Introduction to Logistic Regression

April 21, 2020

Logistic Regression

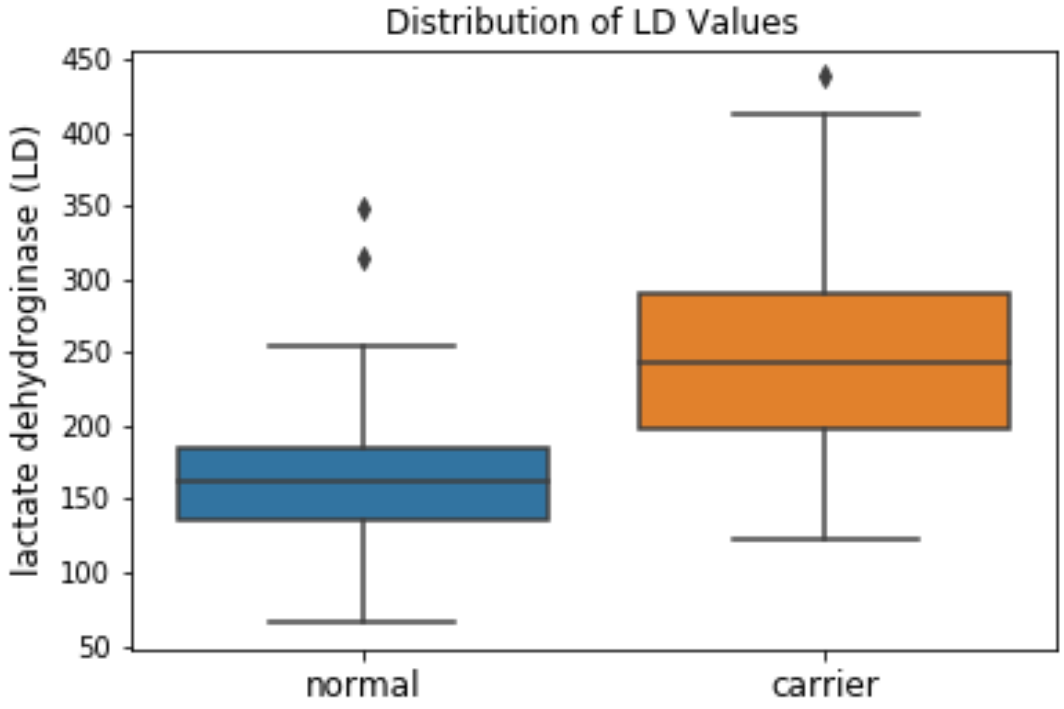
Example: Duchenne Muscular Dystrophy (DMD)

- Genetically-transmitted disease
- Passed from a mother to her children
- Female offspring suffer no apparent symptoms, but male offspring with the disease die at a young age.
- Female carriers tend to exhibit elevated levels of certain serum enzymes or proteins.

Let's say we want to build a model which takes as input the serum levels and outputs our prediction about whether or not a female is a carrier.

Logistic Regression

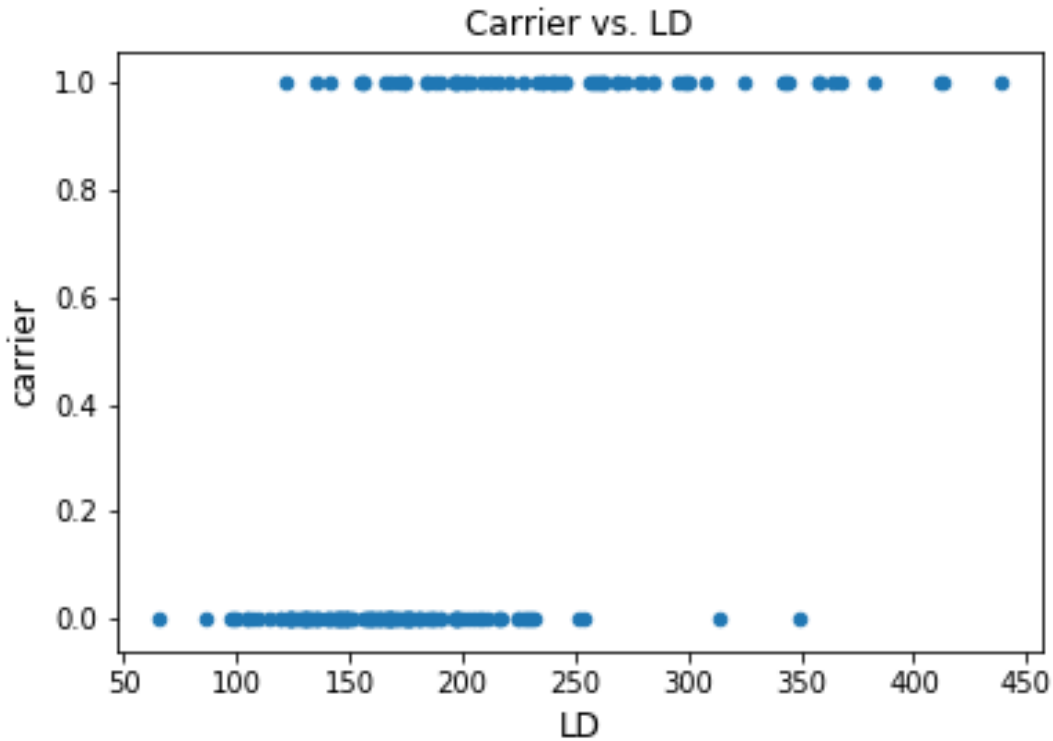
We can start be doing some exploratory analysis.



What we can see is that in our datset, carriers tend to have higher LD values. However, there is some overlap between carriers and non-carriers in the middle values. There is not a single cutoff we can use to classify a person as a carrier or non-carrier.

Logistic Regression

Here is another view of the data, using a scatterplot



Here, we have encoded whether a person is a carrier or not using a numeric 0/1 value. A value of 1 indicates that a person is a carrier.

Logistic Regression

Since there is overlap between carriers and non-carriers, we would probably be best off to not just make a simple prediction of carrier/non-carrier, but instead predict the likelihood or probability that a female is a carrier.

From what we have seen in the plots, females with higher LD values look more likely to be carriers that those with lower values.

Between 175 and 225, it is not clear if a person is a carrier or not, so we might be best off assigning a probability close to 0.5 for people in that range.

Logistic Regression

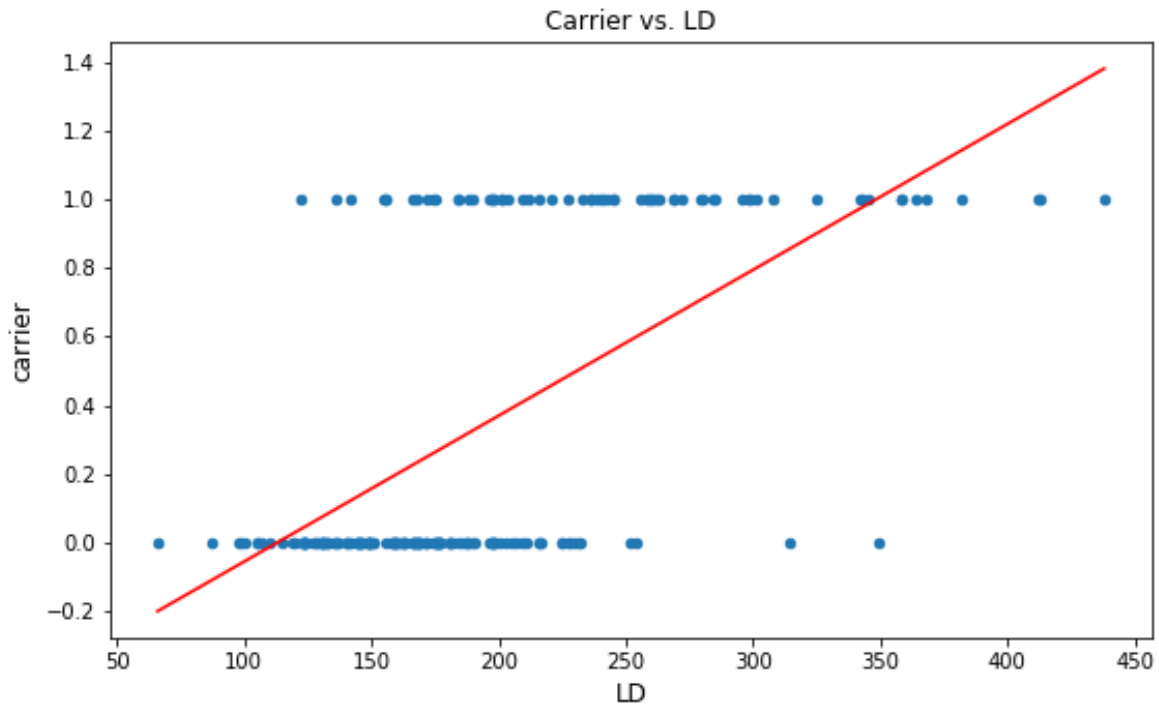
So how do we create our model? We can try using a linear regression model.

Logistic Regression

So how do we create our model? We can try using a linear regression model.

A linear regression model produces the following result:

$$\mathrm{P}(\text{carrier}) = 0.00426 \cdot (\mathrm{LD}) - 0.4829$$



Logistic Regression

This approach has a big problem: probabilities are values between 0 and 1, but this equation has guarantee of outputting values between 0 and 1.

In fact, we can see that for some values, we get predictions less than 0 or greater than 1.

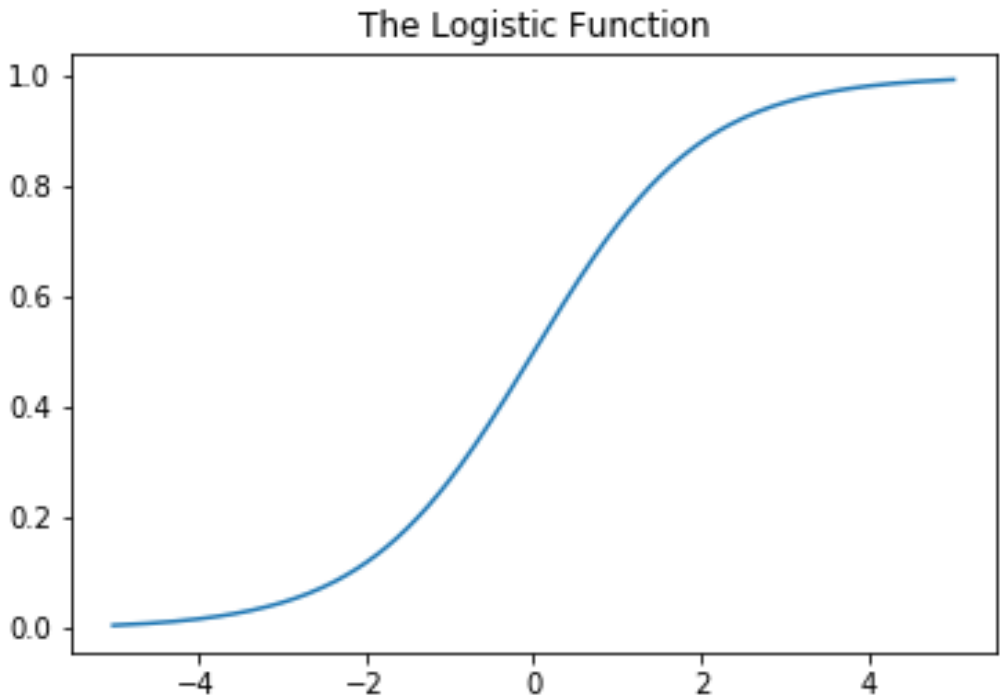
Another problem is that it assumes a fixed change in LD will have a fixed effect on the probability. That is, a change from 60 to 70 will have the same impact as a change from 250 to 260.

Logistic Regression

One possible solution to this problem is to "squash" our output between 0 and 1.

A common way to do this is to pass the output from a linear model into the **logistic function**:

$$l(x) = \frac{1}{1 + e^x}$$



Logistic Regression

This means that instead of our model looking like

$$\mathrm{P}(\text{carrier}) = \beta_1 \cdot (\mathrm{LD}) + \beta_0$$

it will look like

$$\mathrm{P}(\text{carrier}) = \frac{1}{1 + e^{-(\beta_1 \cdot (\mathrm{LD}) + \beta_0)}}$$

Logistic Regression

Fitting this model, we obtain

$$\mathrm{P}(\text{carrier}) = \frac{1}{1 + e^{-(0.02905 \cdot (\mathrm{LD}) - 6.407)}}$$