MET CS 555 Term Project

Hicran Arnold

2/27/2022

ABSTRACT

Many of us are familiar with heart disease and stroke.According to World Health Organization stroke is the second leading cause of death.The stroke data set provides valuable information to test to see the relationship between the input parameters like gender, age, various diseases, and smoking status and predict if a patient is likely to get a stroke based on the input parameters. In my research, I will focus on the questions below; ;
- Q1) Do males have a higher risk of having a stroke than females?
- Q2) Is there any association in between age and bmi level on having a stroke?

INTRODUCTION

Information About the Data Set:

Data Set Name : Stroke Prediction Dataset (11 clinical features for predicting stroke events)
Link to the data source : Kaggle
Data Se Posted By : fedesoriano
Acknowledgements: (Confidential Source) - Use only for educational purposes

Attribute Information:
1. id: unique identifier
2. gender: "Male", "Female"
3. age: age of the patient
4.hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6. ever_married: "No" or "Yes"
7. work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8. Residence_type: "Rural" or "Urban"
9. avg_glucose_level: average glucose level in blood
10. bmi: body mass index
11. smoking_status: "formerly smoked", "never smoked", "smokes"
12. stroke: 1 if the patient had a stroke or 0 if not

Data Cleaning Process: After reviewing the original data set, I found out that some of the data points did not have BMI amount (entered as N/A), and same for the smoking status (entered as unknown). I decided to remove those data points since the original data was large. I sampled 50 to 50 from male and female population because it helped me to visualize and understand the data better for each gender.After cleaning the data I randomly sampled 1000 data points from the original data.

TESTING MALES HAVE A HIGHER RISK OF HAVING A STROKE THAN FEMALES

Many of us think that stroke is usually something that males have to deal with. As a woman, I want to check to see if it is true. To investigate this question I will use the Sample Test of Proportions(z-test.)

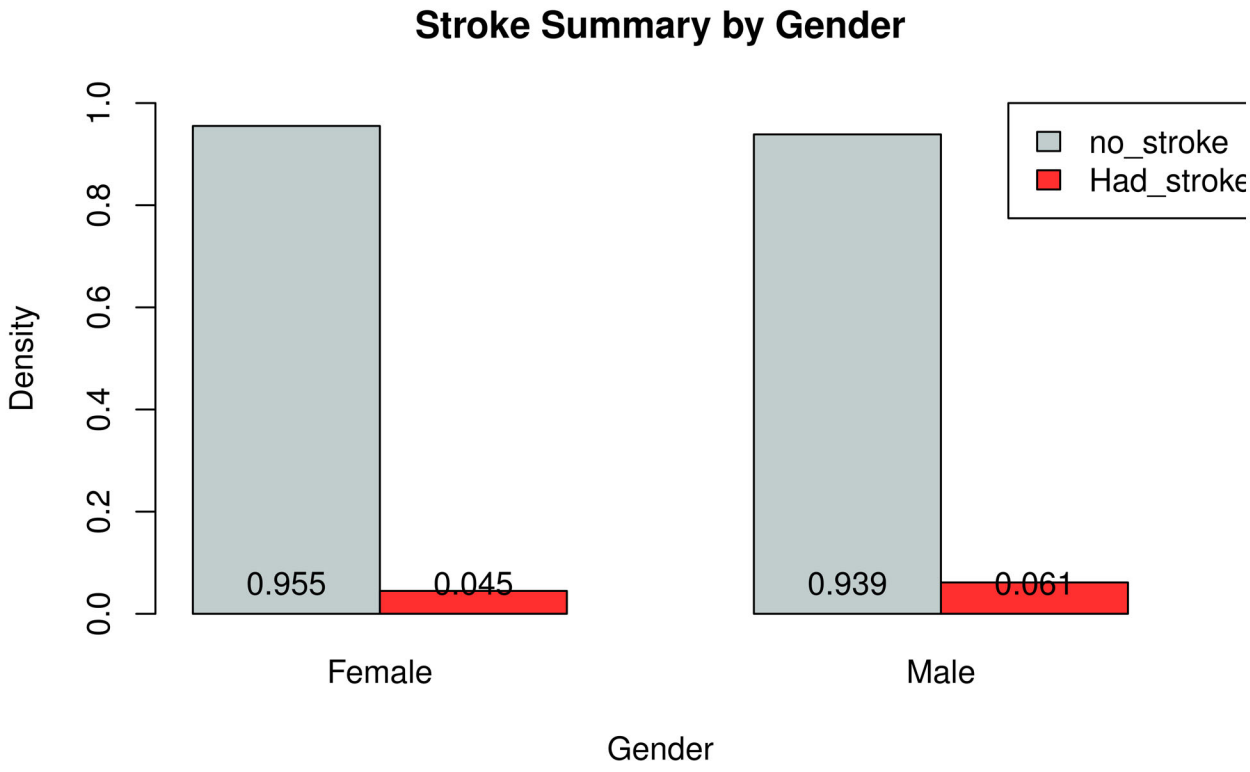
Success : Stroke count (1)
Failure : No stroke (0)

In the table below we see that: 6.14% of men had a stroke and of 4.48% women had a stroke. Male stroke proportion is slightly higher than women.

Table 1: Numerical Summary of Stroke Event by Gender

population_description	Stroke_1	Stroke_0	sampleSize	Sample.proportion
Male	23	352	375	0.0613333
Female	28	597	625	0.0448000

Here is the graphical summary. As we see that the difference is very small.



The estimate of risk difference will help us to quantify the difference between our two

groups(male and female).Also we can use estimate of risk ratio calculation to see if there is a high risk or low risk or there is no risk difference at all. Here is the calculation of the estimated risk difference;

```
RD <-  maleprop - femaleprop
RD
```

```
## [1] 1.843974
```

```
RR <- maleprop/femaleprop
RR
```

```
## [1] 1.393162
```

As we see in the above calculation risk of stroke is 1.85% higher among males as among females.The estimate risk ratio is 1.39. So it means that the risk of stroke event is 1.39 times higher among makes as among females.

test for the global null hypothesis tests the following hypotheses

- H0:p1 = p2 (The underlying population proportion of males who had a stroke is the same as the population proportion of females who had a stroke)
- H1 : p1 != p2(The underlying population proportion of males who had a stroke is is different than the proportion of females who had a stroke)
- alpha = 0.05

We will use the below test statistic

z= ($\hat{p}$1 - $\hat{p}$2)/ sqr($\hat{p}$(1-$\hat{p}$)*((1/n1) + (1/n2))

Decision rule

- Determine the appropriate critical value from the standard normal distribution table associated with a right hand tail probability of alpha = 0.05/2=0.025. Using the Standard Normal Probabilities table, the appropriate critical value is 1.959964
- Decision Rule: Reject H0 if |z|> 1.959964
- Otherwise , do not reject H0

Computing the test statistic and the associated p-value

We can use R to calculate the test statistics

```
#two sample tests for proportions

proptestForGender <- proptestForGender <- prop.test(c(23,28),c(375,625),
                                                    alternative="two.sided",
                                                    conf.level = 0.95, correct = FALSE)
proptestForGender
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(23, 28) out of c(375, 625)
## X-squared = 1.3237, df = 1, p-value = 0.2499
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01266900  0.04573567
## sample estimates:
##     prop 1     prop 2
## 0.06133333 0.04480000
```

```
zval <- sqrt(proptestForGender[["statistic"]][["X-squared"]])
zval
```

```
## [1] 1.15053
```

```
pval <- proptestForGender[["p.value"]]

# population proportion
p <- ((28+23)/(625+375))
```

z is equal to |- 1.1505298|

Conculution

- We are fail to Reject H0 since |1.1505298| is not $>=$ 1.959964 We do not evidence to support that at the alpha =0.05 level that the underlying population proportion of males who had a stroke is different than the population proportion females who had a stroke (Here p value = (0.2499257)) and it is not smaller than alpha level 0.05. Although there is a small risk difference in between males and females we do not have enough evidence to support that it is significant.

TESTING IF THERE IS AN ASSOCIATION IN-BETWEEN AGE AND BMI LEVEL ON HAVING A STROKE?

Age and BMI are the two well known factors in having a stroke event but I want to check to see if that is true. Are they really associated with a having a stroke event?

To investigate this question we will use Chi-squared test and multiple logistic regression model.

Age Summary_

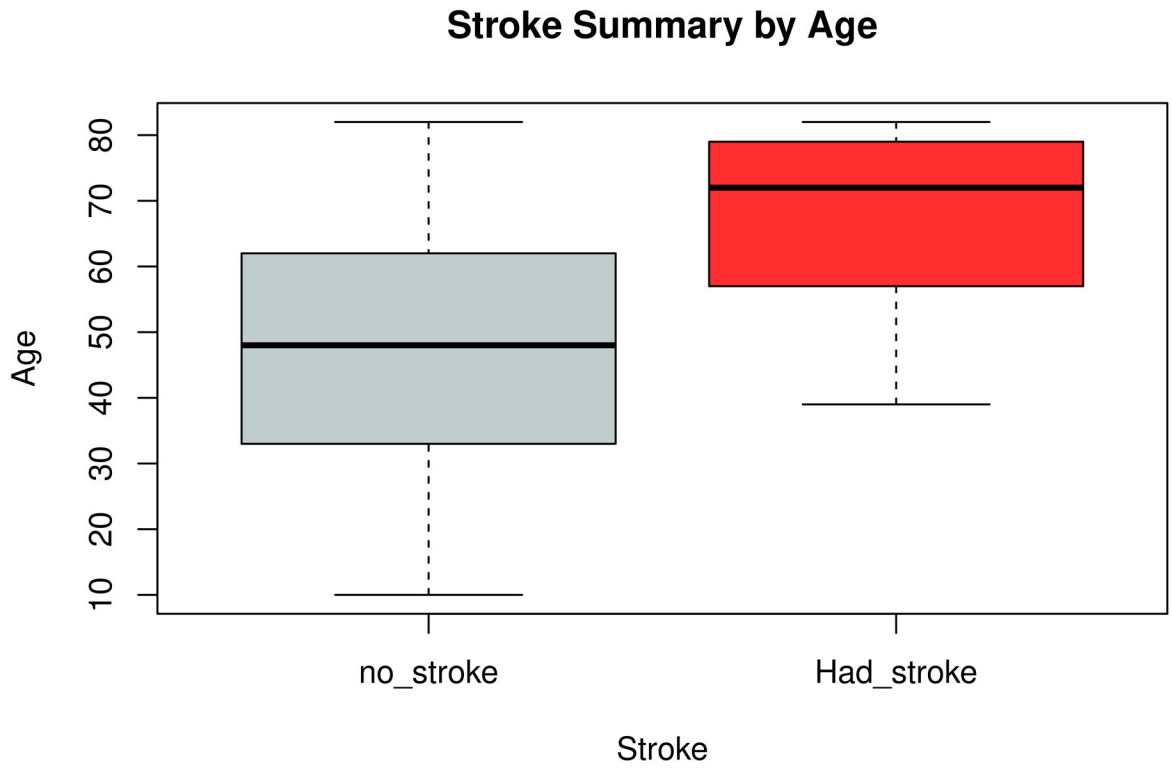
We can summarize the age group that who had stroke

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    39.0    57.0    72.0    68.1    79.0    82.0
```

We can summarize the age group that who did not have stroke

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   33.00   48.00   47.71   62.00   82.00
```

Here is the graph that summarizes our age data. The summary below shows us the age distribution between the two groups (p1: did not have strokes, p2:had a stroke) as we can see that people who had a stroke have concentrated around higher age.



BMI Summary We can summarize the BMI levels by stroke

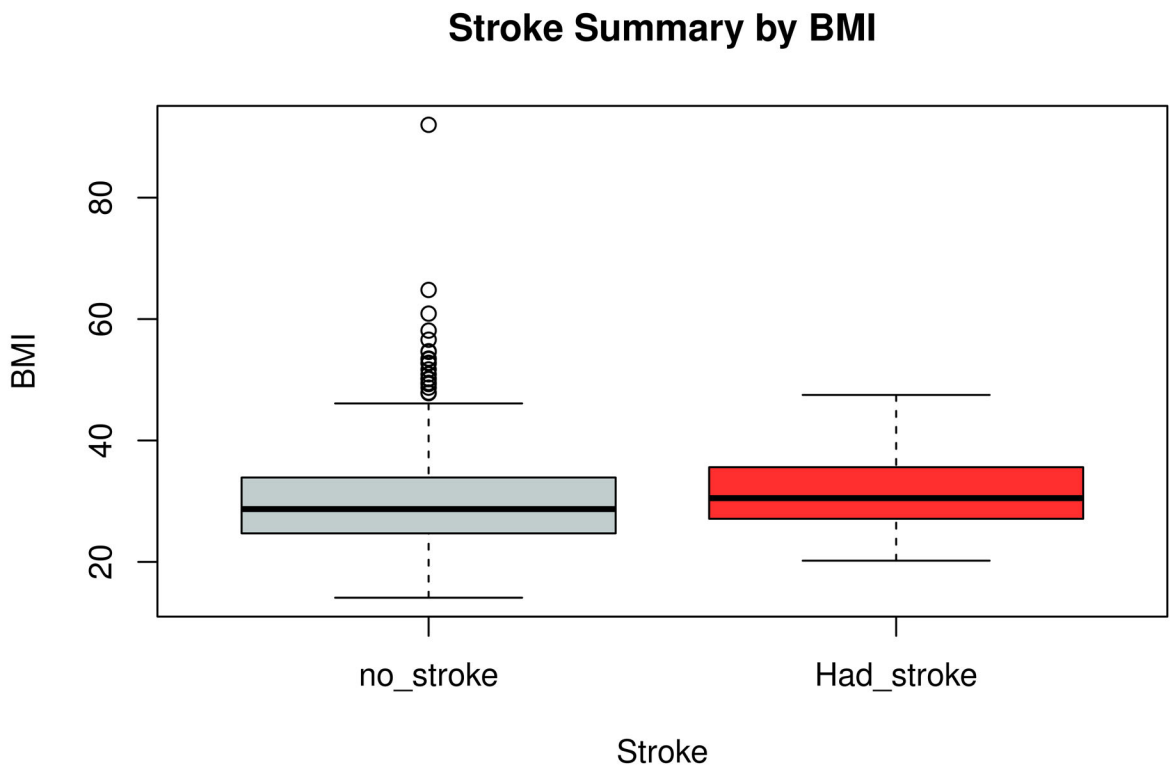
```
summary(data$bmi[data$stroke ==1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.20   27.10   30.50   31.41   35.60   47.50
```

We can summarize the bmi group that who did not have stroke

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.10   24.70   28.70   30.09   33.90   92.00
```

Here is the graph that summarize our BMI data. This shows us that BMI level for people who had stroke starts from 20.20. We had a lot of outliers in the BMI, this might have some effect in our data model but this wont be the same effect if we were using linier regression.



test for the global null hypothesis tests the following hypotheses
H0:βBage=βBbmi=0 (there is no association)
H1: there is at least one βBi != 0 (there is an association)

Test statistic

z = Beta1 / SEBˆ1

Decision rule
- Determine the appropriate value from the standard normal distribution associated with a right hand tail probability of alpha/2 = 0.025 - Using the Standard Normal Probabilities table z, alpha/2 = 1.960 - Decision Rule: Reject H0 if |z| >=1.960 or Reject H0 if p=< alpha
- Otherwise , do not reject H0

Computing the test statistic and the associated p-value

We can use R to calculate the test statistics

```
library(aod)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
m2 <- glm(formula = data$stroke ~
           + data$bmi
          +data$age, family = binomial)

#summary(m2)

wald.test(b=coef(m2), Sigma = vcov(m2), Terms = 2:3)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 45.7, df = 2, P(> X2) = 0.00000000012
```

Conclusion for global test
Since our global test p value 0.00000000012 is smaller than alpha level 0.05 we reject the null hypothesis and conclude that there is at least one there is at least one βBi != 0

Now we can also review each individual parameters and summarize the test result for each of them: Here is the summary;

```
summary(m2)
```

```
##
## Call:
## glm(formula = data$stroke ~ +data$bmi + data$age, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.86890  -0.34534  -0.19480  -0.09848   3.09661
##
## Coefficients:
##             Estimate Std. Error z value            Pr(>|z|)
## (Intercept) -9.04276    1.16105  -7.788 0.00000000000000678 ***
## data$bmi     0.03974    0.01992   1.994              0.0461 *
## data$age     0.08235    0.01219   6.754 0.00000000001435851 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 402.90  on 999  degrees of freedom
## Residual deviance: 332.63  on 997  degrees of freedom
## AIC: 338.63
##
## Number of Fisher Scoring iterations: 7
```

```
# OR
exp(cbind(OR = coef(m2), confint.default(m2)))
```

```
##                        OR         2.5 %      97.5 %
## (Intercept) 0.0001182435 0.00001214765 0.001150966
## data$bmi    1.0405366127 1.00068731300 1.081972788
## data$age    1.0858316155 1.06019274524 1.112090516
```

Age:
The hypotheses tested here are
H0:$\beta Bage$=0 or ORage=1 (there is no association between age and risk for stroke event,

after controlling for BMI)
H0:βBage=0 or ORage=1 (there is an association between age and risk for stroke event, after controlling for BMI)

We see that age is significantly associated with stroke event after adjusting BMI. We will reject the null hypothesis that or after adjusting for BMI since our p value 0.00000000001435851 $<$ 0.05(alpha level). The odds ratio of a stroke is $e^B == 1.086$ for every 1 year increase in age.

BMI:
The hypotheses tested here are
H0:βBbmi=0 or ORbmi=1 (there is no association between BMI and risk for stroke event, after controlling for Age)
H0:βBbmi=0 or ORbmi=1 (there is an association between BMI and and risk for stroke event, after controlling for Age)

we will reject the null hypothesis that or after adjusting for BMI since our p value 0.0461 $<$ 0.05(alpha level). BMI level is significantly associated with having a stroke event. The odds ratio of a stroke is $e^B == 1.04$ for every 1 unit increase in bmi.

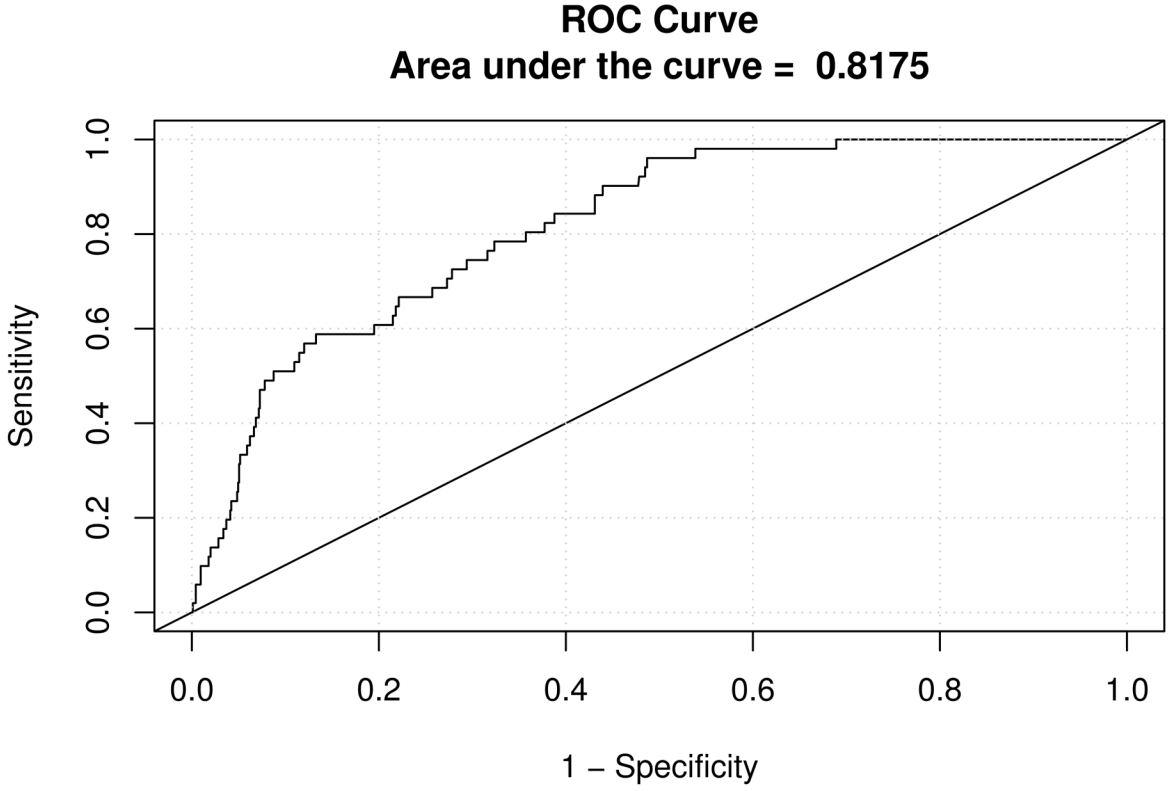
Here is our c statistics. It is 0.82. It is very close to one, so our model is very accurate.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
##
## Call:
## roc.formula(formula = data$stroke ~ data$ProbofTL3)
##
## Data: data$ProbofTL3 in 949 controls (data$stroke 0) < 51 cases (data$stroke 1).
## Area under the curve: 0.8175
```

Here is a graph of our model, we see that it is very close to one, so our model did a good job on prediction.



CONCULUTION

We could not reject the null hypothesis for question one(there is not enough evidence to support that the underlying population proportion are different). So there is no enough evidence to support that stroke is a men problem.
With question two, we learned that age and BMI levels are associated with the stroke event, and our model can predict the odds of having a stroke event with 82% accuracy.The limitation of the logistic regression model is that it might provide a different result if there is an irrelevant parameter, so we need to identify all the relevant independent variables before head.