Report

2022-11-21

# Introduction

Response Variable:

* stroke:

Possible Predictors:

* smoking\_status: The smoking status of the observation. Factor variable with 3 levels: “formerly smoked”, “never smoked”, and “smokes”.

## Cleaning the data

To use this data set, we first converted variables to their correct representations and then omitted any incomplete observations.

### Notable fixes:

* The categorical variable smoking\_status was reformatted to remove the level “Unknown”, as it was used to represent unavailable data. Any cells that previously had “Unknown” were updated to reflect their unavailable status.
* The quantitative variable bmi was fixed by converting the data type from character to numeric. This change does not apply to non-numeric cells, so the cells containing “N/A” were changed to reflect their unavailable status in the now numeric column.

—**still need to list the rest of the variables, short description of data, and the question we want to answer**

# Logistic Regression Model

We chose to use a logistic regression model for our data because our response variable is qualitative with two classes. It was desirable to have a model predict the probability of a person having a stroke using binary classification. If linear regression is used to predict the probability of whether or not a person will have a stroke, then the model may have predicted Y values outside of our intended range of 0-1. The logistic regression model allows us to enforce this restriction of range.

## Model Formula

explain model choice  
write model equation  
test/train models  
test error rate  
interpretation of full model in regards to our question: which factors demonstrate statistical significance in relation to having a stroke?

# Conclusion