# Logistic Regression

## and Maximum Likelihood

Marek Petrik

Feb 09 2017

# So Far in ML

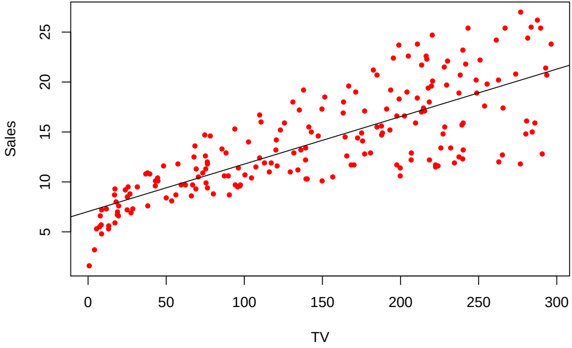
- Regression vs Classification
- Linear regression
- Bias-variance decomposition
- Practical methods for linear regression

# Simple Linear Regression

- We have only one feature

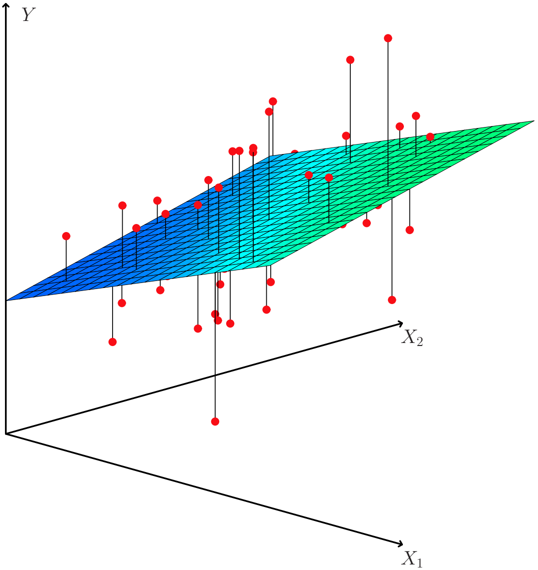
$$Y \approx \beta_0 + \beta_1 X \qquad Y = \beta_0 + \beta_1 X + \epsilon$$

- Example:




$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

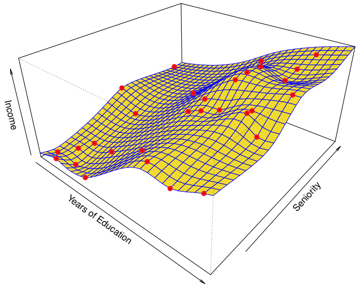
# Multiple Linear Regression



# Types of Function $f$

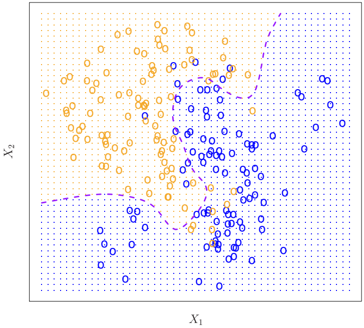
**Regression**: continuous target

$$f : \mathcal{X} \to \mathbb{R}$$




**Classification**: discrete target

$$f : \mathcal{X} \to \{1, 2, 3, \ldots, k\}$$



# Today

- Why not use linear regression for classification
- Logistic regression
- Maximum likelihood principle
- Maximum likelihood for linear regression
- Reading:
  - ISL 4.1-3
  - ESL 2.6 (max likelihood)

# Examples of Classification

1. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?

# Examples of Classification

2. An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the userffs IP address, past transaction history, and so forth.

# Examples of Classification

3. On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

# IBM Watson

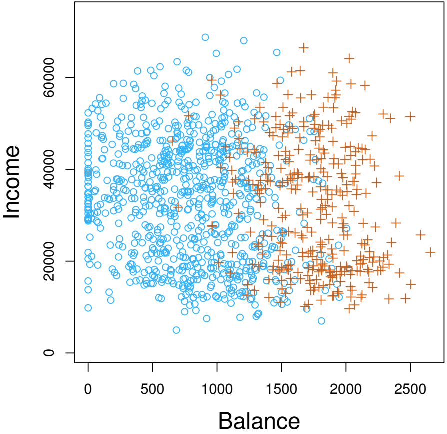
Fair use,

https://en.wikipedia.org/w/index.php?curid=31142331

**Logistic regression** + clever function engineering

# Predicting Default

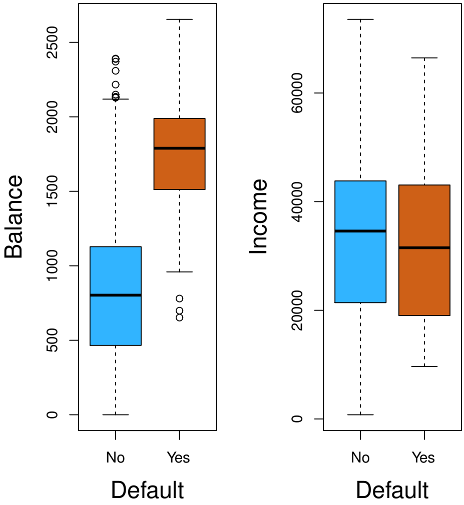
$$\text{default} \approx f(\text{income}, \text{balance})$$



# Predicting Default

$$\text{default} \approx f(\text{income}, \text{balance})$$

Boxplot



# Casting Classification as Regression

- **Regression**: $f : X \to \mathbb{R}$
- **Classification**: $f : X \to \{1, 2, 3\}$

# Casting Classification as Regression

- **Regression**: $f : X \to \mathbb{R}$
- **Classification**: $f : X \to \{1, 2, 3\}$

- But $\{1, 2, 3\} \subseteq \mathbb{R}$
- Do we even need classification?

# Casting Classification as Regression

- **Regression**: $f : X \to \mathbb{R}$
- **Classification**: $f : X \to \{1, 2, 3\}$

- But $\{1, 2, 3\} \subseteq \mathbb{R}$
- Do we even need classification?

- **Yes!**
- **Regression**: Values that are close are similar
- **Classification**: Distance of classes is meaningless

# Casting Classification as Regression: Example

- Predict possible diagnosis:

$$\{\text{stroke}, \text{overdose}, \text{seizure}\}$$

- Assign class labels:

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if overdose} \\ 3 & \text{if seizure} \end{cases}.$$

- Fit linear regression

# Casting Classification as Regression: Example

- Predict possible diagnosis:

$$\{\text{stroke}, \text{overdose}, \text{seizure}\}$$

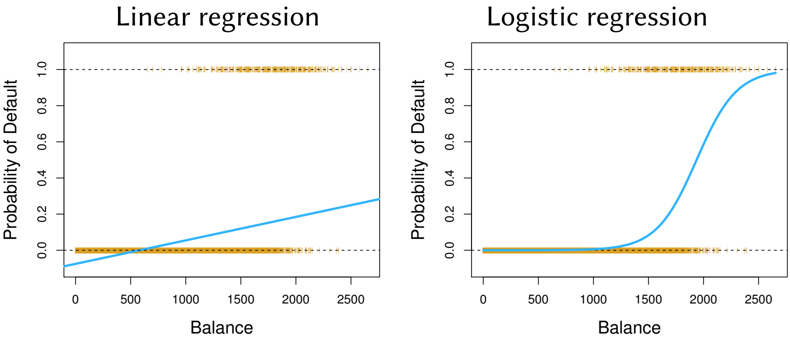
- Assign class labels:

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if overdose} \\ 3 & \text{if seizure} \end{cases}.$$

- Fit linear regression
- **Make predictions**: If uncertain whether symptoms point to stroke or seizure, we predict overdose

# Linear Regression for 2-class Classification

$$Y = \begin{cases} 1 & \text{if default} \\ 0 & \text{otherwise} \end{cases}$$




$$\mathbb{P}[\text{default} = \text{yes} \mid \text{balance}]$$

# Logistic Regression

- Predict **probability** of a class: $p(X)$
- Example: $p(\text{balance})$ probability of default for person with balance
- **Linear regression**:

$$p(X) = \beta_0 + \beta_1$$

- **logistic regression**:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

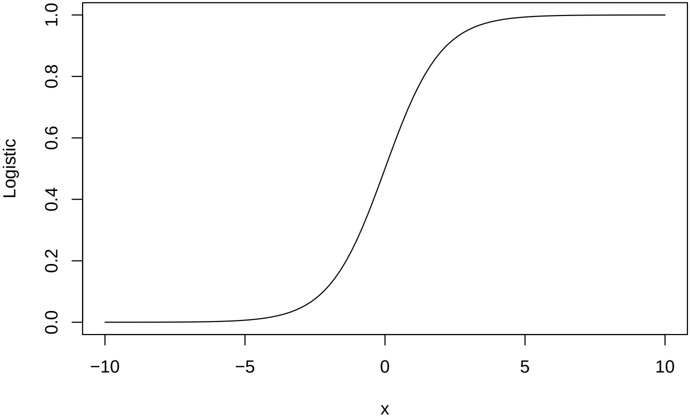
- the same as:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

- <u>Odds</u>: $p(X)/1-p(X)$

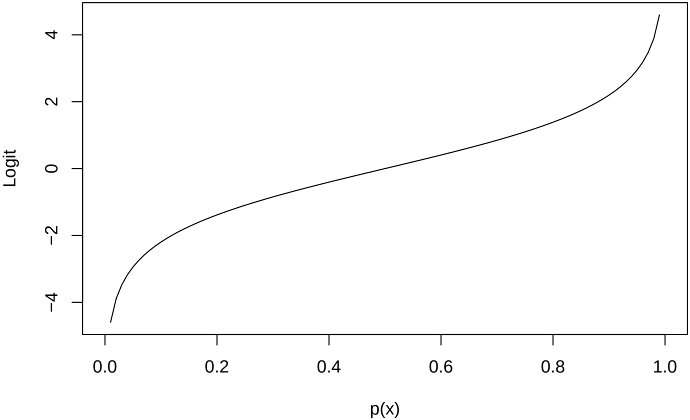
# Logistic Function

$$y = \frac{e^x}{1 + e^x}$$



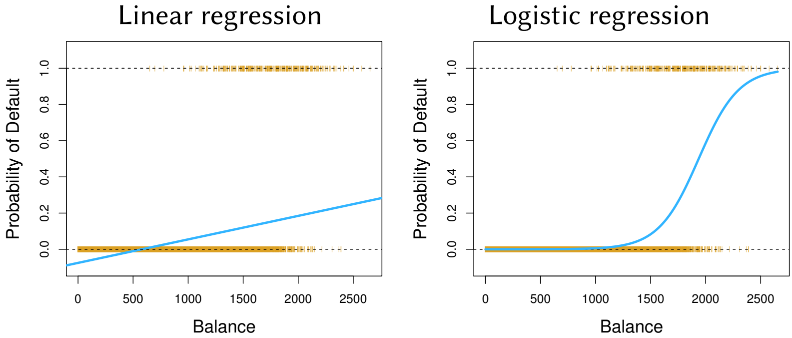
# Logistic Function

$$\log\left(\frac{p(X)}{1-p(X)}\right)$$



# Logistic Regression

$$\mathbb{P}[\textsf{default} = \text{yes} \mid \textsf{balance}] = \frac{e^{\beta_0+\beta_1 \textsf{balance}}}{1 + e^{\beta_0+\beta_1 \textsf{balance}}}$$



# Estimating Coefficients: Maximum Likelihood

- **Likelihood**: Probability that data is generated from a model

$$\ell(\text{model}) = \mathbb{P}[\text{data} \mid \text{model}]$$

- Find the most likely model:

$$\max_{\text{model}} \ell(\text{model}) = \max_{\text{model}} \mathbb{P}[\text{data} \mid \text{model}]$$

- Likelihood function is difficult to maximize
- Transform it using $\log$ (strictly increasing)

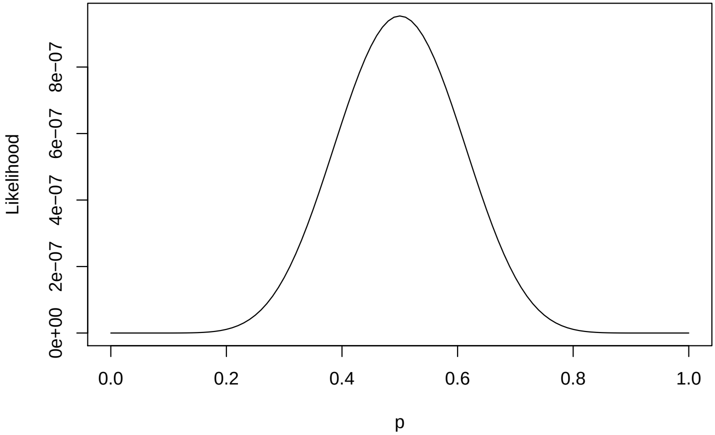
$$\max_{\text{model}} \log \ell(\text{model})$$

- Strictly increasing transformation does not change maximum

# Example: Maximum Likelihood

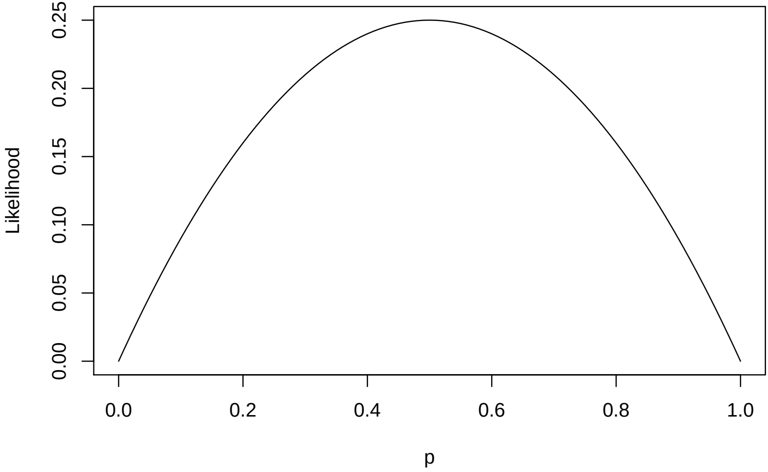
- Assume a coin with $p$ as the probability of *heads*
- **Data**: $h$ heads, $t$ tails
- The likelihood function is:

$$\ell(p) = p^h \, (1-p)^t \; .$$



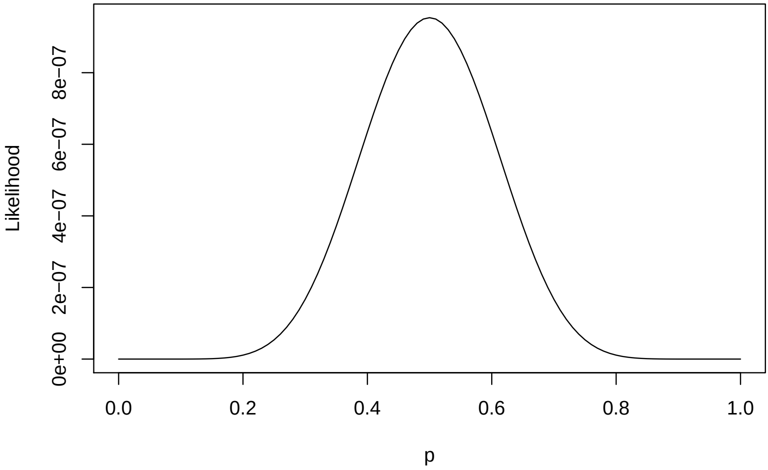
# Likelihood Function: 2 coin flips

**heads** $h = 1$ **tails** $t = 1$



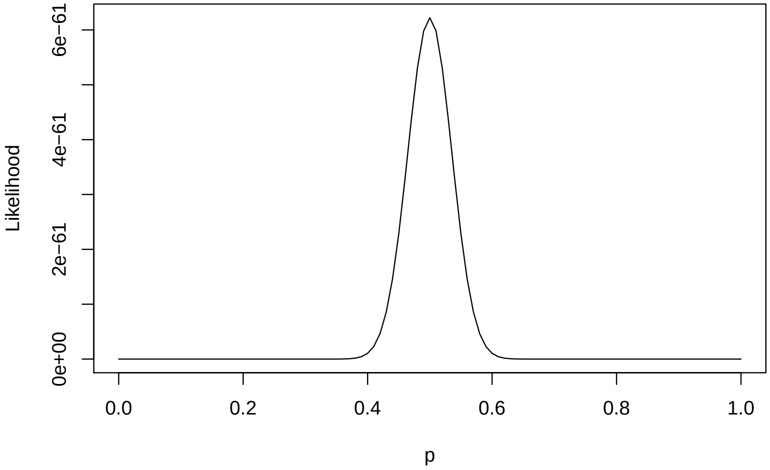
# Likelihood Function: 20 coin flips

**heads** $h = 10$ **tails** $t = 10$



# Likelihood Function: 200 coin flips

**heads** $h = 100$ **tails** $t = 100$



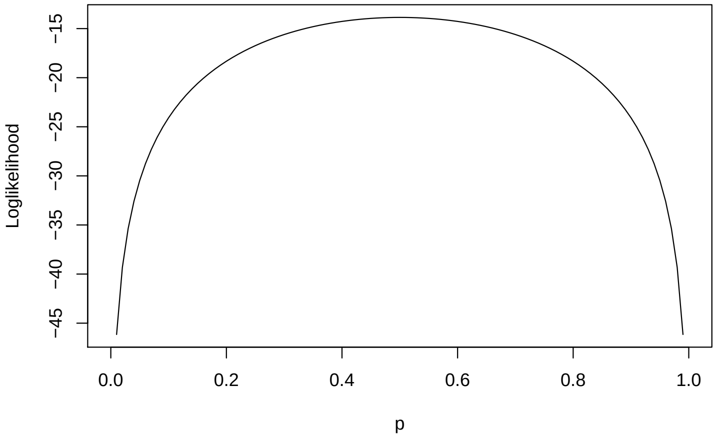
# Maximizing Likelihood

- Likelihood function is not concave: hard to maximize

$$\ell(p) = p^h \, (1-p)^t \; .$$

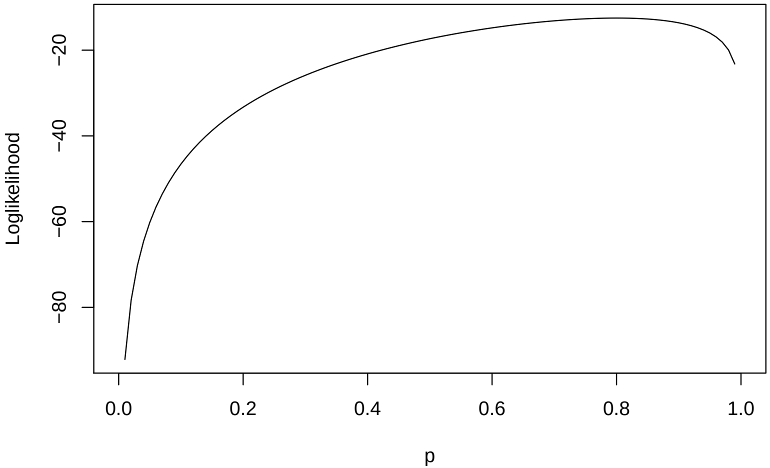
- Maximize the log-likelihood instead

$$\log \ell(p) = h \, \log(p) + t \, \log(1-p) \; .$$



# Log-likelihood: Biased Coin

**heads** $h = 20$ **tails** $t = 50$



# Maximize Log-likelihood

- Log-likelihood:

$$\log \ell(p) = h \ \log(p) + t \ \log(1-p) \ .$$

# Maximize Log-likelihood

- Log-likelihood:

$$\log \ell(p) = h \ \log(p) + t \ \log(1-p) \ .$$

- Maximum where derivative = 0
- Derivative:

$$\frac{d}{dp} h \ \log(p) + t \ \log(1-p) = \frac{h}{p} - \frac{t}{1-p}$$

# Maximize Log-likelihood

- Log-likelihood:

$$\log \ell(p) = h \log(p) + t \log(1-p) \; .$$

- Maximum where derivative = 0
- Derivative:

$$\frac{d}{dp} h \log(p) + t \log(1-p) = \frac{h}{p} - \frac{t}{1-p}$$

- Maximum likelihood solution:

$$p = \frac{h}{h+1}$$

# Max-likelihood: Logistic Regression

- Features $x_i$ and labels $y_i$
- Likelihood:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

- Log-likelihood:

$$\ell(\beta_0, \beta_1) = \sum_{i:y_i=1} \log p(x_i) + \sum_{i:y_i=0} \log(1 - p(x_i))$$

- Concave maximization problem
- Can be solved using gradient descent

# Multiple Logistic Regression

- Multiple features

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_m X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_m X_n}}$$

- Equivalent to:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_n$$

# Multinomial Logistic Regression

- Predicting multiple classes:
  - Medical diagnosis

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if overdose} \\ 3 & \text{if seizure} \end{cases}.$$

  - Predicting which products customer purchases
- Straightforward generalization of simple logistic regression

$$\frac{e^{c_1}}{1 + e^{c_1}} \quad \Rightarrow \quad \frac{e^{c_1}}{e^{c_1} + e^{c_2} + \ldots + e^{c_k}}$$