MATH 560 - Project

Rohit Wason

4/19/2021

Abstract

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Here we use a public dataset of patients to state hypotheses of correlation between various explanatory variables (like, age, bmi, marital status, residence type, etc.) along with the fact that they have had a stroke.

The Dataset

The dataset we use here can be found on Kaggle, a public competetion forum where data scientists and novices collaborate to find answers in complex datasets (kaggle.com/fedesoriano 2021). This dataset contains 5110 data points with 12 variables:

- **id** Unique Identifier. (We will ignore this variable, as we're not interested in the individual patient's details)
- **gender** (”Male”, ”Female” or ”Other”)
- **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”)
- **work__type** (“children”, “Govt__jov”, “Never__worked”, “Private” or “Self-employed”)
- **Residence__type** (“Rural” or “Urban”)
- **avg__glucose__level** (Average Glucose level in blood)
- **bmi** (Body Aass Index)
- **smoking__status** (“formerly smoked”, “never smoked”, “smokes” or “Unknown”)
- **stroke** (1 if the patient had a stroke or 0 if not)

Data Preparation

As is usually the case with real data, there are inconsistencies in this dataset. For example, *numerical* variables, like “bmi” could be classified as *categorical*. Such inconsistencies can hamper our research, hence are cleaned-up as a first step:

We start with analysing the “class” of each variable:

Variable	Class
id	Integer
gender	Categorical
age	Numeric
hypertension	Integer
heart_disease	Integer
ever_married	Categorical
work_type	Categorical
Residence_type	Categorical
avg_glucose_level	Numeric
bmi	Categorical
smoking_status	Categorical
stroke	Integer

We note that the “bmi” variable is classified as “categorical” because there are values like “N/A” for some items. We filter those out:

```
> stroke=stroke[stroke$bmi!=’N/A’,]
> stroke=transform(stroke, bmi=as.numeric(bmi))
> dim(stroke)
[1] 4909   12
```

As a result, out of 5110 data-points, only 4909 are rendered “useful”. We consider this enough data for our research and mark the interesting “numeric” variables: **age**, **bmi** and **average__glucose__level**.

The Analysis

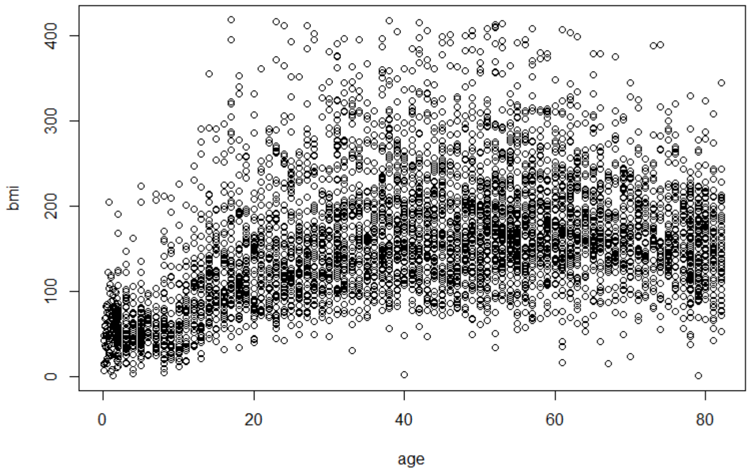
According to stroke.org, a non-profit organization dedicated to public awareness about, and prevention of common causes of stroke, smoking, lack of physical activity, diabetes and obesity are leading factors that could lead to stroke (stroke.org 2021). An important question to analyze would be “Can we predict the onset of stroke given any/all of the variables in this dataset?”

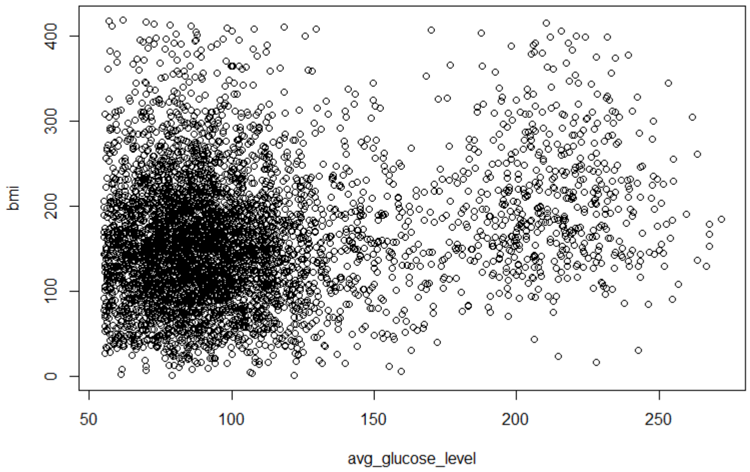
However, since “stroke” is a “categorical” variable, and we want to limit our research to “numeric” values only, we explore the question:

Is “bmi” correlated to any other “numeric” value?

It is customary while analyzing data, to “slice and dice” data in different ways. One such way is to find correlated variables, so their presence doesn’t falsly influence the dependent variable. Therefore, while it might be possible to answer more important questions using this data, the present question is of some value in the big picture.

At this point we plot "bmi" against other "numeric" variables. As a convention, the independent variable is shown horizontally (on the **x-axis**), and the dependent variable, vertically (on the **y-axis**).





Graphically, there seems to be some correlation between "bmi" and each of the "average_glucose_level" and "age". It seems plausible that the latter two variables, also known as **independent variables** could affect the former, the **dependent variable**.

Data Modelling

The technique of Linear Regression is common in research of this type – the dependent variable (“bmi”, in this case) can be modelled with the independent ones. This technique is called “fitting”, and results in a linear equation involving all the variables involved of the kind

$$\hat{y} = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + \cdots + \beta_p(x_p) \tag{1}$$

where β_i are called the **regression coefficients** corresponding to the **independent variables**, x_i $(i = 1, 2, 3, \ldots, p)$.

The **R programming language** offers a powerful way to “fit” this model:

```
> lmStroke=lm(bmi~age+avg_glucose_level , data=stroke)
```

What this created is a “linear model” called “lmStroke” using the 3 interesting variables, for the 4909 data-points. The model could be seen as a projected line in a 3-dimensional space (a dimention representing each of the variables we’re interested in). And what’s fascinating is that these models do not have to be limited to 3-dimentional space only. Most real-data consist of hundreds, or thousands of variables and can be “fitted” in the same way our model is!

We now observe the summary of this model:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        96.90234    2.91577  33.234  < 2e-16
age                 1.08004    0.04556  23.705  < 2e-16
avg_glucose_level   0.17939    0.02313   7.755 1.07e-14
```

One important characteristic of this model is the coefficients (β_i) for each independent variable. Using the specific values of this model we get our version of equation (1) above as:

$$\hat{y} = 96.9023 + 1.0800(x_1) + 0.1793(x_2) \tag{2}$$

which is powerful, since if the independent variables, “age” and “avg_glucose_level” (x_1, x_2 here) correlate with the independent variable, “bmi”, they can be used to predict “bmi” ($\hat{y}$ is the predicted value of y).

ANOVA F-test

An important analysis of this model is called Analysis of variance (ANOVA). It can determine whether the means of three or more groups are different. ANOVA uses **F-tests** to statistically test the equality of means.

The reason this analysis is important is to statistically rate variables on their “importance” in predicting the independent variable.

Hypotheses

We test the **null hypothesis**, H_0 that all independent variables are insignificant in determining the dependent variables. In other words, their coefficients $\beta_1 = \beta_2 = 0$. This is tested against the alternative hypothesis, H_a that such is not the case.

```
Residuals:
    Min      1Q  Median      3Q     Max
-194.40  -49.45  -13.30   35.90  291.67

Residual standard error: 69.96 on 4906 DF
Multiple R-squared:0.1327, Adjusted R-squared:0.1323
F-statistic: 375.2 on 2 and 4906 DF,  p-value:<2.2e-16
```

On examining the above summary of our model, we see that the test statistic

$$F = \frac{MSM}{MSE} = 375.2,$$

which when compared to

$$f^* = 39.4$$

is much greater. Hence we **reject** the hypothesis that all independent variables are insignificant in determining the dependent variables. In other words, the independent variables, “age” and “average_glucose_level” cannot be discarded as playing a role in determining “bmi”.

Summary

We started with examining a health-related dataset from the public domain. We also started with the assumption that the data collection is unbiased and fair.

After an initial cleanup of some anomalies, the data seemed fairly useful with the "numeric" variables. Especially of interest were the "bmi", "age" and "average_glucose_level" variables and we grew interested in finding if these variables exhibit any correlation.

On regressing the latter two variables to "fit" the value of "bmi", we observe that the correlation is significant and that **the two independent variables cannot be discarded, while predicting the "bmi".**

Works Cited

kaggle.com/fedesoriano. “Stroke Prediction Dataset”. 2021. Web. 31/03/2021. <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv>.

stroke.org. “Stroke Risk Factors You Can Control”. 2021. Web. 31/03/2021. <https://www.stroke.org/en/about-stroke/stroke-risk-factors/stroke-risk-factors-you-can-control-treat-and-improve>.