Rohit Wason
Spring 2021

Math 560 Homework (#10, Regression)

Problem 1. Given:
$n = 500$
Mean height of the fathers, $\bar{x} = 67.9$ (explanatory vaiable)
Standard deviation of the height of fathers, $s_x = 2.75$
Mean height of the sons, $\bar{y} = 68.7$ (response vaiable)
Standard deviation of the height of sons, $s_y = 2.83$
The correlation between the heights of the sons and their fathers, $r = 0.5$.

Solution. (a) Give a 99% prediction interval for the height of a son whose father is 70 inches tall.
Using $b_1 = r(s_y/s_x)$ and $b_0 = \bar{y} - b_1\bar{x}$ we have that

$$b_1 = 0.5(2.83/2.75) \approx 0.5145$$

$$b_0 = 68.7 - 0.5145(67.9) \approx 33.7654$$

□

Problem 2.

Year(x)	Kilobits(y)
1971	1
1980	62.5
1987	1000
1993	16000
1998	125000
2000	250000
2002	500000
2004	976562.5

Solution. We build the model

```
lmKilo=lm(Kilobits~Year, data = dram)
lmKilo$coefficients
> (Intercept)        Year
> -40886934.88      20644.12
```

to receive the coefficients $b_0 = -40886934.88, b_1 = 20644.12$. Therefore the equation of the least-square regression line is

$$\hat{y} = -40886934.88 + 20644.12(x)$$

$s_x = 11.6795$
$s_y = 347560.1$
Using $b_1 = r(s_y/s_x)$ and $b_0 = \bar{y} - b_1\bar{x}$ we have that □

Project details

Introduction

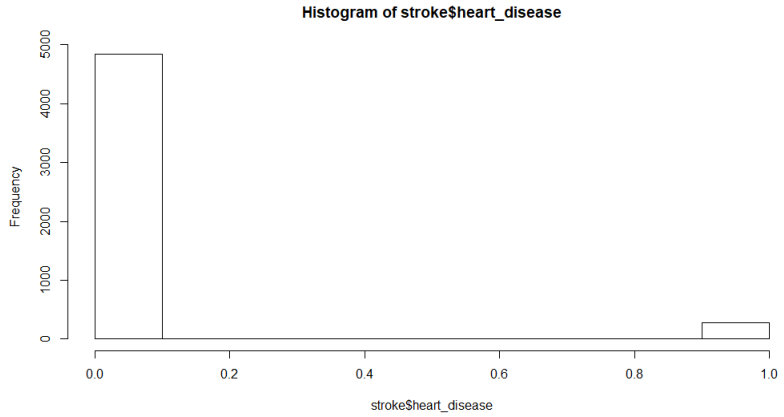
According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking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

1. **id:** unique identifier
2. **gender:** "Male", "Female" or "Other"
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" or "Unknown"*
12. **stroke:** 1 if the patient had a stroke or 0 if not

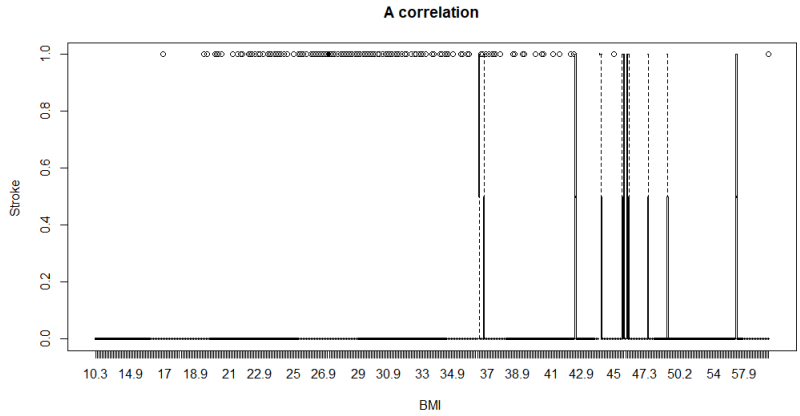
Methods

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).
In particular the following questions are of interest:

1. **Is there a strong correlation between stroke and other conditions, like heart disease, or hypertension?**
   A histogram of whether heart-disease is present in a data point:



2. **Does BMI indicate the occurence of stroke?**



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>.