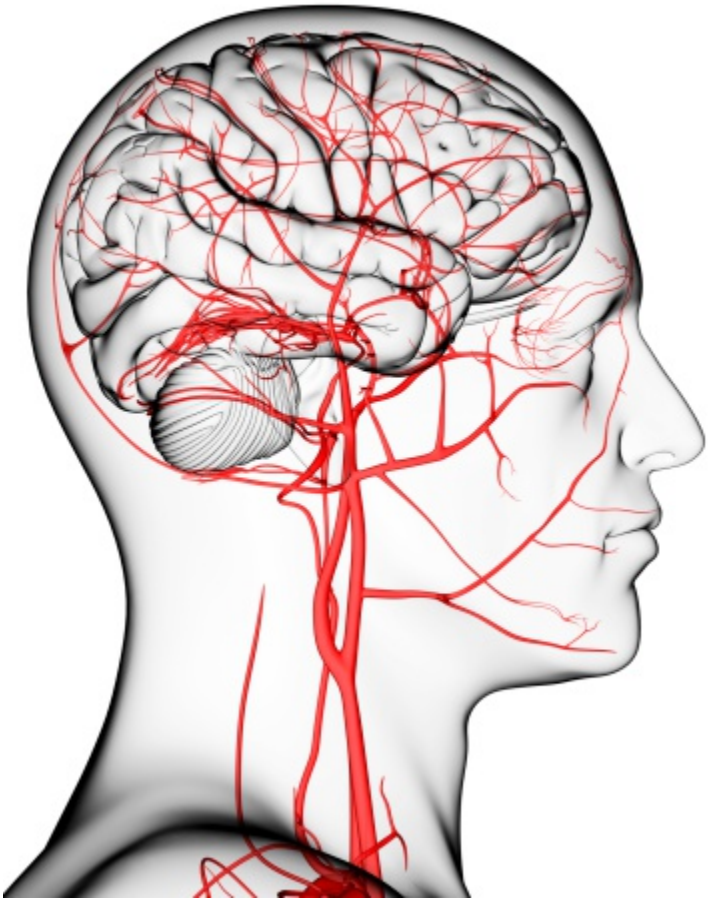
Stroke

Analysis and Prediction



Golden Gate University

BUS 240 Data Analysis for ManagersFall 2022

Justin Yu

December 1st, 2022

Table of Contents

Introduction

A stroke occurs when the flow of blood is restricted to a part of the brain (ischemic stroke) or when a blood vessel in the brain bursts (hemorrhagic stroke). This occurrence can lead to permanent damage of the brain and cause long-term disability. According to the CDC, 795,000 in the United States have a stroke every year and 150,000 of these cases lead to death.

Living a healthy lifestyle and controlling any existing health conditions is key to preventing a stroke. Simple actions such as keeping a healthy weight, not smoking, and treating hypertension and heart disease will significantly decrease the chances of a stroke.

This paper will present the analysis of health data and the prediction of whether or not a person will suffer a stroke. Data cleaning, transformations, and exploration has been done to discover what the data unfolds. It will be established that a multiple regression model falls short when dealing with a categorical response variable. The cleaned and transformed data will provide smooth implementation of logistic regression modeling.

Data Collection

This health dataset was gathered from Kaggle. This website is a platform for data scientists to find and publish datasets to challenge and compete with others to create statistical and machine learning models.

Data Dictionary

The data has 12 variables. The dependent variable is stroke. The categorical predictors are gender, hypertension, heat_disease, ever_married, work_type, Residence_type, and smoking_status. The categorical variables that need to be converted to dummy variables are

gender, ever_married, work_type, Residence_type, and smoking_status. Their dummy variables will be labeled below.

- **id:** unique identifier
- **gender:** "Male", "Female"
    0 - Male, 1 - Female
- **age:** age of the patient
- **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- **ever_married:** "No" or "Yes"
    0 - Not married, 1 - Married
- **work_type:** 5 characteristics. private, self-employed, govt_job, children, never_worked
    0 - private, 1 - self-employed, 2 - Govt_job, 3 - children, 4 - Never_worked
- **Residence_type:** "Rural" or "Urban"
    0 - Rural, 1 - Urban
- **avg_glucose_level:** average glucose level in blood
- **bmi:** body mass index
- **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"*
    0 - never smoked, 1 - formerly smoked, 2 - smoked, 3 - Unknown
- **stroke:** 1 if the patient had a stroke or 0 if not

Data Exploration, Cleaning and Transformations

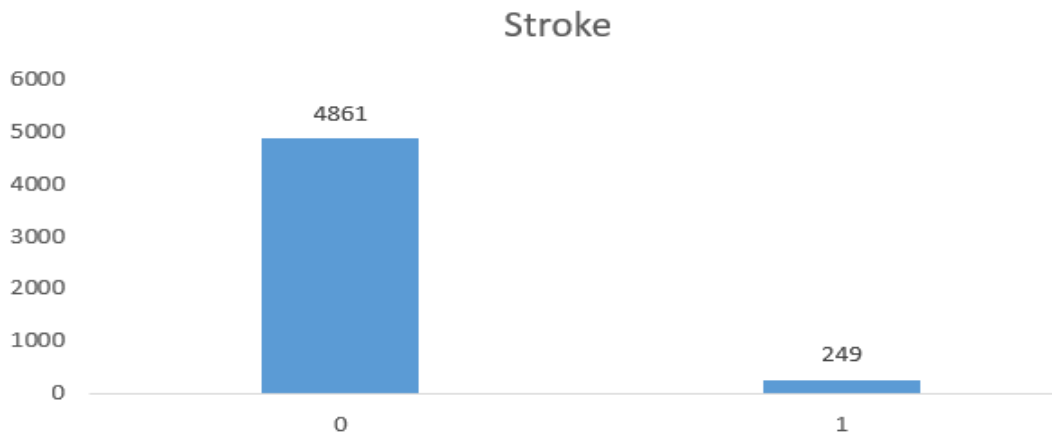
Since most of the variables are categorical, most of the descriptive statistics will have a count of each variable VS. the response variable **stroke**. Each variable will have an explanation on what they are.

Here are the initial descriptive statistics of **age**, **avg_glucose_level** and **bmi**. BMI has 201 N/A values. The column 'old bmi' shows the descriptive statistics with the N/A values. The decision to fill the empty values with the median is to avoid a huge increase in the mean value of bm (since the median is slightly below the mean). As seen in the new bmi column, the average bmi decreased by .03.

	age	*avg_glucose_level*	*old bmi*	*new bmi*
Mean	43.22661448	106.1476771	28.89323691	28.86203523
Median	45	91.885	28.1	28.1
Mode	78	93.88	28.7	28.1
Standard Devi	22.61264672	45.28356015	7.85406673	7.699562319
Minimum	0.08	55.12	10.3	10.3
Maximum	82	271.74	97.6	97.6

Stroke

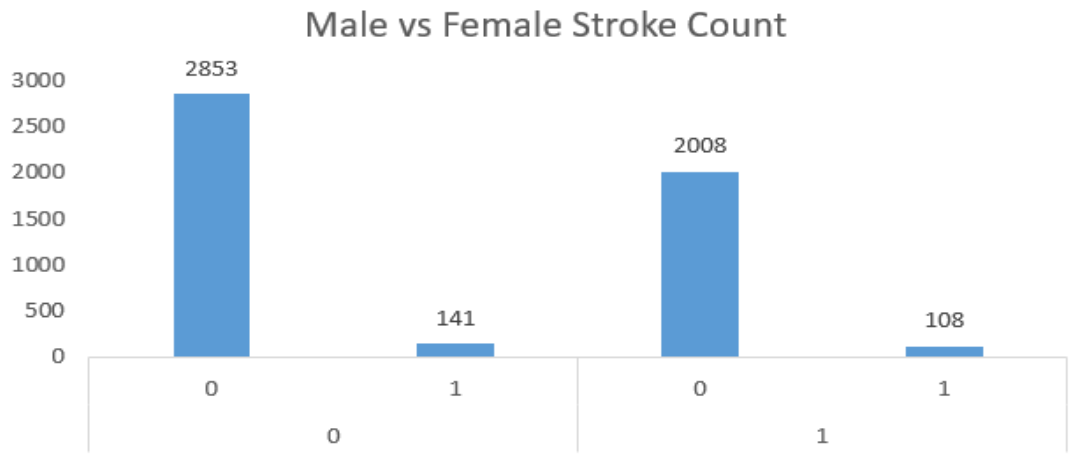
Counts for stroke. 95.13% of people did not have a stroke while 4.87% had a stroke.



Gender

Out of 2994 male patients, 141 suffered from a stroke. The percent is 4.7%

Out of 2116 female patients, 108 suffered from a stroke. The present 5.1%. The difference between these two percentages is not significant.



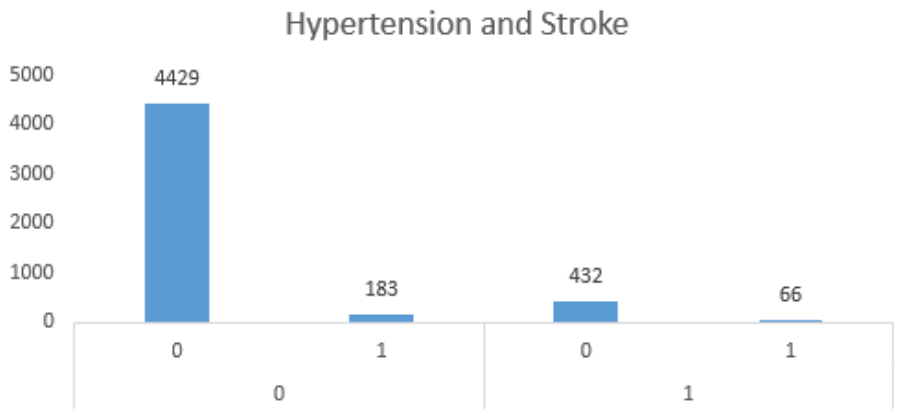
Age

The average age of a person having a stroke is 26 years older than the average person not having a stroke. This falls in line with the statistic by Stanford Health Care which says that the majority of strokes occur in people who are 65 or older.

Stroke	Average of age	Min of age	Max of age
0	41.97154495	0.08	82
1	67.72819277	1.32	82

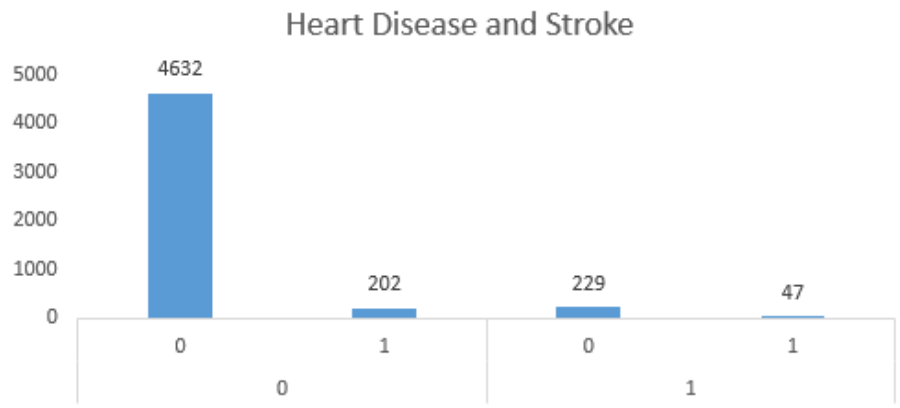
Hypertension

Hypertension is a synonym for high blood pressure. High blood pressure can produce clots in the arteries, preventing blood and oxygen from flowing to the brain. In this data, only 66 people who are diagnosed with hypertension suffered from a stroke.



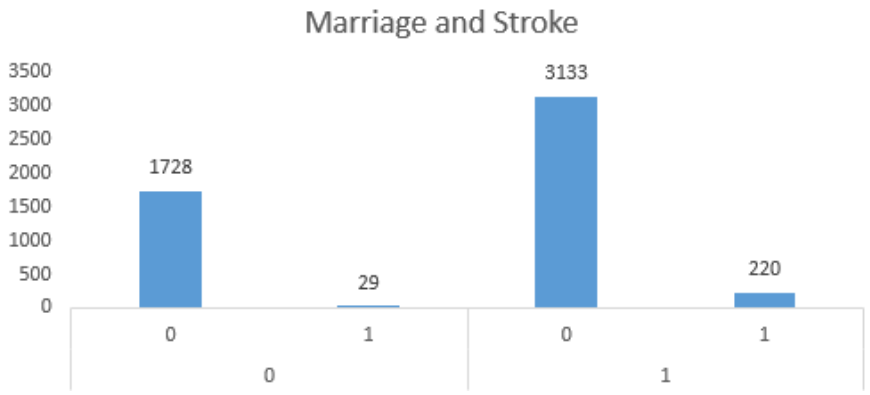
Heart_disease

Heart disease such as coronary artery disease causes plaque to build up in the arteries. This will block the flow of blood to the brain. Furthermore, other heart diseases can cause low pressure which will reduce blood flow to the brain. In this dataset, only 47 people out of 346 have suffered from a stroke. In comparison, 4.3% of people without heart disease versus 20.5% with heart disease have a stroke.



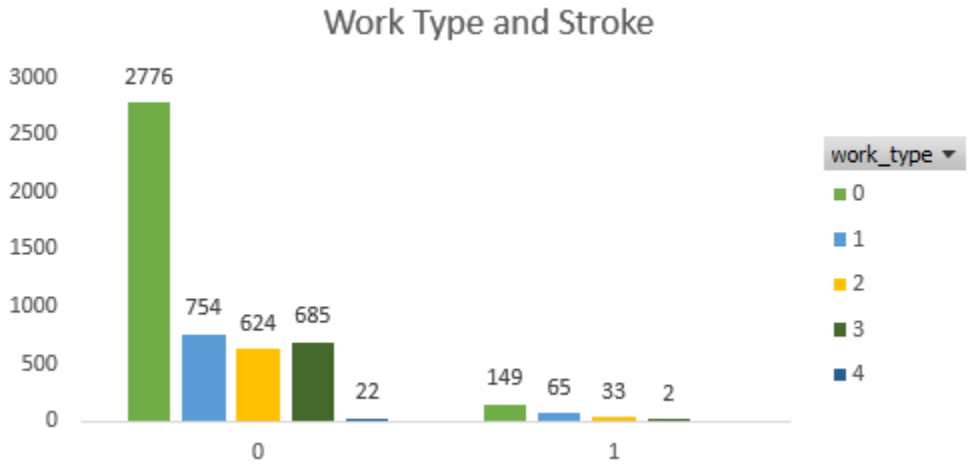
Ever_married

1.6% of people who are not married suffer from a stroke. 7% of people who are married suffer from a stroke.



Work_type

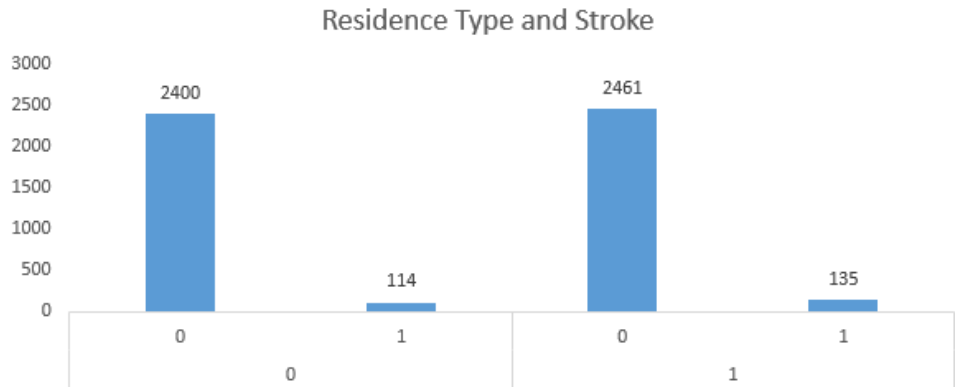
This graph will be accompanied by a frequency table since there are 5 characteristics to look at. The percentages are taken with respect to the **stroke** variable. The percentages of not having a stroke versus having one in the categories private (0), govt_job(2), and never_worked(4) are similar. The larger differences occur in the self-employed (1) and children (3) characteristics.



Stroke	0	1	2	3	4	Grand Total
0	57.11%	15.51%	12.84%	14.09%	0.45%	100.00%
1	59.84%	26.10%	13.25%	0.80%	0.00%	100.00%

Residence_type

The ratio of not having a stroke versus having a stroke depending on if you live in an urban or a rural area is very similar. The values are 4.75% and 5.4% respectively.



Avg_glucose_level

Glucose level means how much sugar is in the blood. A high glucose level can lead to a build up of sugar. This can damage blood vessels and lead to a hemorrhagic stroke. A low glucose level can shock the body, which leads to high blood pressure. Hypertension can cause a stroke. The average of this column with respect to the predictor (104.8 vs 132.5) indicates that a high glucose level is significant in predicting a stroke.

Stroke	Average of avg_glucose_level	Min of avg_glucose_level	Max of avg_glucose_level
0	104.7955133	55.12	267.76
1	132.544739	56.11	271.74

Bmi

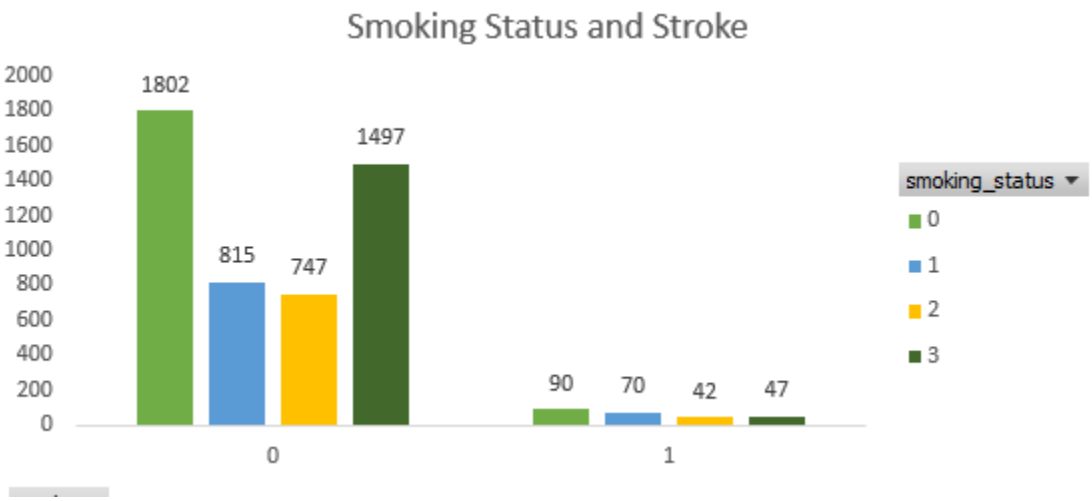
A body mass index greater than 30.0 indicates obesity. Being overweight leads to high blood pressure, high glucose levels, and other consequences. As discussed before, hypertension and high glucose levels can increase the chances of a stroke.

Stroke	Average of bmi	Min of bmi	Max of bmi
0	28.79911541	10.3	97.6
1	30.09036145	16.9	56.6

Smoking_Status

Smoking increases blood and pressure and reduces oxygen in the blood. Furthermore, smoking releases toxic chemicals into the body; specifically into the bloodstream damages blood vessels. This combination of factors can lead to ischemic and hemorrhagic stroke.

The data shows that people who formerly smoked (1) and people who smoke (2) lead to a higher chance of suffering a stroke. People who do not smoke (0) have the same percentage of not having a stroke and having one. The unknown (3) values make up 1544 of the data points. Since it takes up a large proportion of the data, the decision was made to leave this characteristic as unknown. According to the frequency table, this unknown variable accounts for 30.80% of people who do not have a stroke vs 18.88% of people who do. Based on these percentages, this unknown variable may be split into not smoking, formerly smoked, and smokes.



Stroke	0	1	2	3	Grand Total
0	37.07%	16.77%	15.37%	30.80%	100.00%
1	36.14%	28.11%	16.87%	18.88%	100.00%

Pitfall of Multiple Regression Analysis

Performing multiple regression analysis on this dataset is not ideal since the response variable is categorical. This would lead to a violation of the assumptions of linear regression. Recall that the assumptions are

1) Errors are independent of each other.
2) Errors are normally distributed
3) Homoscedasticity (errors have constant variance)

Assumptions 2 and 3 are violated. Errors cannot be normally distributed since the Y (**stroke**) values are either 0 or 1. There is no constant variance since the variance of the errors of Y will decrease and approach 0. Logistic regression does not require these assumptions.

Logistic Regression Analysis and Results

Explanation of the Logistic Model and an Example

The data is modeled using logistic regression. The logistic regression model is

$$\widehat{y} = \frac{e^{\beta_0+\beta_1 x_1+\ldots.+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\ldots.+\beta_n x_n}}$$

using exponent rules we get

$$\widehat{y} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots.+\beta_n x_n)}}$$

This nonlinear equation predicts the probability that Y = 1 for any x.

The range of $\widehat{y}$ is $0 < \widehat{y} < 1$.

For example, recall that our response (stroke) is 0 for no stroke and 1 for having a stroke. After plugging in the coefficients β_i's to obtain a value $\hat{y}$ = 0.2, the prediction for this data point would be 0 (no stroke). Any value of $\hat{y}$ greater than 0.5 will predict a 1.

Implementation of Model

This logistic model has been created in python using the packages sklearn and statsmodels.api. The dataset is split 80/20 into a training set and a test set. This means that 80 percent of the data will be used to create (train) the logistic model. Afterwards the 20 percent of data points in the test set will be inputted into the trained model to predict whether or not the patient had a stroke. This 80/20 split will result in 4088 rows in the training set and 1022 rows in the test set. Therefore, the model will be tested on how many correct predictions will occur in 1022 data points.

Here are the coefficients, z, and p-values from the initial implementation. Assume a significant value of 0.05. The coefficients have the hypothesis

$$H_o: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
age	0.0380	0.0039	9.7897	0.0000	0.0304	0.0456
hypertension	0.7327	0.1685	4.3489	0.0000	0.4025	1.0629
heart_disease	0.6778	0.1948	3.4787	0.0005	0.2959	1.0596
work_type	-0.4140	0.0796	-5.1977	0.0000	-0.5701	-0.2579
avg_glucose_level	0.0016	0.0013	1.2363	0.2163	-0.0009	0.0041
bmi	-0.1526	0.0088	-17.4142	0.0000	-0.1698	-0.1355
smoking_status	-0.1566	0.0553	-2.8311	0.0046	-0.2651	-0.0482
gender_Male	-0.0898	0.1345	-0.6680	0.5041	-0.3534	0.1737
ever_married_Yes	-0.4553	0.1811	-2.5145	0.0119	-0.8102	-0.1004
Residence_type_Urban	-0.2008	0.1270	-1.5806	0.1140	-0.4497	0.0482

Recall that if a p-value > 0.05, then the null hypothesis is not rejected and the coefficient is 0. The predictor **gender_Male** can be taken out of our model since its p-value > .05. Upon closer inspection, the confidence interval for **avg_glucose_leve**l [-.0009, .0041] includes 0. This indicates that there is a chance that the coefficient may be insignificant. However, the function in python assumes a 97.5 percent confidence interval. If a 95 percent CI was used, the interval would be smaller and will likely not include 0. From this information, the predictor **avg_glucose_level** will not be dropped. The final version of the logistic model will be implemented.

Logit Regression Results

Dep. Variable:	stroke	No. Observations:	5110
Model:	Logit	Df Residuals:	5101
Method:	MLE	Df Model:	8
Date:	Wed, 07 Dec 2022	Pseudo R-squ.:	0.06895
Time:	03:38:18	Log-Likelihood:	-926.57
converged:	True	LL-Null:	-995.19
Covariance Type:	nonrobust	LLR p-value:	8.891e-26

	coef	std err	z	P>|z|	[0.025	0.975]
age	0.0381	0.004	9.829	0.000	0.030	0.046
hypertension	0.7316	0.168	4.346	0.000	0.402	1.062
heart_disease	0.6636	0.193	3.430	0.001	0.284	1.043
work_type	-0.4151	0.080	-5.219	0.000	-0.571	-0.259
avg_glucose_level	0.0015	0.001	1.204	0.229	-0.001	0.004
bmi	-0.1537	0.009	-17.827	0.000	-0.171	-0.137
smoking_status	-0.1599	0.055	-2.899	0.004	-0.268	-0.052
ever_married_Yes	-0.4584	0.181	-2.530	0.011	-0.814	-0.103
Residence_type_Urban	-0.2001	0.127	-1.576	0.115	-0.449	0.049

The p-values and the confidence intervals of the coefficients look good. The model's overall fit is assessed by the LLR (log-likelihood ratio) p-value. This is testing the full model used against the model with only the coefficient.

$$H_o : \widehat{y} = \frac{1}{1+e^{-(\beta_0)}}$$

$$H_1 : \widehat{y} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots.+\beta_n x_n)}}$$

The small p-value indicates that the full model is statistically significant. This benchmark for a model is similar to the F-test for linear regression. Another way to calculate the strength of the logistic model is to find the percent of **CORRECT** predictions. The model correctly predicts **95.12 percent** of 1022 strokes. A confusion matrix (contingency table) tallies the correctly labeled (true negative, true positive) predictions and the incorrectly labeled ones (false negative, false positive). The table is attached below.

	FALSE	TRUE
FALSE	971	1
TRUE	49	1

	FALSE	TRUE
FALSE	True Negative	False Positive
TRUE	False Negative	True Positive

To be more specific, the true negatives are the stroke data points labeled '0' for no stroke and the true positives are the points '1' for a stroke. The false negatives and false positives indicate an incorrect prediction.

Here are the first 10 out of 1022 results predicted when using the test data on the trained model.

```
[9.89241755e-01 1.07582452e-02]
[9.58500149e-01 4.14998506e-02]
[9.73783255e-01 2.62167447e-02]
[9.82829626e-01 1.71703741e-02]
[9.91257614e-01 8.74238643e-03]
[9.93337699e-01 6.66230089e-03]
[9.68107286e-01 3.18927136e-02]
[8.87067772e-01 1.12932228e-01]
[9.67695686e-01 3.23043138e-02]
[9.83487557e-01 1.65124430e-02]
```

Both columns indicate a probability. The left column shows the probability that the patient will not suffer from a stroke and the right column shows the probability that the patient will suffer from one. All the probabilities on the left are 88% or greater. Therefore, the first 10 points are predicted to be 0 (no stroke).

Coefficient Analysis

The equation for the logistic model is $\widehat{y} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots.+\beta_n x_n)}}$

As the denominator increases, the value of $\widehat{y}$ decreases.

Using python, the intercept term β_0 = -7.02377607 whereas the rest of the terms are

	coef
age	0.0381
hypertension	0.7316
heart_disease	0.6636
work_type	-0.4151
avg_glucose_level	0.0015
bmi	-0.1537
smoking_status	-0.1599
ever_married_Yes	-0.4584
Residence_type_Urban	-0.2001

From this, hypertension and heart_disease are key figures in causing strokes, while work type and marriage status seem to prevent them.

Conclusion

Excel and python were used to provide descriptive statistics and a logistic regression model for this healthcare data set. Research on what health conditions lead to strokes was done by reading information provided by the CDC. Extra sources beyond our textbook (Doane) were used to gain more insight on the math behind logistic regression. As a result, a clear understanding of how maximum likelihood, log-likelihood, and gradient descent derive the logistic equation could help me understand how to create a more accurate model by reducing bias. Bias is mentioned because I am suspicious of the high accuracy of the correct predictions (95.12%). In conclusion, despite the long process of analyzing factors that lead to strokes and creating a logistic model to make predictions, I believe the model can still be improved with a stronger understanding of statistical modeling.

References

CDC. (2022, October 14). *Stroke facts*. Centers for Disease Control and Prevention. Retrieved December 1, 2022, from https://www.cdc.gov/stroke/facts.htm

Doane, D. P., & Seward, L. W. (2019). *Applied Statistics in business and Economics Sixth Edition*. McGraw-Hill/Irwin.

Fedesoriano. (2020) *Stroke Prediction Dataset*. Retrieved November 22, 2022 from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

R.Yan, Moncada-Torres Arturo (2017, October 12). *Python: How to interpret the result of logistic regression by sm.logit.* Stack Overflow. Retrieved December 4, 2022, from https://stackoverflow.com/questions/46700258/

Stojiljkovic, Mirko (2022, September 1). *Logistic Regression in Python.* Real Python. Retrieved December 9, 2022, from https://realpython.com/logistic-regression-python/