MA930 Assignment 2

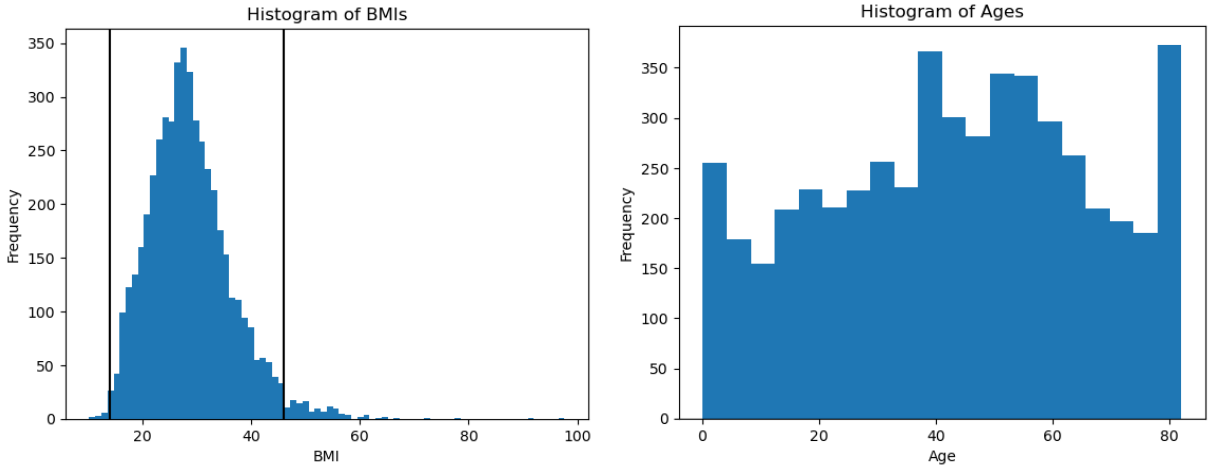
October 2023

Data

My dataset is from Kaggle [1], contains 5110 individuals, and has information that allows predictions about strokes based on other variables like gender, age, various diseases, and smoking status. I have chosen to only look at gender, age and BMI(Body Mass Index) and how they impact stroke probability.

Data Processing

For my analysis having a high density of individuals across my data range was important as it gives a higher confidence. For this reason I restricted my BMI dataset to only include those individuals above 14 and below 46 BMI (shown with black bars in the left histogram below) but for age I didn't need to restrict the data (shown in the right histogram). I also removed all individuals with N/A entries in areas used in my analysis.



Are Males and Females equally likely to have had a stroke?

My null hypothesis is that the set of rv's of male strokes, $X_1, ..., X_{n_{male}}$, is distributed with the same mean as the set of rv's of female stokes, $Y_1, ..., Y_{n_{female}}$. The mean of a Bernoulli rv is p therefore the mean of each set of rv's is equal to the probability of having had a stroke. As $n_{male} > 30$ and $n_{female} > 30$ CLT applies to $\bar{X}$ and $\bar{Y}$, therefore $\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$ where we will approximate the variance of the underlying distributions with the sample variance. Using a 2 tailed hypothesis test we get our p-value $= 2P(\bar{X} - \bar{Y} > \bar{x} - \bar{y}) = 2P(Z > 0.597...) = 0.55$, so we do not reject the null hypothesis for any reasonable significance level and hence say (at least according to this data set) that Males and Females are equally as likely to have had a stroke.

What is the relationship between the probability of having had a stroke and BMI?

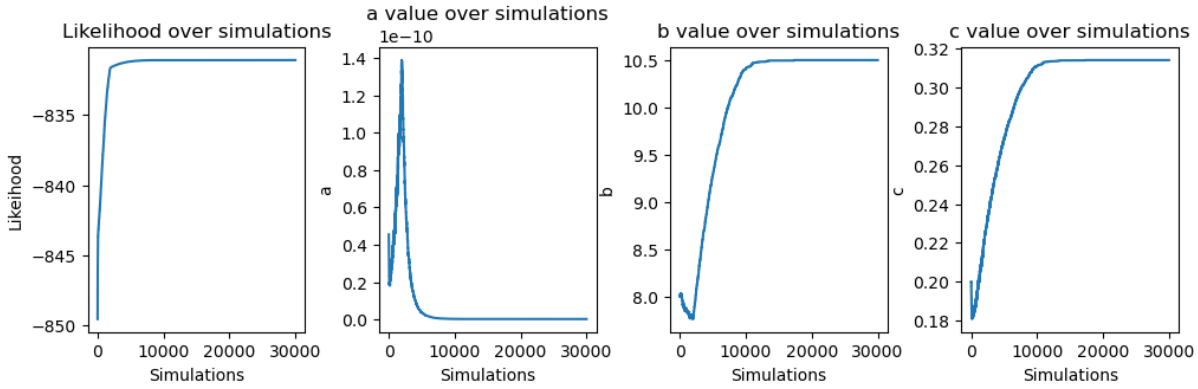
Taking the stroke outcome, x_i, of each individual in the data set as a Bernoulli rv, $X_i \sim Be(p_i)$, then by using each individual's BMI value, y_i, I assume that there is some function, $f(y_i)$ mapping $y_i \to p_i$, then follows $X_i \sim Be(f(y_i))$

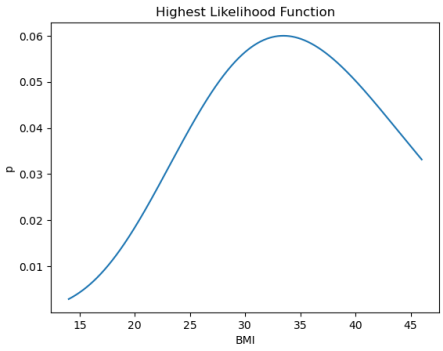
I assumed my function takes the form $ay_i^b e^{-cy_i}$ due to this function always being above 0 and its ability to take on many different shapes within my restricted BMI interval such as linear, exponential and skewed bell

curve giving me confidence that with the correct parameters it would fit the data well. To find the correct parameters I maximised the likelihood function

$$\mathcal{L}(a,b,c|x_1,...,x_n) = P(x_1 = 1|X \sim Be(ay_1^b e^{-cy_1}))...P(x_n = 0|X \sim Be(ay_n^b e^{-cy_n})) = (ay_1^b e^{-cy_1})...(1 - ay_n^b e^{-cy_n})$$

This is difficult to evaluate analytically so instead I used a 3-variable MCMC simulation to find the parameters which maximised the likelihood function. These values were: a = 2.12×10^{-13}, b = 10.5, and c = 0.314 (all 3 s.f.)

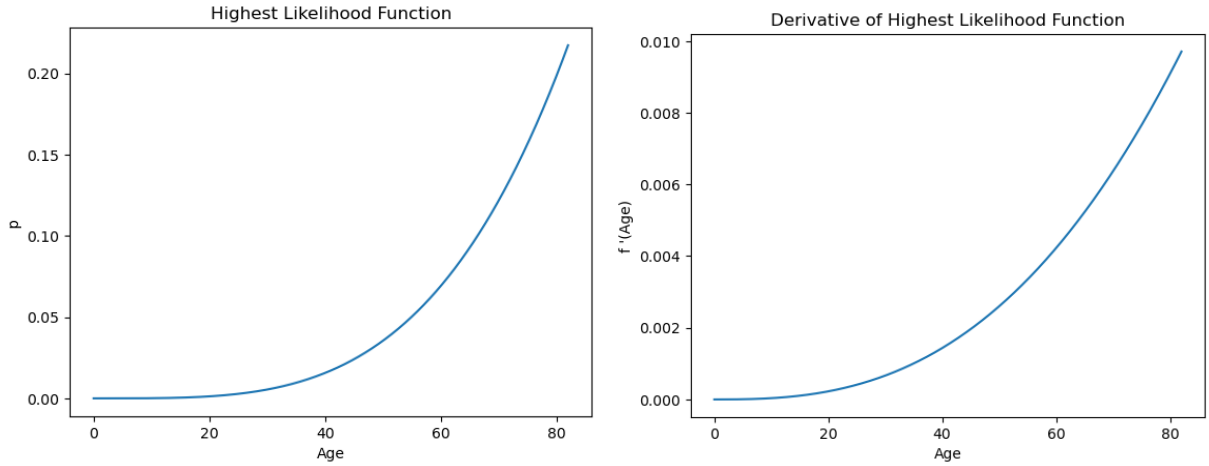




The left figure shows the function f(y) with the most likely parameters of a, b, and c. From this we can see that the probability of having had a stroke fits a quasi-bell curve where BMI increases up until the maximum risk of 32, where you are 3.3 times as likely to have had a stroke as someone with a BMI of 20 (which is in the healthy range [2]), but then interestingly declines as BMI increases beyond this. This could be explained by individuals with morbid obesity being motivated to reduce their BMI after the stoke in order to reduce the chance of having another or more complex biological factors offering very high BMI individuals some stroke protection. This 'obesity paradox' has been shown in other studies [3]

What is the relationship between the probability of having had (and having) a stroke and age?

Here I used the same approach I used to compare BMI and stroke risk but with y_i as age and $f(y_i)$ mapping age $\rightarrow p_i$.
MCMC simulation gave paramter values: a = 2.04×10^{-8}, b = 3.67, and c = 9.99×10^{-5} to maximise the likelihood function.



The left figure shows the function f(y) with the most likely parameters of a, b, and c. From this we can see that the probability of having had a stroke rises quasi-exponentially with age. Given the probability of an individual having had a stoke, p_i, = $\int_0^{Age_i} \frac{dp}{dAge}$ dAge and our function is a function of Age we are able to plot the derivative, $f' = ay^{b-1}e^{-cy}(b - cy)$ shown on the right figure, against age, giving us a plot of the rate of having a stroke against Age. This allows us to compare immediate risk of stroke amongst ages, for example a 60 year old is 6.3x more likely to have a stroke than a 30 year old.

References

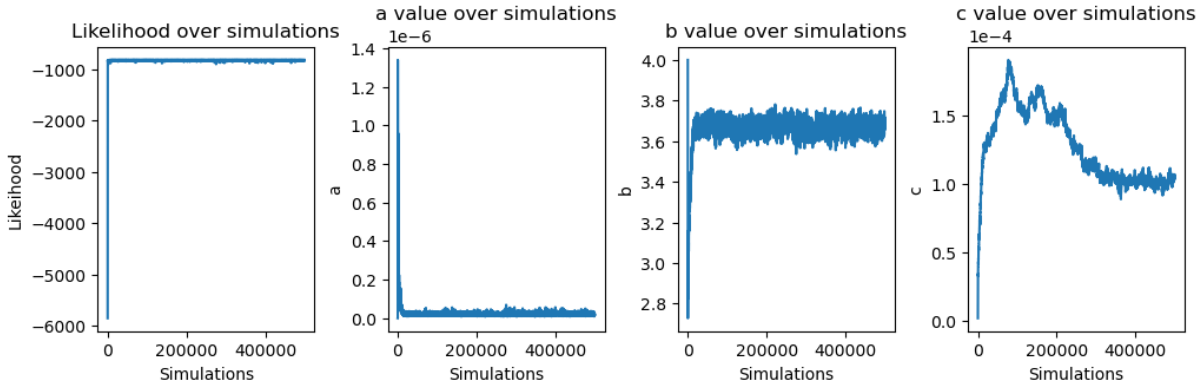
[1] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[2] https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/#:~:text=BMI%20ranges&text=below%2018.5%20%E2%80%93%20you're%20in,re%20in%20the%20obese%20range

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7946563/

Appendix

Plots of parameters and likelihood, for age based stroke risk analysis, during MCMC simulation



Code Url: https://github.com/MaxDButler/MA930Assignment2.git