Linear Regression

Vahid Partovi Nia

Advanced Machine Learning: Lecture 01

February 3, 2018

Outline

- Simple (predictive)
- Interpretable (transparent box)
- Fast to train (big data)
- Works in wide variety of real problems (practical)
- Easy to adapt (generalizable)
- Building block of neural networks (deep learning)

- 80% Supervised learning: relate a predicting variable y to some other measured variables x
- 20% Unsupervised learning: data grouping using some measured variables x.

Some examples

Digit recognition
`http://datawisdom.ca/teaching.htm`

Equivalent terminologies

- y: dependent variable, response variable, output variable
- x: independent variable, explanatory variable, input variable, feature.

- Regression: y is continuous
- Classification: y is discrete

Advertisement

Suppose we have a fixed budget of advertisement to increase sales.

Problem:

How do you distribute advertisement budget between different advertisement methods?

Advertisement

Suppose we have a fixed budget of advertisement to increase sales.

Problem:

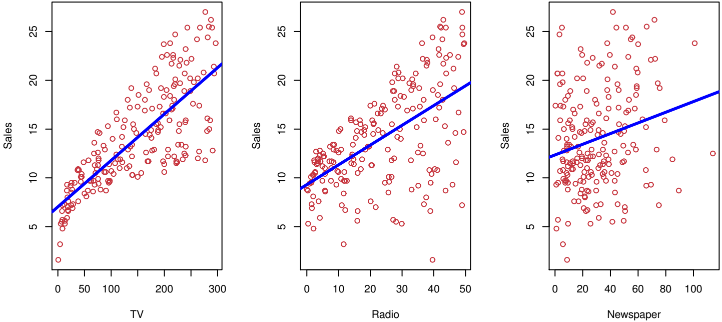
How do you distribute advertisement budget between different advertisement methods?

TV, Radio, Newspaper, Online, etc.

Question:

- Does advertisement affect sale?
- How do we predict sale?
- What is y what is x?
- Is it a regression or a classification?

Sales: y

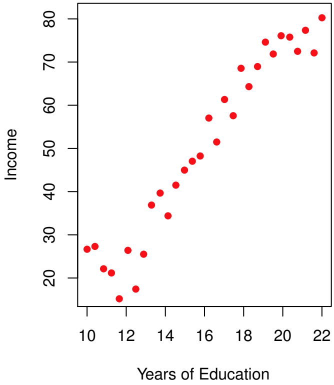


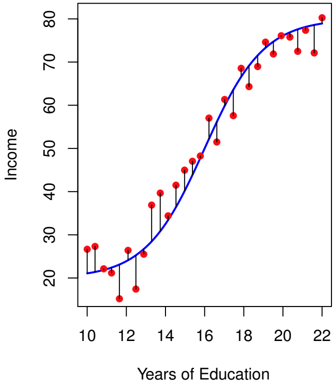
TV: x_1 Radio: x_2 Newspaper: x_3

$$\text{General model: } y \approx f(x_1, x_2, x_3)$$

$$\text{Additive model: } y \approx f_1(x_1) + f_2(x_2) + f_3(x_3)$$

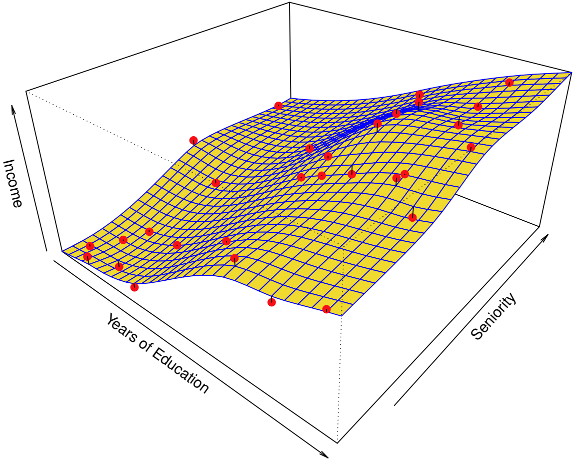
Income: y





Years of Education: x_1

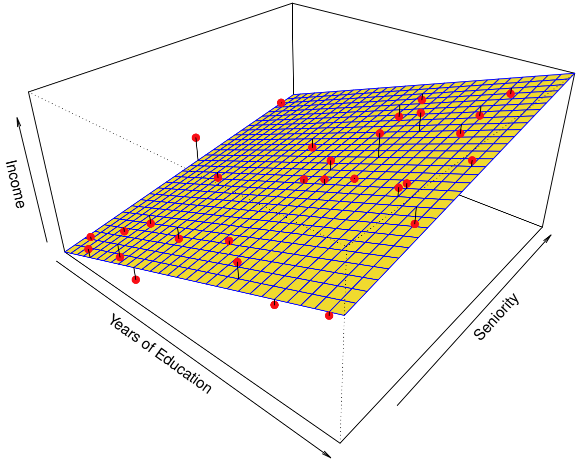
Income: y



Years of Education: x_1 Seniority: x_2

$y \approx f(x_1, x_2)$

Income: y



$$y \approx f_1(x_1) + f_2(x_2)$$
$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Sale prediction simplification

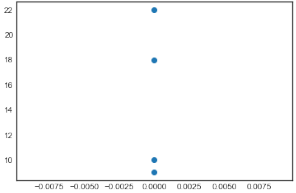
$$
\begin{array}{ccl}
y & \approx & f(x_1, x_2, x_3) \\
 & \Downarrow & \\
y & \approx & f_1(x_1) + f_2(x_2) + f_3(x_3) \\
 & \Downarrow & \\
y & \approx & f_1(x_1) \\
 & \Downarrow & \\
y & \approx & \beta_0 + \beta_1 x_1 \\
 & \Downarrow & \\
y & \approx & \beta_0
\end{array}
$$

Sales prediction

$$y_1 = 22, \quad y_2 = 10, \quad y_3 = 9, \quad y_4 = 18$$

For prediction a probabilistic model is required

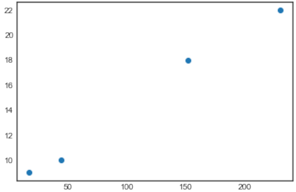
$$y_i = \beta_0 + \varepsilon_i$$



What is $\hat{y}_i$?
LS$\equiv$ MLE
LAD $\equiv$ MLE

$$y_1 = 22 \quad y_2 = 10 \quad y_3 = 9 \quad y_4 = 18$$
$$x_{11} = 230 \quad x_{12} = 44 \quad x_{13} = 17 \quad x_{14} = 151$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$



What is $\hat{y}_i$?
Mean Regression vs Median Regression.

Polynomial

$$\begin{array}{cccc} y_1 = 22 & y_2 = 10 & y_3 = 9 & y_4 = 18 \\ x_{11} = 230 & x_{12} = 44 & x_{13} = 17 & x_{14} = 151 \end{array}$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \cdots, \beta_k x_{1i}^k + \varepsilon_i$$

What is $\hat{y}_i$?

Notation n observation, p features

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

$$\mathbf{X}_{n\times(p+1)} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$$

Matrix differentiation

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ S(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

What is the minimizer of $S(\boldsymbol{\beta})$?

Matrix differentiation

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ S(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

What is the minimizer of $S(\boldsymbol{\beta})$?

$$\begin{aligned}\frac{\partial \mathbf{x}^\top \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} &= \mathbf{x} \Rightarrow \frac{\partial \mathbf{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \\ \frac{\partial \boldsymbol{\beta}^\top \mathbf{A}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} &= (\mathbf{A} + \mathbf{A}^\top)\boldsymbol{\beta}\end{aligned}$$

$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
How do we compute $\hat{\boldsymbol{\beta}}$?

Numerical Consideration

- Cholesky is faster than LU
- LU is faster than QR
- QR is faster than SVD