DATS 6101 Introduction to Data Science

Logistic Regression

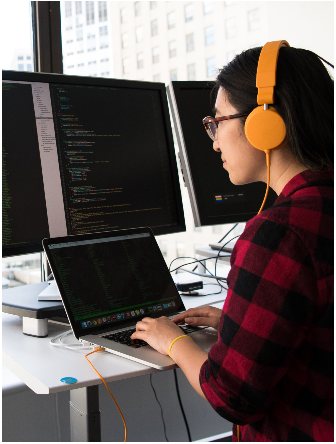
Omer Yalcin

github.com/omerfyalcin/logisticRegression

May 12, 2021

Motivation

Conservation scientist studying a painting [Photo by Richard McCoy from Wikimedia Commons]



Engineer building a credit card fraud detection system [Photo by Christina Morillo from Pexels.]

Logistic Regression

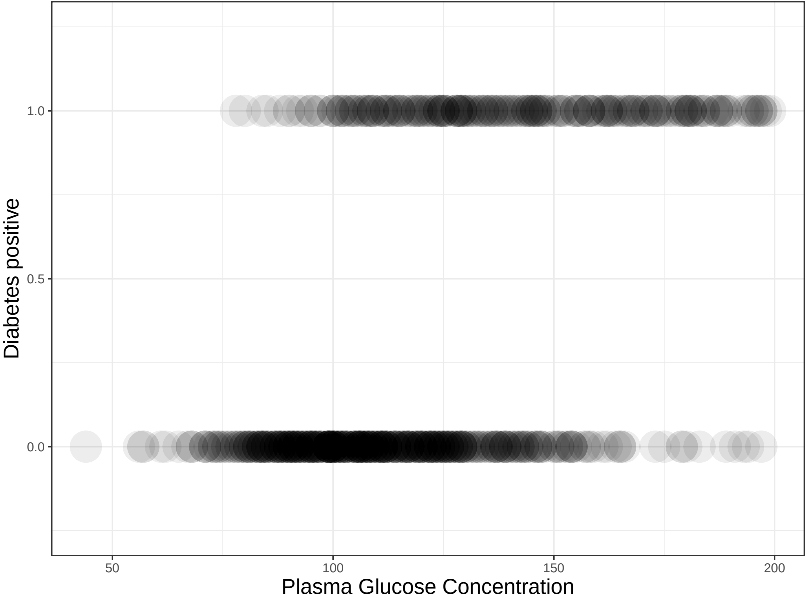
a classification algorithm for **binary outcome variables**

1. A real world problem: diabetes prediction
2. Linear Regression: Solution?
3. Logistic Regression: Extension to $Y \in \{0, 1\}$
4. logit & sigmoid functions
5. Maximum Likelihood: Intuition
6. Fitting a logistic regression model in R

Diabetes Detection

- from National Institute of Diabetes and Digestive and Kidney Diseases (provided by the *mlbench* package in R)
- 768 native American women of the Pima heritage
- age 21 or older
- **outcome variable**:
  - positive (1) or negative (0) for diabetes
  - 268 positive, 500 negative
- **explanatory variables:**
  - plasma glucose concentration (mg/dL)
  - body mass index (kg/m^2)

Plasma Glucose Concentration and Diabetes



Can we use linear regression?

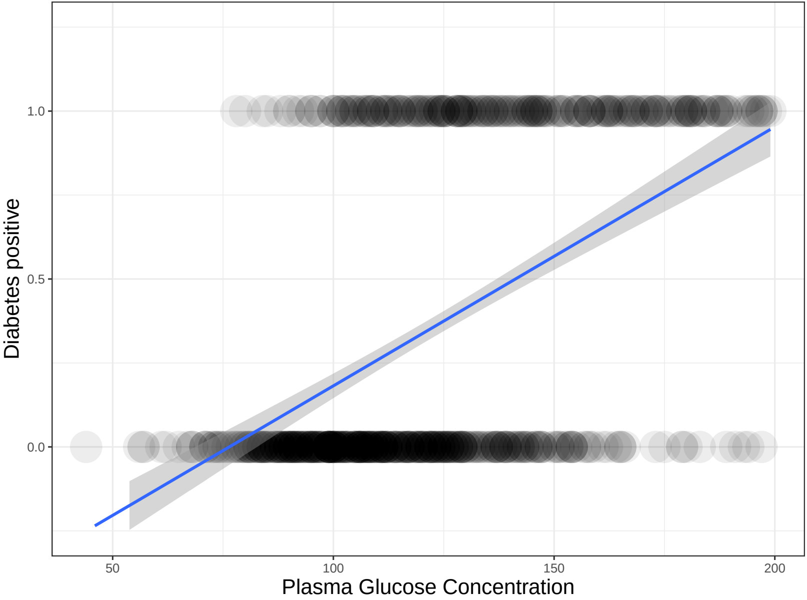
$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} + \epsilon_i$$

$i \in \{1, 2, ..., n\}$: observations

k: the number of explanatory variables

Y_i is in range $(-\infty, \infty)$

Linear Regression



Logistic Regression: Extension to dichotomous Y

Problem: $Y_i \in \{0, 1\}$, and $0 < E[Y_i] < 1$

Solution:

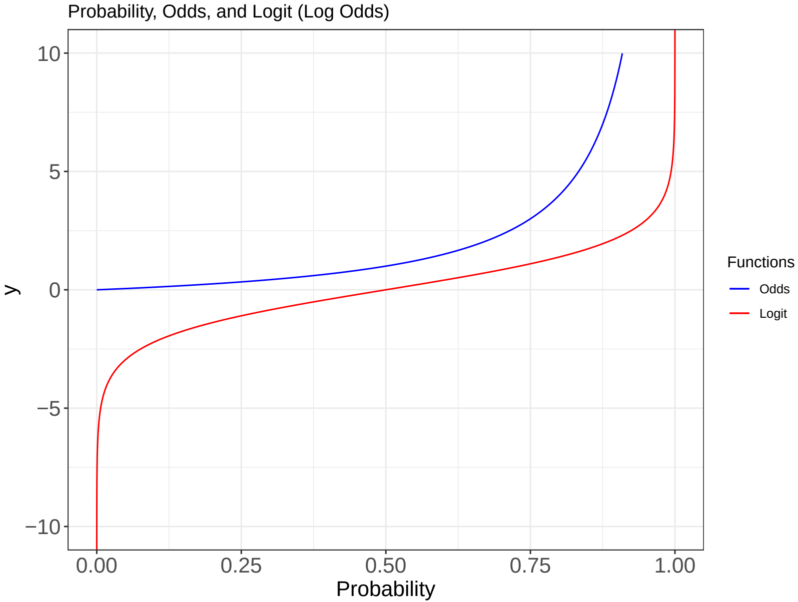
- $p_i = Pr(Y_i = 1)$
- transform p_i so that p_i is in range $(-\infty, \infty)$

Logit Function

$$\underbrace{\boxed{\ln}}_{\text{log}} \left[\underbrace{\boxed{\frac{p_i}{1-p_i}}}_{\text{odds}} \right] = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}$$

- β = constant change in log-odds
- $\exp(\beta)$ = odds ratio, i.e. $\frac{\text{odds}(X_j+1|X_1,...X_k)}{\text{odds}(X_j|X_2,...X_k)}$

Logit Function



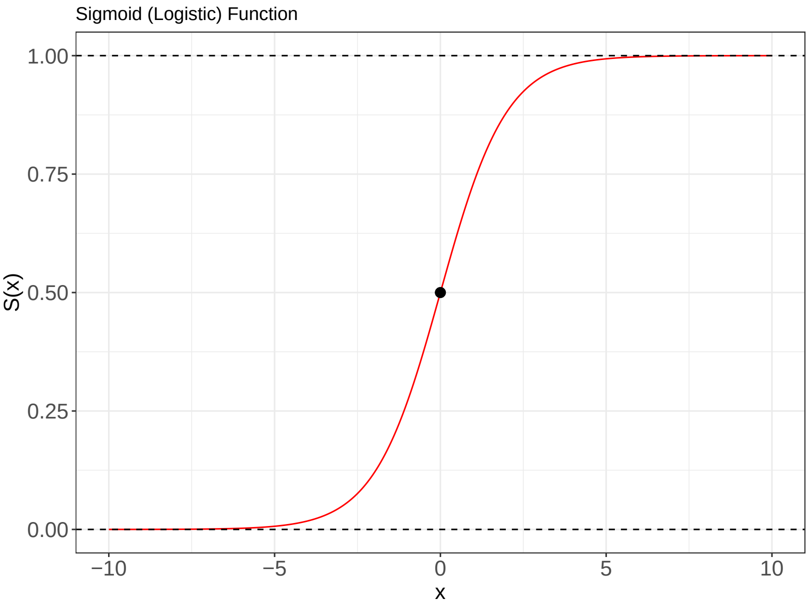
Sigmoid (Logistic) Function

- once we get a predicted log-odds value, plug that back into sigmoid function to get p_i

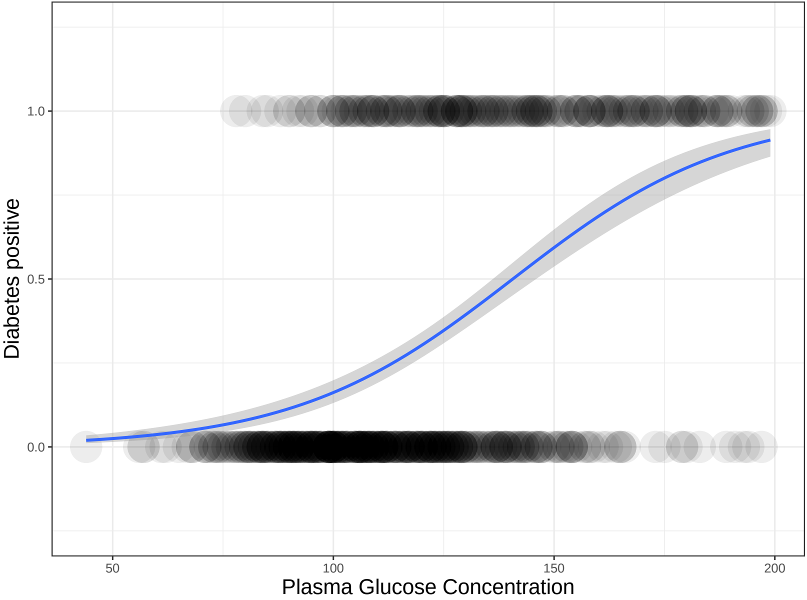
$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

$$p_i = \frac{1}{1 + e^{[-(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik})]}}$$

Sigmoid (Logistic) Function



Logistic Regression



Maximum Likelihood: Intuition

$$Pr(Y_i = 1|X_i) = \frac{1}{1 + e^{[-(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik})]}}$$

Y_i	$Pr(Y_i = 1\|X_i)$	$Pr(Y_i = 0\|X_i)$
1	p_1	$1 - p_1$
0	p_2	$1 - p_2$
1	p_3	$1 - p_3$
1	p_4	$1 - p_4$
0	p_5	$1 - p_5$
0	p_6	$1 - p_6$

$$\mathcal{L}(Y, X|\beta) = (p_1)(1 - p_2)(p_3)(p_4)(1 - p_5)(1 - p_6)$$

find $\beta_0, \beta_1, \ldots, \beta_k$ that maximizes $\mathcal{L}(Y, X|\beta)$

Implementing in R

You can follow along with the **logisticRegresssion.Rmd** file in https://github.com/omerfyalcin/logisticRegression