Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Introduction

Georgia Tech

About This Lesson

Georgia Tech

Yes/No Questions

- How likely is it that users will like a new layout of our website?
- Will my customers leave my wireless service at the end of their subscription?
- What financial characteristics can be used to predict whether or not a business will go bankrupt?

➔ **Model the probability of 'Yes'**

Georgia Tech

Linear Regression

Model: $\boldsymbol{Y}_i = \beta_0 + \beta_1 \boldsymbol{X}_{i,1} + \beta_2 \boldsymbol{X}_{i,2} + \cdots + \beta_p \boldsymbol{X}_{i,p} + \varepsilon_i \; i = 1, \cdots, n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $\mathrm{E}(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent random variables
- *Normality Assumption*: $\varepsilon_i \sim \mathrm{Normal}$

Georgia Tech

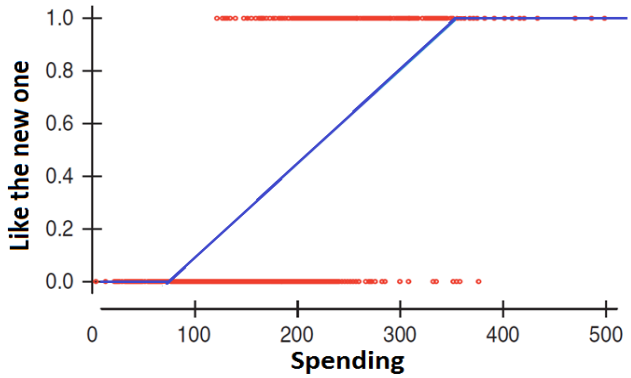
Linear Regression for Yes/No Question?

- Uber recently changed their logo.

- You are asked to model whether Uber users will like the new logo based on how much they spent in the last 3 months using Uber.

Georgia Tech

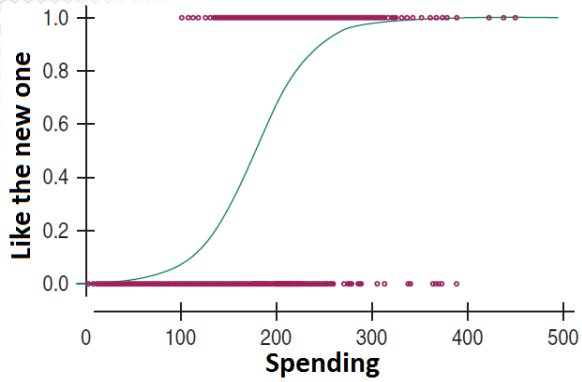
What Is Wrong with Linear Regression?



Customers will not behave like this!

Georgia Tech

S-shaped Curve



Logistic Regression Model

Data: $\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Model: We model the *probability of success given the predictor(s)*

$$\mathrm{p} = \mathrm{p}(X_1, \cdots, X_p) = \Pr(Y = 1 \mid X_1, \cdots, X_p)$$

by linking p to the predicting variables through a nonlinear *link function* g:

$$\mathrm{g}(\mathrm{p}) = +\mathrm{g}\left(\mathrm{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

There is no error term!
What are the model assumptions?

Logistic Regression Model

Data: $\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Assumptions:

- *Linearity Assumption*: $\mathrm{g}\left(\mathrm{p}\left(X_1, \cdots, X_p\right)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- *Independence Assumption*: $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are independent random variables
- *Logit Link Function:*

$$\mathrm{g(p)} = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right)$$

Georgia Tech

Summary

Georgia Tech