

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Introduction



About This Lesson



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'**



Linear Regression

Model: $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots + \beta_p X_{i,p} + \varepsilon_i \quad i = 1, \dots, n$

Assumptions:

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



Linear Regression for Yes/No Question?

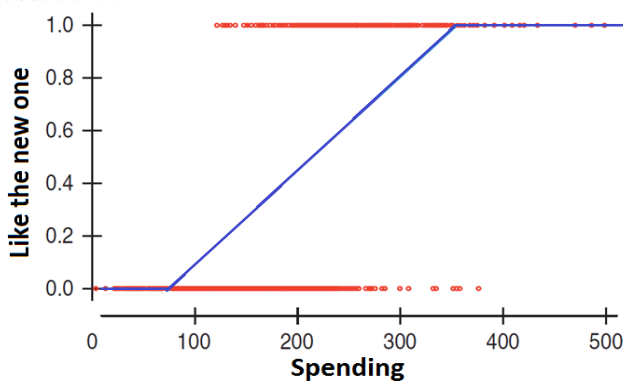
- Uber recently changed their logo.



- 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

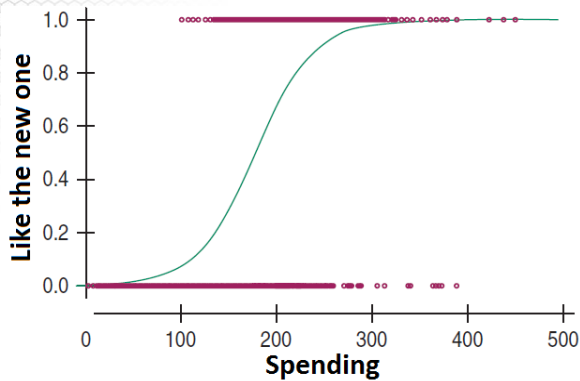
What Is Wrong with Linear Regression?



Customers will not behave like this!

Georgia
Tech

S-shaped Curve



Logistic Regression Model

Data: $\{(X_{1,1}, X_{1,2}, \dots, X_{1,p}), Y_1\}, \{(X_{2,1}, X_{2,2}, \dots, X_{2,p}), Y_2\}, \dots, \{(X_{n,1}, X_{n,2}, \dots, X_{n,p}), Y_n\}$
 where Y_1, \dots, Y_n are *binary* responses

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

$$p = p(X_1, \dots, X_p) = \Pr(Y = 1 \mid X_1, \dots, X_p)$$

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

$$g(p) = +g(p(X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

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

Logistic Regression Model

Data: $\{(X_{1,1}, X_{1,2}, \dots, X_{1,p}), Y_1\}, \{(X_{2,1}, X_{2,2}, \dots, X_{2,p}), Y_2\}, \dots, \{(X_{n,1}, X_{n,2}, \dots, X_{n,p}), Y_n\}$
 where Y_1, \dots, Y_n are *binary* responses

Assumptions:

- **Linearity Assumption:** $g(p(X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$
- **Independence Assumption:** Y_1, \dots, Y_n are independent random variables
- **Logit Link Function:**

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

Summary

