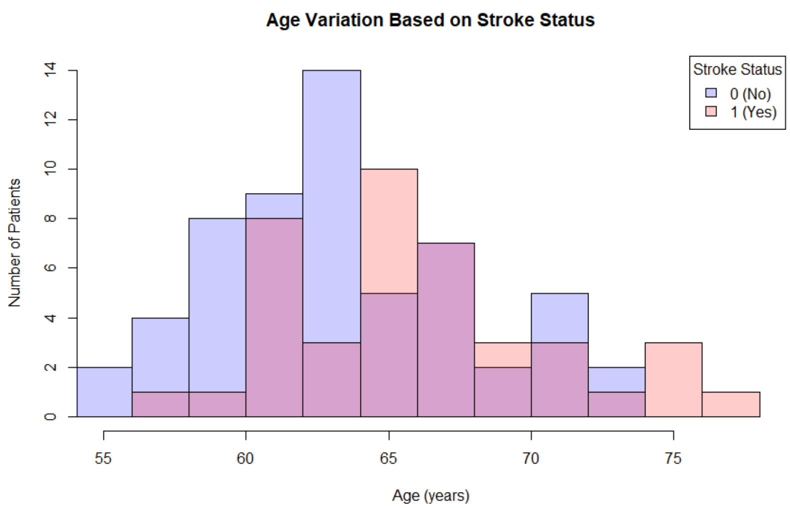
Statistical Methods

Initially, I began to analyze our data by exploring the variables. Out of the 100 patients, 52 were male and 48 were female. The table below shows a summary of the other variables involved in the study while the graphs help to visualize some of the questions and data we are looking at.

	Minimum	Median	Maximum	Mean	Missing Values
LKW (hours)	2	11	23	12.2	3
RACE score	0	1	5	1.592	2
Age (years)	55	64	77	64.91	1

The violin plot shows that a higher RACE score is connected to those that had a stroke versus those that did not have a stroke. Those that did not have a stroke scored a lower RACE score, with an average of 1.125, while those that did suffer from a stroke scored a higher race score, with an average of 2.214.





The histogram shows the relationship between age and stroke status. As you can see, those that suffered from a stroke have slightly higher ages with an average of 66.46 years old versus those that did not suffer from a stroke who averaged 63.81 years old.

To answer the questions at hand, I fit the data with logistic regression models within R (See Appendix for Code). The first logistic regression model was fit with stroke status as the response variable with RACE score, age, gender, and LKW as the explanatory variables with no interaction terms between them. The second logistic regression model was the same except this time with an added interaction term between age and RACE score. Logistic regression is the most fitting for this data because we are testing it against stroke status which can only be yes or no in this case. Using these models we can see their influence on stroke status based on the p-values and the odds ratios. The p-value represents if there is a relationship between the clinical variables and stroke status. If the p-value is very small then it gives stronger evidence that there is a relationship between the variables. The odds ratio represents the relationship between the variables. If it is equal to 1 then it shows there is no association between the two variables.

Below we can see how I set up the logistic regression models with each Xi variable representing RACE score, age, gender, and LKW respectively. Each βi represents the coefficient of the variable in the logistic regression model and β0 represents the intercept. The calculation for the odds ratio is also given.

Logistic Regression:

$$log(\frac{P(Outcome)}{1 - P(Outcome)}) = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + ...$$

Odds Ratio:

$$\frac{P(Outcome)}{1 - P(Outcome)} = exp(\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + ...)$$

Results

Association Between RACE and Stroke Given All Other Clinical Variables

$$log(\frac{P(Stroke = 1)}{1 - P(Stroke = 1)}) = \beta 0 + \beta 1 * RACE + \beta 2 * LKW + \beta 3 * Male + \beta 4 * Age$$

	Coefficient Estimate	Std. Error	Z value	Pr(>|Z|)	Odds Ratio	95% Odds Ratio Confidence Interval
RACE	0.7045	0.1977	-2.521	0.00036	2.0230	[1.4018, 3.0644]
LKW	-0.0282	0.0396	3.564	0.4771	0.9721	[0.8974, 1.0501]
Male	-0.3095	0.4999	-0.619	0.5358	0.7338	[0.2702, 1.9489]
Age	0.1231	0.0541	2.277	0.0228	1.1310	[1.0211, 1.2650]

Answering the question of if there is an association between RACE and stroke status given all other clinical variables, we can see that the p-value of RACE is 0.00036 and the odds ratio is 2.023. From this, we can conclude that the odds of having a stroke for someone who had a RACE score over 2.5 is 2.023 times greater than those with a RACE score under it. Additionally, age has an odds ratio of 1.131 meaning that the older subjects were slightly more likely to have a stroke. The p-value for both RACE and age is less than 0.05 meaning that they are statistically significant. LKW and gender have very large p-values meaning that they are not statistically significant. Looking at the odds ratio confidence intervals, for both RACE and age the lower bound is greater than 1 meaning that a higher RACE score and older age both increase the risk of having a stroke.

Association Between Race and Stroke if it Varies by Age

$$log(\frac{P(Stroke = 1)}{1 - P(Stroke = 1)}) = \beta 0 + \beta 1 * RACE + \beta 2 * LKW + \beta 3 * Male + \beta 4 * Age + \beta 5 * Age * RACE$$

	Coefficient Estimate	Std. Error	Z value	Pr(>|Z|)	Odds Ratio	95% Odds Ratio Confidence Interval
RACE	-3.5287	3.4279	-1.029	0.303	0.02934	[0.00002, 13.888]
LKW	-0.0243	0.0400	-0.607	0.544	0.9760	[0.9005, 1.0549]
Male	-0.2524	0.5068	-0.498	0.618	0.7769	[0.28277, 2.0956]
Age	0.0240	0.0936	0.256	0.798	1.0242	[0.8489, 1.2312]
Age*RACE	0.0661	0.0538	1.229	0.219	1.0683	[0.9710, 1.2010]

Adding in the interaction term between age and RACE in the logistic regression model make all of the p-values very large meaning that none of them are statistically significant when including the interaction term between. This means that with an interaction term, we cannot conclude any of the clinical variable's relationship with stroke status. This means that the association between RACE and stroke does not vary by age.

Discussion

In conclusion, there is an association between an increased RACE score and an increased chance of having a stroke from the data. However, this relationship does not vary based on the age of the patients.

The data set we are given is rather thorough but there still are some null or missing values for LKW, RACE, and Age for certain patients. Additionally, there are only data for 100 individuals in this data set which is limiting to being certain about any conclusions we make. It would be hard to apply this to the population as a whole without knowing more about where the data came from and how it was taken as we were given limited information on the data itself. The 100 EEG (Electroencephalography) signals taken from each individual were also not used in this study which can help to determine whether or not one of them suffered from a stroke or not.