Chances of getting stroke: Difference between Male and Female

Hugo TING

2022-12-14

Please find the link to my GitHub repo **here**

Introduction

There is a growing recognition of the clinical and public health importance of stroke in women. From an **article** published by the Nationl Library of Medicine, it is said that age-specific stroke incidence and mortality rates are higher in men than in women, stroke affects a greater number of women because of their increased longevity and the fact that stroke event rates increase substantially in the oldest age groups.

To address this topic, I regress the probability of getting stroke between male and female. Additional confounding variables such as age, smoking habit, marriage status and blood glucose level are included in the model to support the study. Variables are described as following:

Dependent variable: - Chance of getting stroke. It is equal to 1 if the record have had stroke, 0 otherwise Independent variables: - A binary column is created for male. It is equal to 1 if the record is a male, 0 if the record is a female. Confounding Variables: - Age: Numeric column, only contains records who are 18 or above - Smoker: Binary column. It is equal to 1 if the person has smoked or is a smoker. - Married: Binary column. It is equal to 1 if the person is married or was married. - Glucose level: Numeric column.

Hypothesis

Based on the background information we have, we assume that gender difference has no effect on chance of getting stroke as below.

$$H_0 : \beta_1(male) = \beta_2(female)$$

So once we have the results showing if gender difference have a robust effect on the chance of getting stroke, we can deny the null hypothesis if the beta fulfils the following hypothesis.

$$H_1 : \beta_1(male) \neq \beta_2(female)$$

Data Cleaning

Source

The data for this study has been retrieved from a confidential source as stated by the author. The selected data set is published on Kaggle - Stroke Prediction Dataset **here**.

Filters

Records which are below 18 years old are filtered as adults and married people with different parameters are focused in this analysis. Moreover, records which have a "Other" value in the gender column is removed. All unnecessary columns for this analysis, such as residential type and work type are removed. The BMI column is also removed because there are too many NAs and a lot of values were spotted the same, the author did not clarify if the bmi data is legit, which could be found here, so I decided not to include the BMI column as one of the independent variables. After filtering, 4253 records are left, thus N = 4253.

X, Y and Z Variables

A dummy variable "stroke" is transformed from character to numeric although it is already 1 and 0 in the data, where 1 represents the person have had stroke and 0 otherwise. In addition, I created a binary variable for "male" as the explanatory variable, it is equal to 1 if the record is a male and 0 is the record is a female. I added more binary variables for the confounding variables such as "married" and "smoker". 1 in "married" means the record is married or has married before, 0 otherwise. 1 in the "smoker" means the record is a smoker or has smoked before, 0 otherwise. While the other z-variables such as "age" and "avg_glucose_level" is preserved as numeric.

Summary Statisitcs

The summary statistics table shows that 6% of the respondents have had stroke. 39% of the respondents are male. The average age of the respondents is around 50 years old. 18% of the records are/were smokers. 79% of the respondents are/has married. The average glucose level of the respondents are 108.51.

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
Stroke	0.06	0.00	0.23	0.00	1.00	0.00	1.00
Male	0.39	0.00	0.49	0.00	1.00	0.00	1.00
Age	50.21	51.00	17.83	18.00	82.00	21.00	79.00
Smoker	0.18	0.00	0.39	0.00	1.00	0.00	1.00
Married	0.79	1.00	0.41	0.00	1.00	0.00	1.00
Glucose Level	108.51	92.44	47.77	55.12	271.74	60.77	219.71

Correlation Matrix

A correlation matrix is used to show a big picture of the association among dependent, independent and z-variables. The correlation matrix is shown in the Appendix. The correlation matrix shows that having a stroke is positively correlated with male, age, married, average glucose level and smoker. While surprisingly, smokers and having stroke have almost no correlation, let see if it is really further in the case study.

Model

The main hypothesis of this analysis is that male has lower chance of getting stroke. So a simple linear probability model with the chance of getting stroke as a dummy variable regressed on male (when the value = 1 in "male" column). Below are the probability models tested.

Model 0

$$Stroke^P = \alpha + \beta(male)$$
$$Stroke^P = \alpha + \beta(age)$$

Table 2 in the appendix shows the probability of getting a stroke. Column 1 (lpmmale) indicates that male are 95% confident to have 1.04% more chance to get a stroke while comparing to a female, without considering other independent variables. From the second column we can see that 1 year increase in age since 18 years old will have 0.33% higher chance of getting stroke, without taking other independent variables into account. For graphical visualization, from Figure 2 we can see that the regression line is in upward slope, which means if the person is a man, his chance of getting stroke is higher. The same can be observed in Figure 3 for the relationship between age and the chance of getting stroke.

Figure 2 : male's Probablility of getting stroke

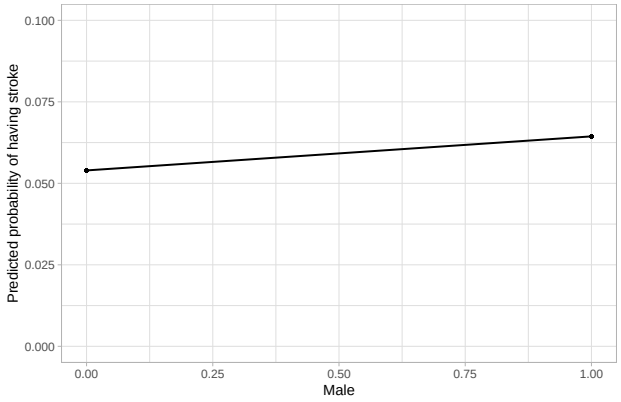
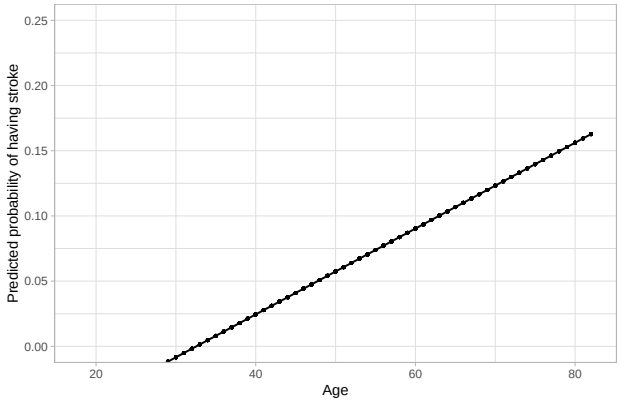


Figure 3 : Person's age Probablility of getting stroke



Model 1

$$Stroke^P = \alpha + \beta_1(male) + \beta_2(age)$$

As the R-squared of male shown in Table 1 is only 0.048%, it is too low to take it into account, and it means other independent variables are also affecting the probability of a person to get stroke (i.e. age). Adding the age into account at the same time as male, we can see that the R2 increased to 6.29% significantly in Table 2 in the appendix. We can also interpret that in this model, male only have 0.4% higher chance of getting stroke than a female of the same age, which is 1% less chance of male getting stroke that we got from Model 0 (1.04%)

Model 2

$$Stroke^P = \alpha + \beta_1(male) + \beta_2(age) + \beta_3(smoker)$$

As we can see in the correlation Matrix, smokers almost have no correlation with getting stroke. However smoking is scientifically that will increase the chance of having stroke. I would like to look at how smoking might affect the chance of getting stroke, so smoker as a parameter is put into account. From the Table 3 in the appendix, we can see that the Pseudo R2 only increased by 0.001 in both logit and probit model. It means that the scalability of data did not improve by much.

Model 3

In Model 3, females who are 18 or above is taken as the base category for comparison. 2 more z-variables, married and glucose level, are added into the model as the last 2 parameters. From Table 3 in the appendix, we can see that both the R2 of logit and probit model increased by 0.9%, in total around 17%. The R2 of Model 3 is the higest in all the models, so this will be the preferred model for the analysis.

$$Stroke^P = \alpha + \beta_1(male) + \beta_2(age) + \beta_3(smoker) + \beta_4(married) + \beta_5(glucose)$$

From Table 4 in the appendix, we can see the results of the lpm, logit, logit margin, probit and probit margin regressions. From column 1 - LPM, we can see that male on average is 0.2% more likely to get a stroke than women. However the 95% confidence interval of the probability of male than female to get stroke is [-1.2% , 1.6%], this means we cannot casually conclude that male has a 0.2% higher chance then female getting a stroke, because the 95% confidence level is much broader than that. It is also the case for smokers, although the mean chance of smokers from the data set to get stroke is 0.6% higher, the 95% CI is [-1.2% , 2.4%]. It is strange to see that the increase or decrease in glucose level does not affect the chance of getting stroke. We can possibly interpret this phenomenon as the robustness of glucose level on the effect of affecting stroke chance is statistically insignificant.

However, for age, as the SE of the LPM coefficient is 0, we can say that it is 95% confident that with 1 year increase in age, there will be 0.3% higher chance to get stroke, which aligns with what we got in other models as well. So we can interpret that age is one of the dominant and the most robust factor in determining the chance of getting stroke.

We can also interpret that marriage is also another robust factor which changes the chance of getting stroke. IN a 95% CI, married people are likely to have [1.9% , 5.5%] lower chance to get stroke than the people who are not married.

Robustness check

The results from the regression justify that our null hypothesis stands as the effect of gender on the chance of getting stroke is statistically insignificant enough to say that male have higher or lower chance of getting stroke than female. To further check the robustness of keeping the null hypothesis, logit and probit regressions are run for Model 3. From Table 4 in the appendix, we can see in Column 2 and 3, the logit coefficients are around 20 times larger than the logit margin. As well as the probit coefficient and margin, the probit coefficient is 10 times larger than the probit margin. While the logit and probit margin are pretty similar to the LPM value of male, age and glucose level, the logit and probit margin of smoker and marriage is far different, and from this difference, we can probably interpret that the logit and probit model is not too close or similar to the LPM 45 degree line. However, we can still take away 3 key findings, which are, age has a positive correlation with chance of getting stroke, marriage has a negative relationship with chance of getting stroke, and gender does not really affect the chance of stroke, at least based on this dataset.

Conclusion

Based on the result of regression analysis, it can be said that the preferred model is Model 3. Pseudo R-squared is 17.1% which is not too bad for getting an overall picture of the data. Across all models it is evident that gender indifference is not likely to have an effect on stroke chance. Thus, the Null hypothesis stands, whcih means there is no significant difference between beta (male) and beta (female).

However if we take the factor of age into the regression between gender and chance of getting stroke, and we can see that age has an upward sloping regression line with chance of getting stroke, the **difference in life expectancy** between male and female may explain why the reports from scientists said female have higher chance of getting stroke. From a data retrieved from **WorldData.info**, it says that male have on average

70.6 years of life while women have 75.1 years in 2021. In light of this, we can conclude that, there actually is a gender difference in chance of getting stroke, but the gender itself is not the robust reason, instead, longer life expectancy of females is the reason. As female have higher life expectancy of on average 4.5 years than male, according to the Table 2 above, female have around 1% more chance of getting stroke in their lifetime than male.

Appendix

Correlation Matrix

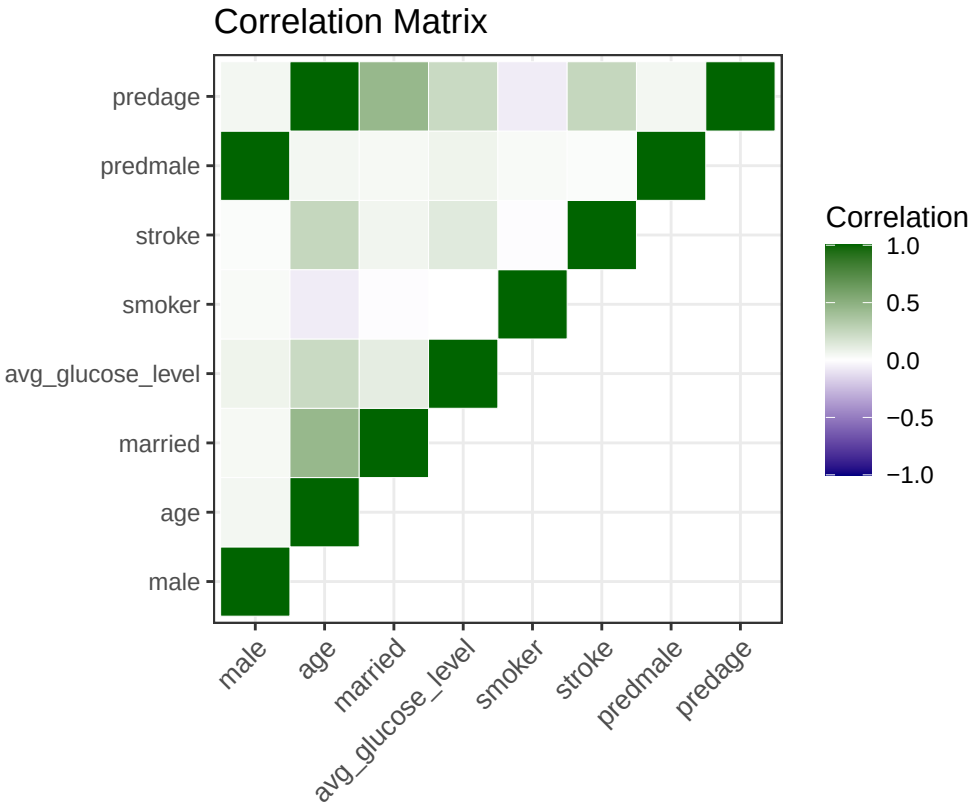


Table 2: Probability of getting stroke by gender and age status

Dependent Var.:	stroke	stroke	stroke
Constant	0.0540***	-0.1072***	-0.1084***
	(0.0045)	(0.0088)	(0.0091)
Male	0.0104		0.0039
	(0.0075)		(0.0072)
Age		0.0033***	0.0033***
		(0.0002)	(0.0002)
S.E. type	Hete.-rob.	Hete.-rob.	Hete.-rob.
Observations	4,253	4,253	4,253
R2	0.00048	0.06292	0.06299

Table 3: Logit, Probit with Pseudo R2

	(M1) logit	(M1) Probit	(M2) logit	(M2) Probit	(M3) logit	(M3) Probit
Constant	−7.382**	−3.752**	−7.477**	−3.805**	−7.738**	−3.932**
	(0.365)	(0.165)	(0.374)	(0.170)	(0.418)	(0.191)
male	0.135	0.055	0.125	0.049	0.083	0.029
	(0.138)	(0.070)	(0.138)	(0.070)	(0.140)	(0.070)
age	0.076**	0.037**	0.077**	0.037**	0.075**	0.036**
	(0.005)	(0.002)	(0.005)	(0.003)	(0.005)	(0.003)
smoker			0.249	0.132	0.253	0.130
			(0.184)	(0.091)	(0.185)	(0.092)
married					−0.196	−0.108
					(0.224)	(0.112)
avg_glucose_level					0.005**	0.002**
					(0.001)	(0.001)
Num.Obs.	4253	4253	4253	4253	4253	4253
RMSE	0.23	0.22	0.23	0.22	0.22	0.22
PseudoR2	0.159	0.161	0.160	0.162	0.169	0.171

* $p < 0.05$, ** $p < 0.01$

Table 4: The Probability of getting stroke across races- LPM, Logit, and Probit models

	(1)LPM	(2) logit coeffs	(3) logit Marg	(4) probit coeffs	(5) probit Marg
Constant	−0.128**	−7.738**	−7.738**	−3.932**	−3.932**
	(0.012)	(0.418)	(0.418)	(0.191)	(0.191)
male	0.002	0.083	0.004	0.029	0.003
	0.002	0.083	0.004	0.029	0.029
	0.002	0.083	0.083	0.029	0.003
	0.002	0.083	0.083	0.029	0.029
	(0.007)	(0.140)	(0.007)	(0.070)	(0.007)
	(0.007)	(0.140)	(0.007)	(0.070)	(0.070)
	(0.007)	(0.140)	(0.140)	(0.070)	(0.007)
	(0.007)	(0.140)	(0.140)	(0.070)	(0.070)
age	0.003**	0.075**	0.004**	0.036**	0.004**
	0.003**	0.075**	0.004**	0.036**	0.036**
	0.003**	0.075**	0.075**	0.036**	0.004**
	0.003**	0.075**	0.075**	0.036**	0.036**
	(0.000)	(0.005)	(0.000)	(0.003)	(0.000)
	(0.000)	(0.005)	(0.000)	(0.003)	(0.003)
	(0.000)	(0.005)	(0.005)	(0.003)	(0.000)
	(0.000)	(0.005)	(0.005)	(0.003)	(0.003)
smoker	0.006	0.253	0.013	0.130	0.013
	0.006	0.253	0.013	0.130	0.130
	0.006	0.253	0.253	0.130	0.013
	0.006	0.253	0.253	0.130	0.130
	(0.009)	(0.185)	(0.010)	(0.092)	(0.010)
	(0.009)	(0.185)	(0.010)	(0.092)	(0.092)
	(0.009)	(0.185)	(0.185)	(0.092)	(0.010)
	(0.009)	(0.185)	(0.185)	(0.092)	(0.092)
married	−0.037**	−0.196	−0.010	−0.108	−0.011
	−0.037**	−0.196	−0.010	−0.108	−0.108
	−0.037**	−0.196	−0.196	−0.108	−0.011
	−0.037**	−0.196	−0.196	−0.108	−0.108
	(0.009)	(0.224)	(0.013)	(0.112)	(0.012)
	(0.009)	(0.224)	(0.013)	(0.112)	(0.112)
	(0.009)	(0.224)	(0.224)	(0.112)	(0.012)
	(0.009)	(0.224)	(0.224)	(0.112)	(0.112)
avg_glucose_level	0.000**	0.005**	0.000**	0.002**	0.000**
	0.000**	0.005**	0.000**	0.002**	0.002**
	0.000**	0.005**	0.005**	0.002**	0.000**
	0.000**	0.005**	0.005**	0.002**	0.002**
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Num.Obs.	4253	4253	4253	4253	4253
RMSE	0.23	0.22	0.22	0.22	0.22

* $p < 0.05$, ** $p < 0.01$