Predictive Modelling

Linear Regression

Jonathan Mwaura

Khoury College of Computer Science

July 16, 2024

The Roux Institute
at Northeastern University

Introduction

Textbook

Reading: Chapter 3 of: Gareth James et al (2021) . An Introduction to Statistical Learning (2nd Edition) .
`https://www.statlearning.com/`

Acknowledgements

These slides have been adapted from the following Professors:
1) Andrew Ng - Stanford
2) Eric Eaton - UPenn
3) David Sontag - MIT
4) Alina Oprea - Northeastern

Supervised Learning: Regression

Training

Data → Pre-processing → Feature extraction → Learning model

- **Data**: Labeled, $x_i, y_i \in R$
- **Pre-processing**: Normalization
- **Feature extraction**: Feature Selection
- **Learning model**: Regression, $f(x)$

Testing

New data → Learning model → Predictions $y' = f(x') \in R$

- **New data**: Unlabeled, x'
- **Learning model**: $f(x)$
- **Predictions**: Response variable, Regression

The Roux Institute
at Northeastern University

Steps to Learning Process

- Define problem space
- Collect data
- Extract feature
- Pick a model (hypothesis)
- Develop a learning algorithm
  - Train and learn model parameters
- Make predictions on new data
  - Testing phase
- In practice, usually re-train when new data is available and use feedback from deployment

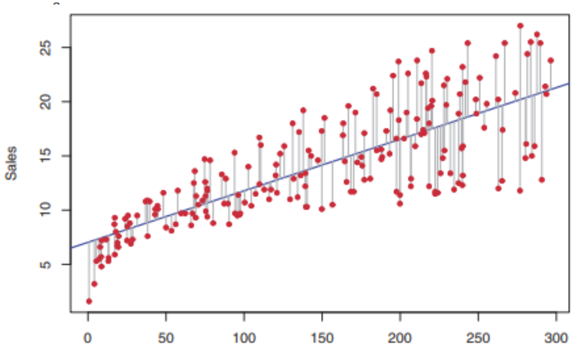
Linear regression

- One of the most widely used techniques
- Fundamental to many complex models
  - Generalized Linear Models
  - Logistic regression
  - Neural networks
  - Deep learning
- Easy to understand and interpret
- Efficient to solve in closed form
- Efficient practical algorithm (gradient descent)

Linear regression

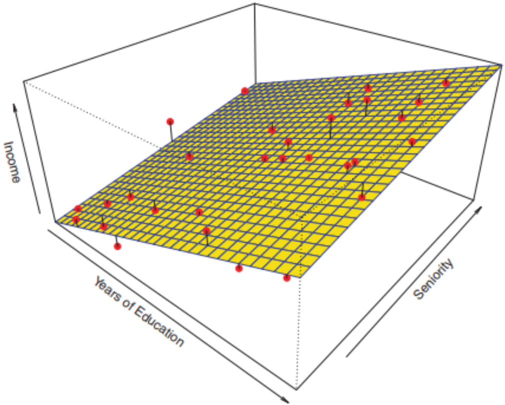
Given:

- Data $X = \{x_1, \dots x_N\}$, where $x_i \in R^d$ Features
- Corresponding labels $Y = \{y_1, \dots y_N\}$, where $y_i \in R$ Response variables



Simple Linear Regression: 1 predictor

Income Prediction



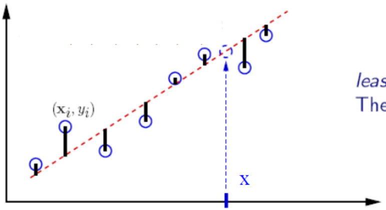
Linear Regression with 2 predictors
Multiple Linear Regression

Hypothesis: linear model

- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Simple linear regression
Regression model is a line with 2 parameters: θ_0, θ_1

- Fit model by minimizing sum of squared errors



least squares (LSQ)
The fitted line is used as a predictor

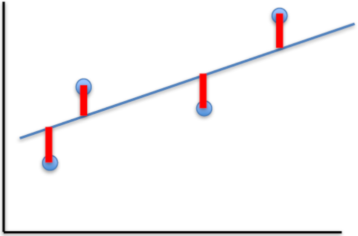
Least-Squares Linear Regression

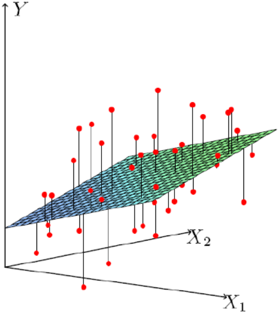
- Cost Function

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

Mean Square Error (MSE)

- Fit by solving $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

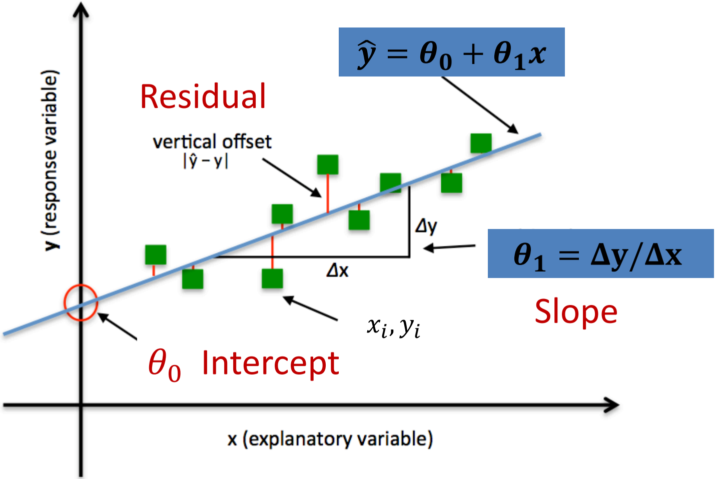




Terminology and Metrics

- Residuals
  - Difference between predicted values and actual values
  - Predicted value for example i is: $\widehat{y_i} = h_\theta(x_i)$
  - $R_i = \left|y_i - \widehat{y_i}\right| = |y_i - (\theta_0 + \theta_1 x_i)|$
- Residual Sum of Squares (RSS)
  - $RSS = \sum {R_i}^2 = \sum\left[y_i - (\theta_0 + \theta_1 x_i)\right]^2$
- Mean Square Error (MSE)
  - $MSE = \frac{1}{N}\sum R_i^2 = \frac{1}{\mathrm{N}}\sum\left[y_i - (\theta_0 + \theta_1 x_i)\right]^2$

Interpretation



$$h_\theta(x) = \theta_0 + \theta_1 x$$

MSE$= \frac{1}{N} \sum_{i=1}^{N} [h_\theta(x_i) - y_i]^2$

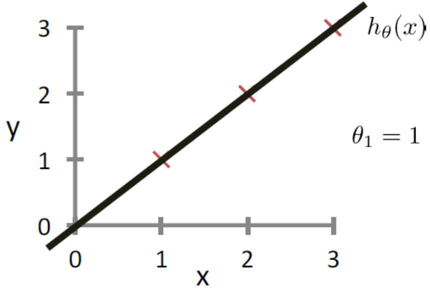
Intuition on MSE

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

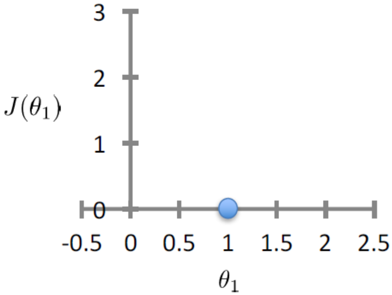
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta_1)$

(function of the parameter θ_1)



Fix θ_0 =0

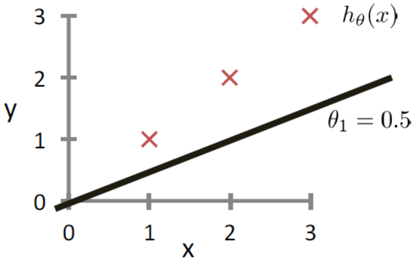
Intuition on MSE

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

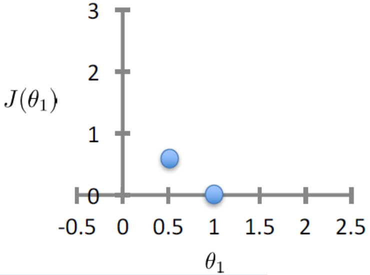
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta_1)$

(function of the parameter θ_1)



$$J([0, 0.5]) = \frac{1}{2 \times 3}\left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right] \approx 0.58$$

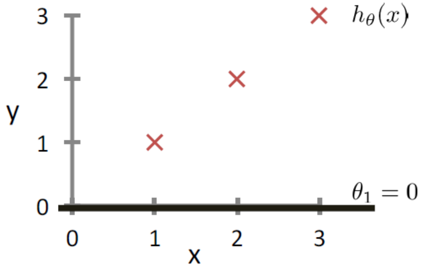
Intuition on MSE

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

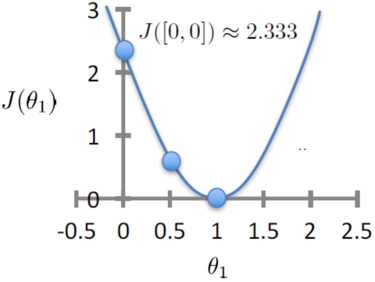
$h_\theta(x)$

(for fixed θ_1, this is a function of x)

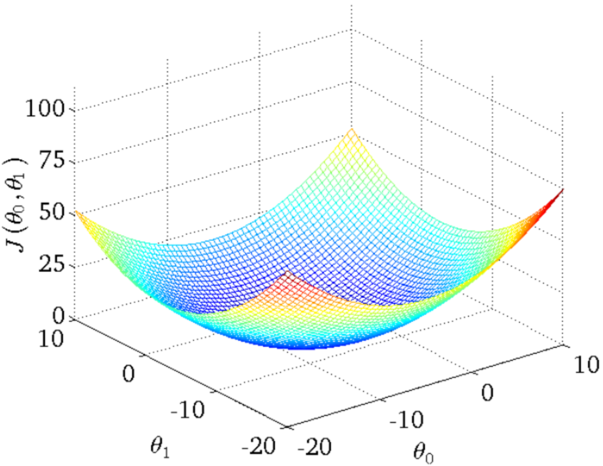


$J(\theta_1)$

(function of the parameter θ_1)



MSE function

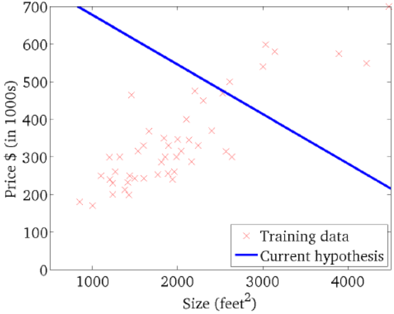


The Roux Institute
at Northeastern University

Relation between h and J

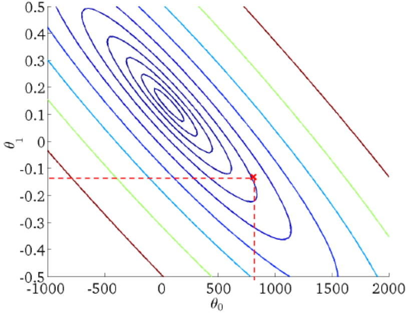
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

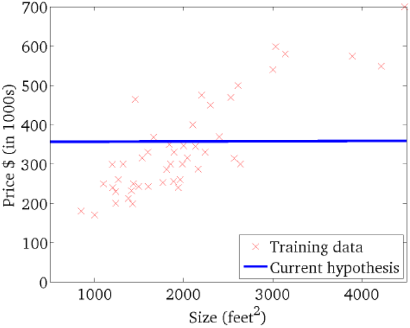
(function of the parameters θ_0, θ_1)



Relation between h and J

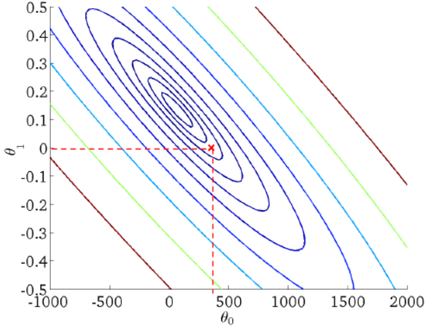
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

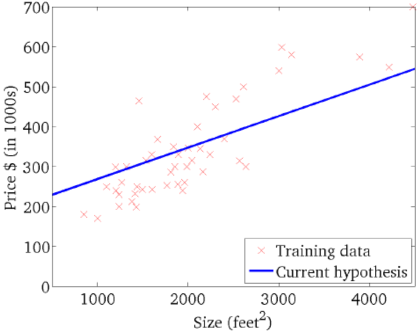
(function of the parameters θ_0, θ_1)



Relation between h and J

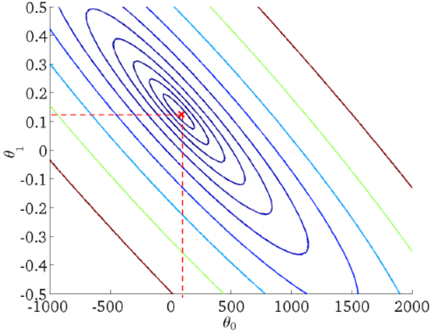
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



Find optimal model parameters θ to minimize MSE J

The Roux Institute at Northeastern University

Statistical perspective

- Response has linear dependence on input with Normal noise
  - $y_i = \theta_0 + \theta_1 x_i + \epsilon_i$, $\epsilon_i \in N(0, \sigma^2)$ noise
  - $y_i | x_i \sim N(0, \sigma^2)$
  - $f(y_i | x_i; \theta, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{-\frac{1}{2\sigma^2}[y_i - (\theta_0 + \theta_1 x_i)]^2}$ PDF
  - One training example
- Training dataset
  - $f(y_1, \dots, y_N | x_1, \dots, x_N; \theta, \sigma) = \prod_{i=1}^{N} f(y_i | x_i; \theta, \sigma)$
  - Assume independence

Maximum Likelihood Estimation (MLE)

Given training data $X = \{x_1, \dots, x_N\}$ with labels
$Y = \{y_1, \dots, y_N\}$

What is the likelihood of training data for parameter θ?

Define likelihood function

$$Max_\theta \, L(\theta) = P[Y|X;\theta] = f(y_1, \dots, y_N | x_1, \dots, x_N; \theta)$$

Assumption: training points are independent!

$$L(\theta) = \prod_{i=1}^{N} P[y_i | x_i; \theta]$$

Log Likelihood

- Max likelihood is equivalent to maximizing log of likelihood

$$L(\theta) = \prod_{i=1}^{N} P[y_i|x_i, \theta]$$

$$\log L(\theta) = \sum_{i=1}^{n} \log P[y_i|x_i, \theta]$$

- They both have the same maximum

MLE for Linear Regression

$$L(\theta) = \prod_{i=1}^{N} P[y_i | x_i; \theta] = \prod_{i=1}^{N} f(y_i | x_i; \theta, \sigma)$$

$$\log L(\theta) = -\mathrm{c} \sum_{i=1}^{N} [y_i - (\theta_0 + \theta_1 x_i)]^2$$

Max likelihood θ is the same as Min MSE θ!

The MSE metric has statistical motivation

Solution for simple linear regression

- Dataset $x_i \in R, y_i \in R, h_\theta(x) = \theta_0 + \theta_1 x$
- $J(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\theta_0 + \theta_1 x_i - y_i)^2$ MSE / Loss

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{2}{N}\sum_{i=1N}(\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{2}{N}\sum_{i=1}^{N} x_i(\theta_0 + \theta_1 x_i - y_i) = 0$$

- Solution of min loss

$- \theta_0 = \bar{y} - \theta_1 \bar{x}$

$- \theta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

How Well Does the Model Fit?

- Correlation between feature and response
  - Pearson's correlation coefficient

$$\rho = Corr(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{\mathrm{Cov}(\mathrm{X},\mathrm{Y})}{\sigma_X \sigma_Y}$$

- Measures linear dependence between X and Y
- Positive coefficient implies positive correlation
  - The closer to 1 the coefficient is, the stronger the correlation
- Negative coefficient implies negative correlation
  - The closer to -1 the coefficient is, the stronger the correlation
- $\theta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
- If σ_X = σ_Y, then $\theta_1 = \mathrm{Corr}(\mathrm{X},\mathrm{Y})$

Regression vs Correlation

- Correlation
  - Find a numerical value expressing the relationship between variables
- Regression
  - Estimate values of response variable on the basis of the values of fixed variable.
- The slope of linear regression is related to correlation coefficient
- Regression scales to more than 2 variables, but correlation does not