The Research of Influence Factors that Possibly lead to Stroke

Wenbo Wang*

Limai Chinese International School, Beijing, 100024, China

*Corresponding author: wenbo20060529@outlook.com

Abstract. Although previous studies have demonstrated that the incidence of stroke is associated with hypertension, heart disease. In this research, the method Binary Probit Model is used to deal with the data from Kaggle named “Diabetes, Hypertension and Stroke Prediction” which was published 2015 for 70692 survey responses. It is concluded that although stroke has no connection with Gender, it has a relatively strong connection with heart disease and hypertension. Many of the factors having a possibly strong connection with stroke have never appeared in previous studies, such as heart disease which provides some new perspectives to study the pathogenesis. Advances in pathogenesis help improving the treatment of stroke and reach out for further research afterward. This paper uses Probit regression to study the relation between the three diseases from the data found in Kaggle, and the result seems to support a positive correlation between them which the three diseases are closely related to each other.

Keywords: Binary probit model; influence factors; stroke.

1. Introduction

Stroke is a disease that affects the arteries leading to and within the brain which it is also called cerebrovascular accident [1]. Stroke is also classified into two groups such as Ischemic strokes and Hemorrhagic strokes. Ischemic strokes are strokes caused mainly by blockage of an artery, which about 87 percentage of all strokes are ischemic [2]. Hemorrhagic strokes are strokes caused by bleeding, which only about 13 percentage among all strokes are hemorrhagic [3]. According to the fact that stroke affects nervous system within the brain, the disease is acute and crucial. Although stroke is severe to human health, patients need to obtain timely medication, which it is possible to reduce the sequela from stroke. Also the symptoms of stroke is sometimes difficult to discover, so many patients might neglect the symptoms and could not receive formal medical care in a timely manner. Since stroke could be very difficult to be aware and difficult to treat, it is necessary to learn more about the factors that cause stroke. There are many kinds of potential factors that might cause stroke such as diseases, inheritance, and some other external factors.

Generally, it is possible that disease other than stroke might be possible relating factors. Hypertension is one of the most prevalent risk factor for especially ischemic stroke because hypertension causes high blood pressure that causes more damages to vascular wall [4]. The accumulating damages lead to bleed within the brain in theory, so hypertension is one of the potential factors considered in this paper [5]. Heart disease is another potential factor considered in this paper, because heart disease and stroke all could be related to damage of the vascular system [6]. Heart disease and stroke also seems to share some relation due to their similarities [7]. Among the three diseases, hypertension is relatively the mild one because it barely cause instant death [8]. However, heart disease and stroke all commonly cause significant damage to human health because they directly damage important organs of the human body [9]. Furthermore, due to the fact stroke is a terrible disease, it is important to prevent stroke. Regular physical examination: Middle aged and elderly people are prone to symptoms of stroke, so regular physical examinations can be used to prevent stroke [10]. During physical examination, blood samples need to be drawn for blood lipid and blood sugar tests, blood pressure needs to be measured, and if necessary, brain CT scans can be performed [11]. Theoretically, many factors are related to diseases, but many factors have been fully studies such as the relation between smoking and stroke. This paper stands on the point of view which tries to uncover the factors related to diseases. The influential factors of disease types this paper studies are heart disease and hypertension, because these two diseases seem to have strong connection

through daily experience and theory. Nevertheless, scientific facts require strict experiments to prove it and stand on a solid foundation.

In brief, the relation between these three diseases could be explained and they seem to be closely related through daily experience, but the true condition is still unknown, so more survey and experiments are needed to ensure the true relation [12]. This reflects the objective of the paper, which this literature relies on data of the survey on stroke to do statistical analysis in order to find some useful results to uncover more information about the relationship between these three diseases. After consideration and optimization, this paper will use the Probit Regression model to study the effect of hypertension and heart disease on stroke, whether they are factors that possibly lead to stroke.

2. Methods

2.1. Data Sources

The data for this paper is collected from the Kaggle website, which Shwetanshu Goel, Sola Fajobi and Kaushal Krishna are the top three contributor to the compile from the 70,692 survey responses from cleaned BRFSS 2015.

2.2. Variable Selection

The data used in this paper count a total of 40910 people from the 70692 survey responses, including the people who gets stroke or not, of whom 22710 are male and 18197 are women. The patients' ages range from 0 to 103 years. The data contains 2 variables (Hypertension and Heart Disease).

Table 1. Logogram and numbers of the 2 factors

Elements	Logogram	Number1	Stroke1
Hypertension	x_1	8748	6532
Heart Disease	x_2	5225	4142

*Number 1: The number of people suffering from the disease.
**Stroke1: The number of stroke patients suffering from the disease.

Table 1 shows the number of people and stroke patients who have the disease. As shown in Table 1, for the convenience of writing, the logogram of the factors is as above. The sample of data is 40910, of which 20460 have stroke.

Table 2. Gender distribution by age group

Age	[0-19]	[20-39]	[40-59]	[60-79]	[80-99]	[100-199]
Number2	3170	9618	8336	11192	4306	28
Stroke2	1873	4432	5563	5577	2929	28

*Number 2: The number of people in each age group.
**Stroke 2: The number of people in each age group who have stroke.

Table 2 shows the stroke patients in each age group that range from 0 to 199. The majority of people contributed to stroke patient age from the range 60 to 79. Aged people are more common to be stroke patients from the data.

2.3. Research Protocol

This paper uses the Probit Regression model, whether or not to have stroke is the dependent variable(Y), and the 2 factors are the independent variables(X), where 0 represents no and 1 represents yes. Next, this paper uses SPSSAU to analyze the relationship between the effect of X on Y, i.e., the relationship between the 2 factors on stroke.

2.4. Model Principle

Similar to logistic regression, Probit regression is also a regression method for fitting $0-1$ dependent variables, converting variables with values distributed in the real range into probability values distributed in the $(0,1)$ interval through a cumulative probability function. The link function of Probit regression can use the logit cumulative probability function or the Inverse function of the cumulative probability function of the standard Normal distribution Φ^{-1}, which is the Probit function. Probit regression is used for statistical analysis of data whose dependent variable is a categorical variable, which is similar to Logistic regression. Deforming Logistic mode, it can be seen that:

$$P = \frac{e^{X\beta}}{1+e^{X\beta}} \tag{1}$$

The right-hand side of the above equation ($\frac{e^{X\beta}}{1+e^{X\beta}}$) happens to resemble the probability distribution function of the standard growth distribution which is also known as logistic distribution. The Probit model assumes that the probability distribution function on the right is similar to that of the standard normal distribution:

$$P = \int_{-\infty}^{X} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X\beta)^2}{2}} dx \tag{2}$$

2.5. Model Testing

It can be seen from Table 3 that as model fitting quality is judged by model prediction accuracy, the overall prediction accuracy of the research model is 65.52%, and the model fitting is acceptable. When the true value is 0, the prediction accuracy is 85.20%. In addition, when the true value is 1, the prediction accuracy rate is 45.86%.

Table 3. Binary probit regression prediction accuracy summary

		predicted value		forecast accuracy	Prediction error rate
		0	1		
true value	0	17424	3026	85.20%	14.80%
	1	11078	9382	45.86%	54.14%
Summary				65.52%	34.48%

3. Results and Discussion

Figure 1 shows a series of factors that cause stroke in the proportion of people with stroke. These factors that cause stroke are hypertension and heart disease. There are a total of 40910 samples, the proportion of stroke samples that were identified from the samples and classified according to different influencing factors was determined to determine the probability that each risk would cause stroke.

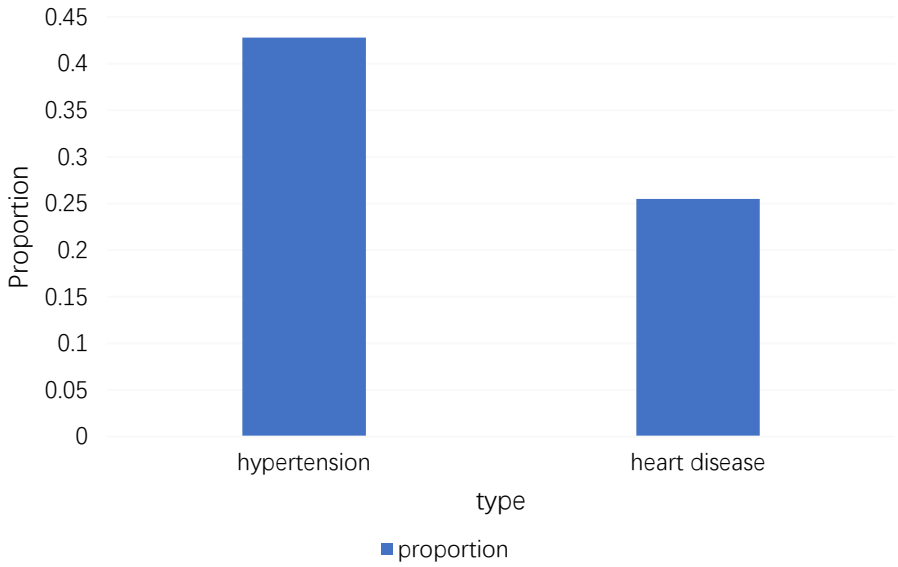


Fig. 1 Factors in the proportion of people with stroke

In above Figure 1, it is not difficult to see that hypertension accounts for the higher proportion, up to 42.8%. Heart disease only accounts for a smaller proportion, only up to 25.5%. Thus, according to the chart data, hypertension leading to a greater hidden risk of stroke.

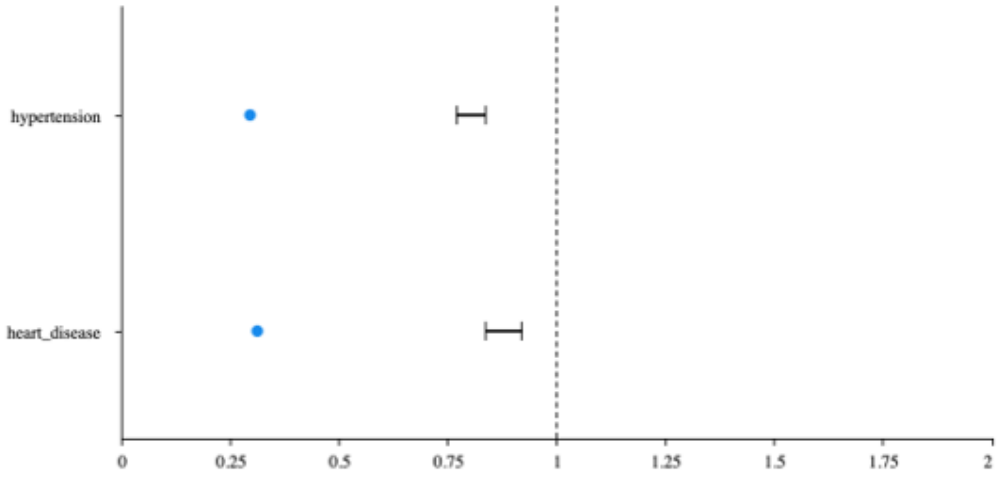


Fig. 2 Coefplot

As shown in figure 2, Coefplot directly presents the regression result of Probit Regression in graphs. This manner is more intuitive to express the conclusions of the paper. CI in the chart is the acronym of confidence interval. Confidence interval approximates the range of the parameter that represents the overall data. Therefore, CI can be used to approximate the overall parameter and replace hypothesis-testing to test the effectiveness of the model. In the case of this paper, Coefplot directly prove that Probit Regression is relatively an effective model.

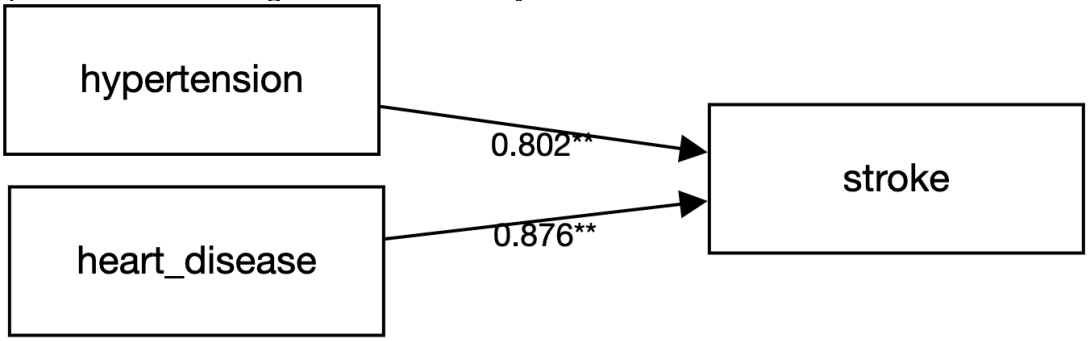


Fig. 3 Variable-related schematic

As shown in Figure 3, the research introduced various factors that might be related to stroke into the model, including hypertension and heart disease. These two factors are connected to stroke by arrow with the regression coefficient of each factor, which 0.802 for hypertension and 0.876 for heart disease. Through calculations, this study yields the final linear regression equation:

$$Probit(p) = -0.266 + 0.802x_1 + 0.876x_2 \quad (3)$$

Where p represents the probability of stroke being 1.

Table 4. Model results

Elements	Model coefficient	S. E.	z value	p-value	95% CI
x_1	0.802	0.005	53.942	0.000	0.283 ~ 0.305
x_2	0.876	0.006	48.634	0.000	0.297 ~ 0.322

*Dependent variable: Stroke

Table 4 is about marginal effect which marginal effect refers to the additional effect caused by adding one unit and often used more in the field of econometrics. First, the above table shows the marginal Effect size at the mean value and the corresponding test and 95% confidence interval value. Second, if the marginal Effect size value is significant, it means that there is a significant marginal effect, otherwise it means that there is no significant marginal effect. Third, if the marginal Effect size is significant and greater than 0, it means that the increase of X will bring about positive effect changes. If the marginal Effect size is significant and less than 0, it means that the increase of X will bring about negative effect changes.

It is manifest from Table 4 that the model gets the conclusion that by judging whether the p-value is greater than 0.05. If the value p is larger than 0.05, then the model is not useful. In the case of this paper, p value for heart disease and hypertension is smaller than 0.05. Probit Regression is an effective model for analyzing the correlation between hypertension and heart disease with stroke, and this study can further infer what factors may be associated with stroke. Compared to previous studies, they tend to use specific analysis of a single factor, which is limited to known possible factors for developing stroke, such as smoking, genetics, and air quality of the patient's living environment. In contrast, the factors covered in this study are more extensive and comprehensive, which can not only effectively avoid the errors caused by not controlling for a single variable, but also broaden the ideas of future research on stroke, helping medical staff to identify more directions for treatment and detect stroke earlier so that timely treatment can be obtained, because the result of stroke patients who failure receiving timely treatment is very serious even deadly.

4. Conclusion

The current study selects diverse data and focuses on influencing factors that may be associated with developing stroke. It was concluded that having stroke may be related to hypertension and heart disease.

It cannot be denied that due to the limited amount of data, this model may have errors in addition to the factors, and the sample may cover only a small amount of factors. However, the study still has a lot of value and merits. First of all, the approach taken in this study is innovative. On the one hand, a graphical approach was used to visually analyze the differences in the proportion of each factor in the population with stroke and the population without stroke. This enables visualization of the experiment and makes the results clearer and more intuitive. On the other hand, instead of using a single-factor analysis method as in many previous experiments, multiple linear regression was chosen for this experiment, making it more comprehensive. Secondly, it has some positive effects on stroke treatment. In addition to the factors known to be associated with stroke such as smoking, more factors may be related to stroke that deserve attention and consideration, such as heart disease, hypertension

so on. Whether these factors are associated with the development of stroke requires further medical investigation, which means that these findings will point the way to further relevant research in the future. Once a new causative factor other than the one already identified is discovered, this can help people detect stroke earlier, and treat it as soon as possible, improving survival rates and patients' quality of life.

References

[1] Ge Ying. Causes and Symptoms of Cerebral Stroke. Health Literature, 2020, 112.

[2] Jauch E C, et al. Guidelines for the Early Management of Adults with Ischemic Stroke. Circulation, 2013.

[3] Andersen K K, et al. Hemorrhagic and ischemic strokes compared: stroke severity, mortality, and risk factors. Stroke, 2009, 2068-2072.

[4] Evenson, et al. Concordance of stroke symptom onset time: The second delay in accessing stroke healthcare (DASH II) study. Ann Epidemenol, 2001.

[5] Evenson, Kelly R, et al. Concordance of Stroke Symptom Onset Time. Ann Epidemenol, 2001, 11(3): 202-207.

[6] Brunner, Hans R, et al. Essential hypertension: renin and aldosterone, heart attack and stroke. N Engl J Med, 1972, 441-449.

[7] Giuseppe, et al. Hypertension: a harbinger of stroke and dementia. Hypertension, 2013.

[8] Sidney, Stephen, et al. The Heart Disease and Stroke Statistics-2013 Update and the Need for a National Cardiovascular Surveillance System. Circulation, 2012, 21-23.

[9] Smith, Sidney C. Reducing the Global Burden of Ischemic Heart Disease and Stroke. Circulation, 2011.

[10] Fiebach N H. Blood pressure, stroke, and coronary heart disease. Part 2, Short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context. Lancet, 1990, 827-838.

[11] Pearson, Thomas A, et al. AHA Guidelines for Primary Prevention of Cardiovascular Disease and Stroke: 2002 Update: Consensus Panel Guide to Comprehensive Risk Reduction for Adult Patients Without Coronary or Other Atherosclerotic Vascular Diseases. American Heart Association Scienc. Circulation, 2002.

[12] VOGT. Social networks as predictors of ischemic heart disease, cancer, stroke and hypertension: incidence, survival and mortality. Journal of Clinical Epidemiology, 1992, 659-666.