Introduction to Machine Learning for Social Scientists

Class 5: Intro to GLM

Edgar Franco Vivanco

Stanford University
Department of Political Science

edgarf1@stanford.edu

Summer 2018

Homework 3 Due Friday July 20th at midnight

Available tomorrow Start early!

In-class midterm Monday July 23rd

Final project

We'll talk more about this after midterm but you should start forming teams: 6 teams of 5 in total

All teams should be organized by Wed July 25th

Issues with the Room!

Next Monday old room, this one during the rest of the quarter

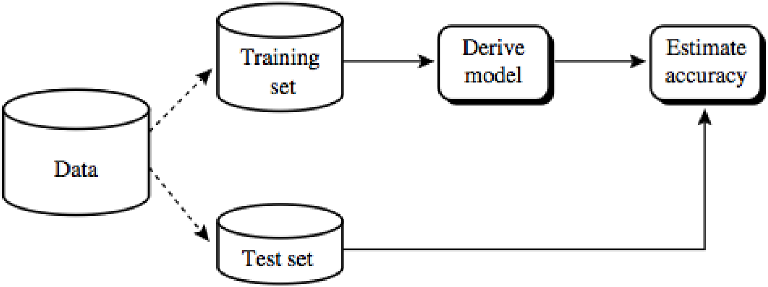
Today's Goals

1. Key concepts
    - Training and test sets
    - Linear Probability Model
    - Logit function and logit inverse function, logistic regression
2. Key techniques and R functions
    - glm
    - Natural logarithm, log
    - table

Divide your data

- **Training set:** A set of examples used to fit the parameters and learn. These are already classified by human coders, or produced in a semi-automated way, to create a 'gold standard'.
  Conventionally 80% of available data.
- **Test set:** A set that follows the same probability distribution and is used to test de model.
  Conventionally 20% of available data.

Divide your data



R!

Regression vs Classification

1. *Regression*
   - Quantitative responses
   - Example: Age, height, salary, price, vote share, etc.

Regression vs Classification

1. *Regression*
   - Quantitative responses
   - Example: Age, height, salary, price, vote share, etc.
2. *Classification*
   - Qualitative responses
   - Example: Election result (win, lose), fake news (yes, no, maybe).

Regression vs Classification

1. *Regression*
   - Quantitative responses
   - Example: Age, height, salary, price, vote share, etc.
2. *Classification*
   - Qualitative responses
   - Example: Election result (win, lose), fake news (yes, no, maybe).

NOTE: The distinction is not always that clear-cut: Logistic regression (a type of non-linear regression) is often used for classification.

Classification

For simplicity, in this class we are going to classify binary outcomes.

- Vote: Yay / Nay?

Classification

For simplicity, in this class we are going to classify binary outcomes.

- Vote: Yay / Nay?
- Email: Spam / Not Spam?

Classification

For simplicity, in this class we are going to classify binary outcomes.

- Vote: Yay / Nay?
- Email: Spam / Not Spam?
- Online transaction: Faudulent (Yes / No)?

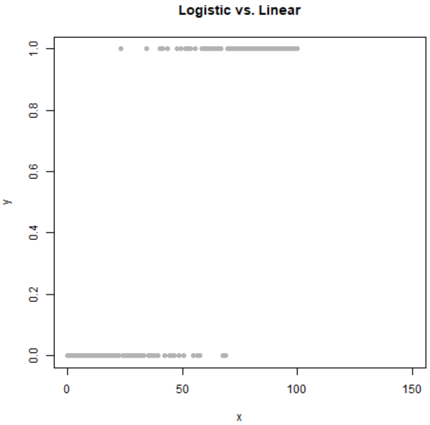
Classification

For simplicity, in this class we are going to classify binary outcomes.

- Vote: Yay / Nay?
- Email: Spam / Not Spam?
- Online transaction: Faudulent (Yes / No)?

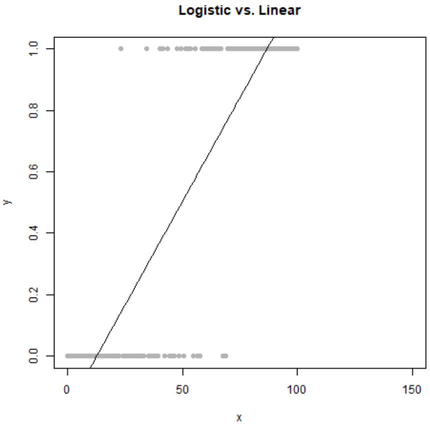
$$y \in \{0, 1\} \qquad \begin{bmatrix} 0: & \text{"Negative class"} \\ 1: & \text{"Positive class"} \end{bmatrix}$$

Predicting qualitative responses



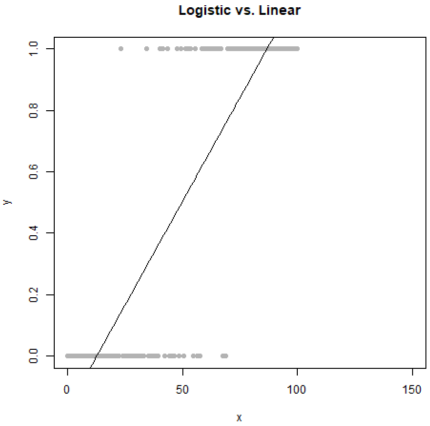
- $y \in \{0, 1\}$, X continuous

Predicting qualitative responses



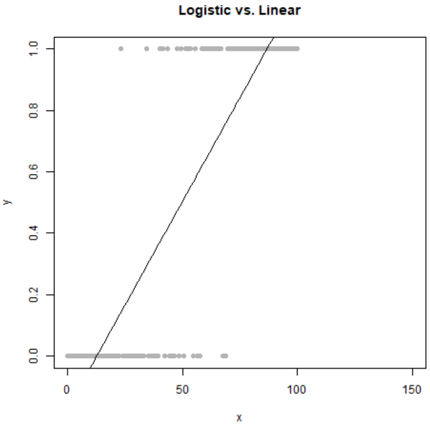
- $y \in \{0, 1\}$, X continuous
- We can always fit a linear model as we did before

Predicting qualitative responses



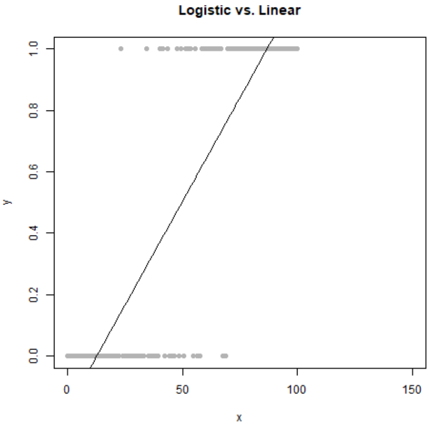
- $y \in \{0, 1\}$, X continuous
- We can always fit a linear model as we did before
- We call this the linear probability model because we can associate its predictions to probabilities.

Predicting qualitative responses



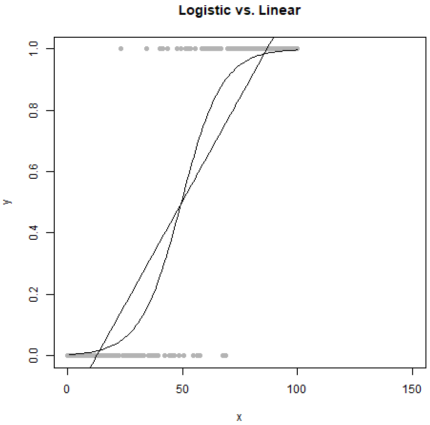
- $y \in \{0, 1\}$, X continuous
- We can always fit a linear model as we did before
- We call this the linear probability model because we can associate its predictions to probabilities.
- PROBLEMS?

Predicting qualitative responses



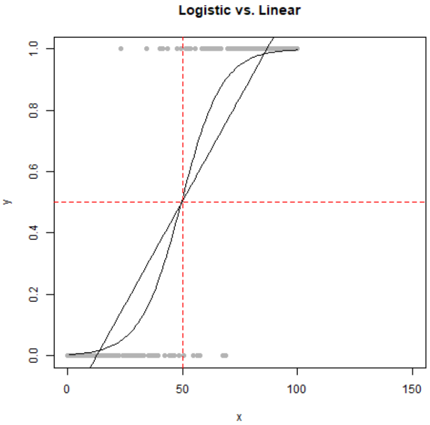
- $y \in \{0, 1\}$, X continuous
- We can always fit a linear model as we did before
- We call this the linear probability model because we can associate its predictions to probabilities.
- PROBLEMS?
- Predictions smaller than zero and larger than 1!

Predicting qualitative responses



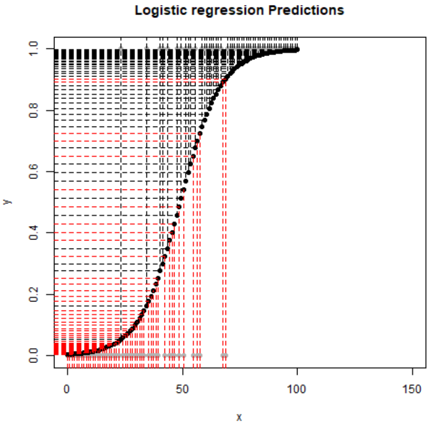
- Logistic model
- Estimates the probability of yes: $Pr(\hat{Vote_i} = 1|x_i)$ using a logarithmic transformation

Predicting qualitative responses



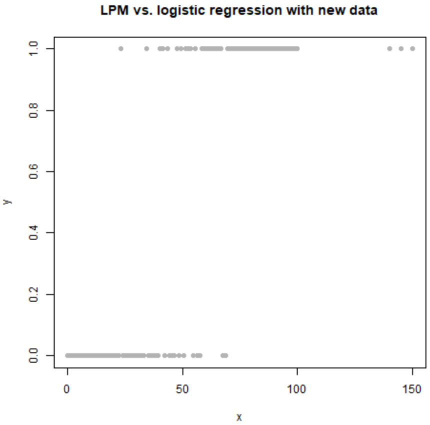
- Logistic model
- Estimates the probability of yes: $\hat{Pr(Vote_i = 1|x_i)}$
- The linear model predictions intersect when the probability equals 0.5.

Predicting qualitative responses



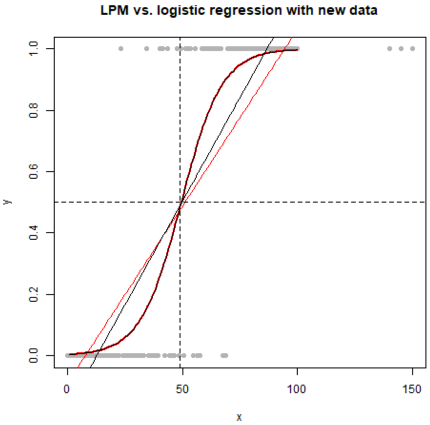
- Logistic model
- Estimates the probability of yes: $Pr(\hat{Vote_i} = 1|x_i)$
- The linear model predictions intersect when the probability equals 0.5.
- We can associate each point to a probability

Predicting qualitative responses



- Logistic model usually more stable to outliers

Predicting qualitative responses



- Logistic model usually more stable to outliers

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)
- $\log_{10} 1000 = 3$ (because $10^3 = 1000$)

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)
- $\log_{10} 1000 = 3$ (because $10^3 = 1000$)

Some rules of logarithms

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)
- $\log_{10} 1000 = 3$ (because $10^3 = 1000$)

Some rules of logarithms

- $\exp(\log(a)) = e^{\log(a)} = a$

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)
- $\log_{10} 1000 = 3$ (because $10^3 = 1000$)

Some rules of logarithms

- $\exp(\log(a)) = e^{\log(a)} = a$
- $\log(a \times b) = \log(a) + \log(b)$ (!!!!!!)

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)
- $\log_{10} 1000 = 3$ (because $10^3 = 1000$)

Some rules of logarithms

- $\exp(\log(a)) = e^{\log(a)} = a$
- $\log(a \times b) = \log(a) + \log(b)$ (!!!!!!)
- $\log(\frac{a}{b}) = \log(a) - \log(b)$

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)
- $\log_{10} 1000 = 3$ (because $10^3 = 1000$)

Some rules of logarithms

- $\exp(\log(a)) = e^{\log(a)} = a$
- $\log(a \times b) = \log(a) + \log(b)$ (!!!!!!)
- $\log(\frac{a}{b}) = \log(a) - \log(b)$
- $\log(a^b) = b \log(a)$

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)
- $\log_{10} 1000 = 3$ (because $10^3 = 1000$)

Some rules of logarithms

- $\exp(\log(a)) = e^{\log(a)} = a$
- $\log(a \times b) = \log(a) + \log(b)$ (!!!!!!)
- $\log(\frac{a}{b}) = \log(a) - \log(b)$
- $\log(a^b) = b \log(a)$
- $\log(1) = 0$

A Brief Reminder About (Natural) Logarithms

Logarithm log is a class of functions.

- $\log_e z = x$ if $e^x = z$.
- We'll call $\log_e$ natural logarithm. And we'll assume $\log = \log_e$
- $\log e = 1$ (because $e^1 = e$)
- $\log_{10} 1000 = 3$ (because $10^3 = 1000$)

Some rules of logarithms

- $\exp(\log(a)) = e^{\log(a)} = a$
- $\log(a \times b) = \log(a) + \log(b)$ (!!!!!!)
- $\log(\frac{a}{b}) = \log(a) - \log(b)$
- $\log(a^b) = b \log(a)$
- $\log(1) = 0$

Call $p_i = \Pr(\text{Vote}_i = 1|\boldsymbol{x}_i)$

Call $p_i = \Pr(\text{Vote}_i = 1|\boldsymbol{x}_i)$

$$\text{Vote}_i \sim \text{Bernoulli}(p_i)$$

Call $p_i = \Pr(\text{Vote}_i = 1 | \boldsymbol{x}_i)$

$$\begin{aligned} \text{Vote}_i &\sim \text{Bernoulli}(p_i) \\ p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \end{aligned}$$

Call $p_i = \Pr(\text{Vote}_i = 1 | \boldsymbol{x}_i)$

$$\begin{aligned}
\text{Vote}_i &\sim \text{Bernoulli}(p_i) \\
p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \\
\log\left(\frac{p_i}{1 - p_i}\right) &= \boldsymbol{\beta} \cdot \boldsymbol{x}_i
\end{aligned}$$

Call $p_i = \Pr(\text{Vote}_i = 1|\boldsymbol{x}_i)$

$$
\begin{aligned}
\text{Vote}_i &\sim \text{Bernoulli}(p_i) \\
p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \\
\log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{\beta} \cdot \boldsymbol{x}_i \\
p_i &= \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}
\end{aligned}
$$

Call $p_i = \Pr(\text{Vote}_i = 1|\boldsymbol{x}_i)$

$$
\begin{aligned}
\text{Vote}_i &\sim \text{Bernoulli}(p_i) \\
p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \\
\log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{\beta} \cdot \boldsymbol{x}_i \\
p_i &= \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)} \\
&= \frac{1}{1+\exp(-\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}
\end{aligned}
$$

Call $p_i = \Pr(\text{Vote}_i = 1 | \boldsymbol{x}_i)$

$$\begin{aligned}
\text{Vote}_i &\sim \text{Bernoulli}(p_i) \\
p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \\
\log\left(\frac{p_i}{1 - p_i}\right) &= \boldsymbol{\beta} \cdot \boldsymbol{x}_i \\
p_i &= \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)} \\
&= \frac{1}{1 + \exp(-\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}
\end{aligned}$$

Important functions:

Call $p_i = \Pr(\text{Vote}_i = 1 | \boldsymbol{x}_i)$

$$\begin{aligned}
\text{Vote}_i &\sim \text{Bernoulli}(p_i) \\
p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \\
\log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{\beta} \cdot \boldsymbol{x}_i \\
p_i &= \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)} \\
&= \frac{1}{1+\exp(-\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}
\end{aligned}$$

Important functions:

$$\text{odds}(p) = \frac{p}{1-p}$$

Call $p_i = \Pr(\text{Vote}_i = 1 | \boldsymbol{x}_i)$

$$
\begin{aligned}
\text{Vote}_i &\sim \text{Bernoulli}(p_i) \\
p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \\
\log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{\beta} \cdot \boldsymbol{x}_i \\
p_i &= \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)} \\
&= \frac{1}{1+\exp(-\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}
\end{aligned}
$$

Important functions:

$$
\begin{aligned}
\text{odds}(p) &= \frac{p}{1-p} \\
\text{log odds or logit}(p) &= \log\left(\frac{p}{1-p}\right)
\end{aligned}
$$

Call $p_i = \Pr(\text{Vote}_i = 1|\boldsymbol{x}_i)$

$$\begin{aligned}
\text{Vote}_i &\sim \text{Bernoulli}(p_i) \\
p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \\
\log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{\beta} \cdot \boldsymbol{x}_i \\
p_i &= \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)} \\
&= \frac{1}{1+\exp(-\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}
\end{aligned}$$

Important functions:

$$\begin{aligned}
\text{odds}(p) &= \frac{p}{1-p} \\
\text{log odds or logit}(p) &= \log\left(\frac{p}{1-p}\right) \\
\text{logistic function or logit}^{-1}(a) &= \frac{1}{1+\exp(-a)}
\end{aligned}$$

More on tutorial

R!

Some Context: Iraq Vote

- In 2002 President George Bush announced the Joint Resolution to Authorize the Use of United States Armed Forces Against Iraq.
- Congressional opposition.

NEXT

- Assessing model predictions for classification
  - Precision
  - Accuracy
  - Recall