Chapter 2

Linear Regression

Linear Regression

- Regression: Algorithm to ***predict a*** RESPONSE variable y …
  - based on a set of REGRESSORS or EXPLANATORY VARIABLES

Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

Linear Regression

- Regression: Algorithm to ***predict a response*** variable y...
  - based on a set of **regressors** or **explanatory variables**

y Response

Regressor x

$x(k)$ — Regressor

$y(k)$ — Response

k

Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

Linear Regression

- In general the ***regressor*** $\bar{\mathbf{x}}$ can be an n −dimensional vector
  - x_1 is the cost of ____TV____
  - x_2 is the cost of ____Radio____ and so on….

x_3 : Newspaper

y

multiple regressors.

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Linear Regression

- In general the ***regressor*** $\bar{\mathbf{x}}$ can be an n −dimensional vector
  - x_1 is the cost of **TV advertising**
  - x_2 is the cost of **Radio advertising** and so on….

Regression vector

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{matrix} \text{TV} \\ \text{Radio} \\ \text{Newspaper} \end{matrix}$$

$y(k)$

$x_1(k)$

$x_2(k)$

$x_3(k)$

k

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Linear Regression

- In general the ***regressor*** $\bar{\mathbf{x}}$ can be an n −dimensional vector
  - x_1 is the cost of **TV advertising**
  - x_2 is the cost of **Radio advertising** and so on….

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Regression: Other Examples

Microsoft

- **Example 1**
- $y(k)$ = Price of ***particular stock*** at time k

Facebook Amazon Google

  - $x_1(k), x_2(k), \ldots, x_N(k)$: Prices of **related stocks** at time k

Linear Regression Model

$$y(k) = h_0 \times 1 + h_1 x_1(k) + h_2 x_2(k) + \cdots + h_n x_n(k) + \epsilon(k)$$

Regressors

Bias

Regression coefficients

Bias.

Model Error

Linear Regression Model

$$y(k) = h_0 + h_1 x_1(k) + h_2 x_2(k) + \cdots + h_n x_n(k) + \epsilon(k)$$

$$= \underbrace{\begin{bmatrix} 1 & x_1(k) & x_2(k) & \cdots & x_n(k) \end{bmatrix}}_{\bar{x}^T(k)} \underbrace{\begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_n \end{bmatrix}}_{\bar{h}} + \epsilon(k)$$

Model Error

$$= \bar{x}^T(k) \cdot \bar{h} + \epsilon(k)$$

Linear Regression Model

$$y(k)$$
$$= h_0 + h_1 x_1(k) + \cdots + h_n x_n(k) + \epsilon(k)$$

$$= \underbrace{\overbrace{[1 \quad x_1(k) \quad x_2(k) \quad \ldots \quad x_n(k)]}^{1\times(n+1)}}_{\bar{x}^T(k)} \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_n \end{bmatrix} + \epsilon(k)$$

Model Error

Regression vector

Regressors

$$y(k) = \bar{\mathbf{x}}^T(k)\bar{\mathbf{h}} + \epsilon(k)$$

Regression coefficient Vector

Linear Regression

- This is termed Linear Regression
- $h_0, h_1, \ldots, h_n$ are termed the **Regression coefficients**

We have to learn the Regression coefficient vector $\bar{h}$

Linear Regression

- This is termed ***Linear Regression***
- $h_0, h_1, \ldots, h_n$ are termed the **Regression coefficients**

Training Data

- The ***regression coefficients*** can be computed as follows
- Consider the availability of Training Pairs $\left(y(k), \bar{\mathbf{x}}(k)\right)$
  - for $k = 1,2, \ldots, M$

$y(1), \bar{x}(1)$

$y(2), \bar{x}(2)$

$\vdots$

$y(M) \quad \bar{x}(M)$

Training Data

- The ***regression coefficients*** can be computed as follows
- Consider the availability of **training pairs** $\left(y(k), \bar{\mathbf{x}}(k)\right)$
  - for $k = 1,2, \dots, M$

Model Computation

- The ***training set*** can be expressed as

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(M) \end{bmatrix} = \begin{bmatrix} \bar{x}^T(1) \\ \bar{x}^T(2) \\ \vdots \\ \bar{x}^T(M) \end{bmatrix} \bar{h} + \begin{bmatrix} \epsilon(1) \\ \epsilon(2) \\ \vdots \\ \epsilon(M) \end{bmatrix}$$

Model Computation

- The ***training set*** can be expressed as

M Responses

M Regression Vectors.

M errors.

Regression Learning Model.

$$\underbrace{\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(M) \end{bmatrix}}_{\bar{\mathbf{y}}} = \underbrace{\begin{bmatrix} \bar{\mathbf{x}}^T(1) \\ \bar{\mathbf{x}}^T(2) \\ \vdots \\ \bar{\mathbf{x}}^T(M) \end{bmatrix}}_{\mathbf{X}} \bar{\mathbf{h}} + \underbrace{\begin{bmatrix} \epsilon(1) \\ \epsilon(2) \\ \vdots \\ \epsilon(M) \end{bmatrix}}_{\bar{\boldsymbol{\epsilon}}}$$

M x 1

M x (n+1)

M x 1

Least Squares

$$\bar{y} = X\bar{h} + \bar{\epsilon}$$

$$\bar{\epsilon} = \bar{y} - X\bar{h}$$

Least Squares.

min Error

$\min \|\bar{\epsilon}\|^2$

$$\min \|\bar{y} - X\bar{h}\|^2 .$$

Least Squares

$$\min \left\| \underbrace{\bar{\mathbf{y}} - \mathbf{X}\,\bar{\mathbf{h}}}_{\bar{\epsilon}} \right\|^2$$

Model Computation

- The ***regression coefficients*** are given as

$$\bar{b} = (X^T X)^{-1} X^T \bar{y}$$

Regression Vector

Model Computation

- The ***regression coefficients*** are given as

X : Tall matrix

$\Rightarrow$ NOT square.

Regression coeff vector

$$\bar{\mathbf{h}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\bar{\mathbf{y}}$$

$(X^T X)^{-1} X^T$: Pseudoinverse

$(X^T X)^{-1}(X^T \cdot X) = I$

The Boston Housing Dataset

- The **Boston Housing Dataset** is a derived from information collected by the U.S. Census Service USA.
- Concerning housing in the area of Boston MA.

- The following describes the dataset columns:
- **CRIM** - per capita crime rate by town
- **ZN** - proportion of residential land zoned for lots over 25,000 sq.ft.
- **INDUS** - proportion of non-retail business acres per town.
- **CHAS** - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- **NOX** - nitric oxides concentration (parts per 10 million)
- **RM** - average number of rooms per dwelling
- **AGE** - proportion of owner-occupied units built prior to 1940
- **DIS** - weighted distances to five Boston employment centres
- **RAD** - index of accessibility to radial highways
- **TAX** - full-value property-tax rate per $10,000
- **PTRATIO** - pupil-teacher ratio by town
- **MEDV** - Median value of owner-occupied homes in $1000's

Crime Rate

Pollution

Number of rooms

Response

Regressors:

11 Regressors.

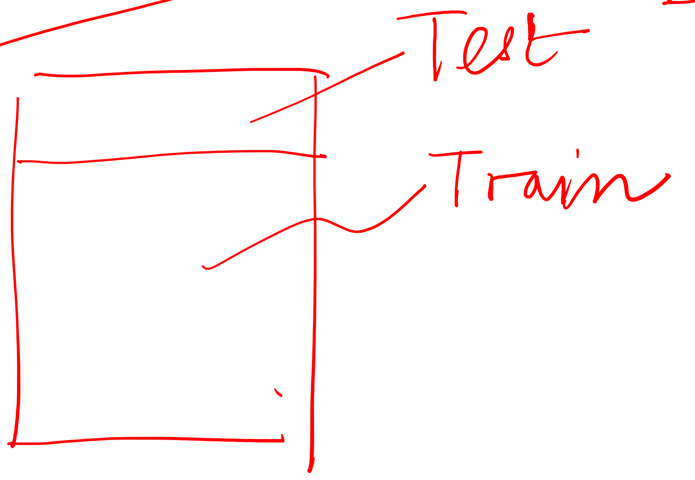
```python
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
import pandas as pd
```

Scikit.learn

Linear Regression

import numpy

Data Handling



```
BosData = pd.read_csv('BostonHousing.csv')
X = BosData.iloc[:,0:11]
y = BosData.iloc[:, 13] # MEDV: Median value of owner-occupied homes in $1000s
```

Pandas

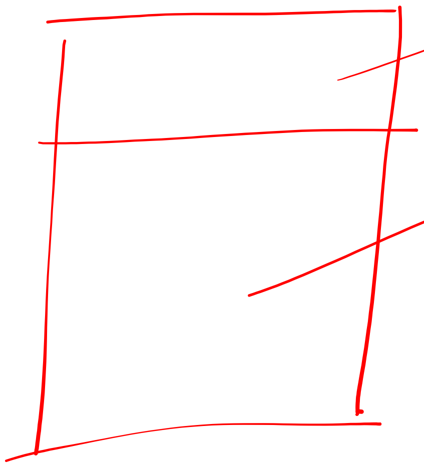
Read CSV File
Boston housing data

Response
Median home price

11 Regressors
in each Regression vector

```python
X_train, X_test, y_train, y_test =\
 train_test_split(X, y, test_size = 0.2, random_state=5)
reg = LinearRegression()
reg.fit(X_train, y_train)
```

Linear Regression module



Test 20%

Train 80%

Fits Linear Regression model For traing set

```
y_train_predict = reg.predict(X_train)
rmse = np.sqrt(mean_squared_error(y_train,y_train_predict));
r2 = r2_score(y_train, y_train_predict)
```

Root MSE

Root mean square error

using model learnt Predict Response for Training set

R2 score

$$RMSE = \sqrt{\underbrace{\frac{1}{M}\sum_{k=1}^{M}\left(y(k)-\hat{y}(k)\right)^2}_{MSE}}$$

mean squared error Between training response & Prediction frome model.

```python
print('Train RMSE =', rmse)
print('Train R2 score =', r2)
print("\n")
```

RMSE

R2 score

```python
y_test_predict = reg.predict(X_test)
rmse = (np.sqrt(mean_squared_error(y_test, y_test_predict)))
r2 = r2_score(y_test, y_test_predict)
print('Test RMSE =', rmse)
print('Test R2 score =', r2)
```

compute RMSE for test

compute R2 score for test

Run prediction For Test set

Print RMSE for test set

Print R2 score for test set

```
Train RMSE = 5.511467677842388
Train R2 score = 0.6463832866583819


Test RMSE = 4.287105260205546
Test R2 score = 0.7652527354155101

In [9]:
```

Test set

Training set

Instructors may use this white area (14.5 cm / 25.4 cm) for the text.
Three options provided below for the font size.

Font: Avenir (Book), Size: 32, Colour: Dark Grey

Font: Avenir (Book), Size: 28, Colour: Dark Grey

Font: Avenir (Book), Size: 24, Colour: Dark Grey

Do not use the space below.