5th Week Lecture Note

Basic Mobile Lab 2 (521200-1)

Sehwan (Paul) Kim, Ph.D

Associate Professor, **Dept. Biomedical Engineering, School of Medicine, Dankook University, Korea**
Vice President, **Beckman Laser Institute Korea, Dankook University, Korea**
Senior Specialist, **Beckman Laser Institute, University of California, Irvine, USA**

Contents of this week

- Review & More : Terminologies
- Linear Regression

Terminologies!

Let’s assume that you want to predict or explain ‘sales’,
using budgets of ‘TV’, ‘Radio’, and ‘Newspaper’

‘TV’, ‘Radio’, ‘Newspaper’ : **Input variable**, typically denoted using x_1, x_2, x_3
Inputs can have various different names :
predictors, independent variables, features, or just variables

‘Sales’ : **Output variable**, typically denoted using y
output can have various different names :
response variable, dependent variable, target

Linear Model : Linear Regression

Terminologies : supervised learning

When we do supervised learning, we assume that there is some relationship between **y** and **X** = $(x_1, x_2, x_3, .. , x_p)$

$$y = f(X) + \varepsilon$$

f : fixed, unknown function of **X**, relationship between **y** and **X**.
systematic information that **X** provides about **y**.

ε : random error term, noise of the data itself, irreducible.
cannot be predicted using **X**.

Terminologies : supervised learning

When we want to make a prediction or inference,
we are making $\hat{f}$ **as an estimate for** f

$$y = f(X) + \varepsilon$$

$$\hat{y} = \hat{f}(X)$$

$\hat{f}$: an estimated relationship (a ML model you choose)
$\hat{y}$: prediction for y

Linear Model : Linear Regression

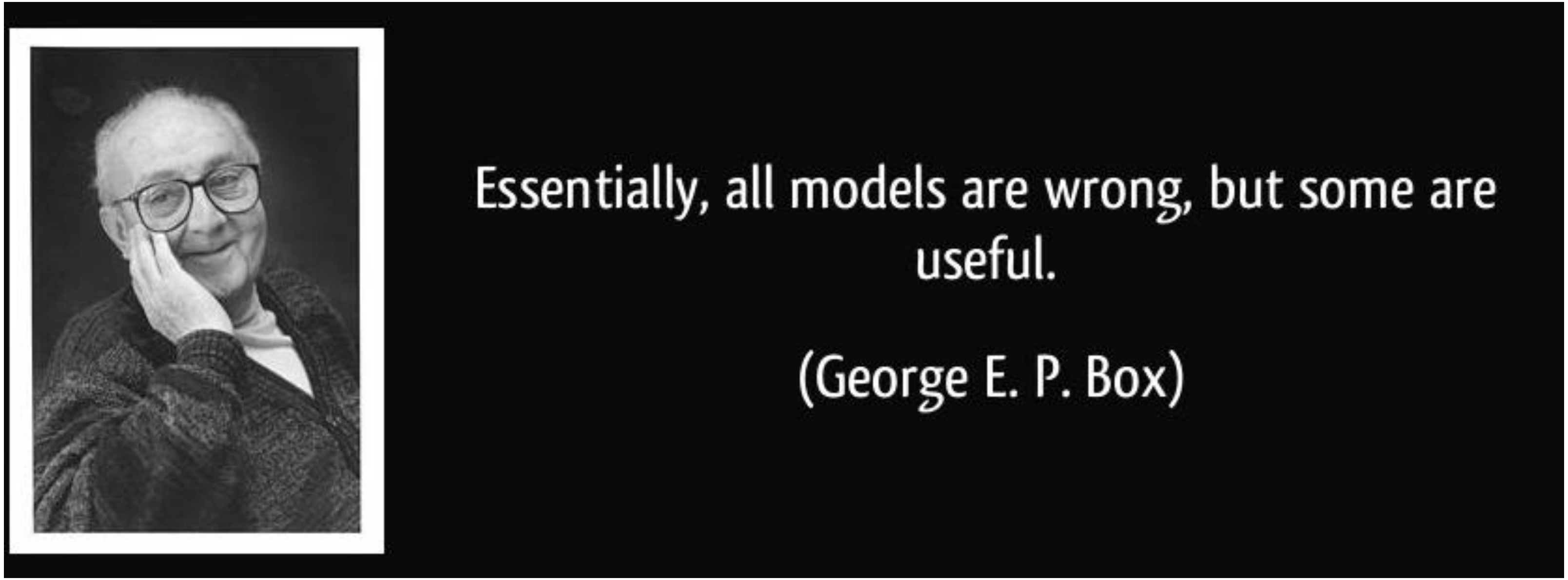
Terminologies! (+advanced)

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\operatorname{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

1. The more diffrence between $\hat{f}$ and true f, the bigger reducible part
2. The irreducible part depends on the structure of true f.
   In other words, if the structure of f is assumed differently,
   the irreducible part also changes.
3. **It is usually hard to believe that the true relationship is linear**

Linear Model : Linear Regression

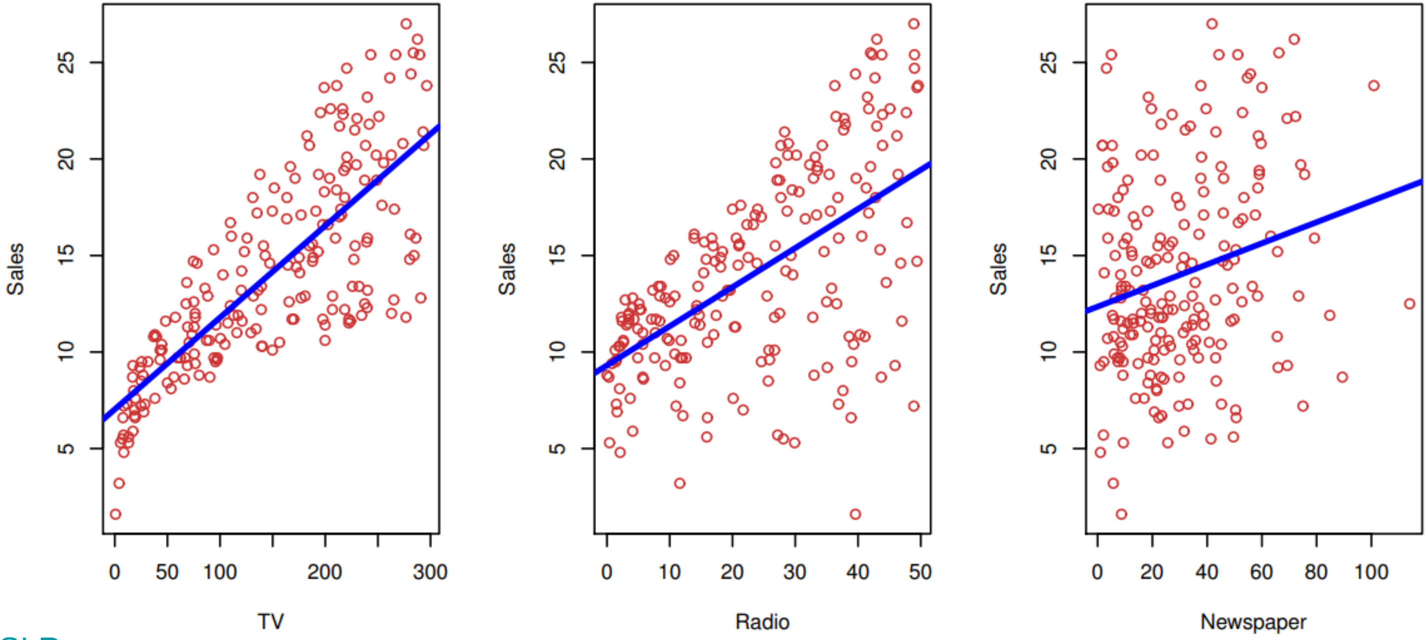
Keep in mind, we need an ‘useful’ model.



Linear Model : Linear Regression

Let's use 'Advertising' data set as an example.

Dataset consists of the **sales** of the product, along with advertising budgets for three diffent media, : **TV**, **radio**, and **newspaper**.



reference : ISLR

Questions we might seek to address :

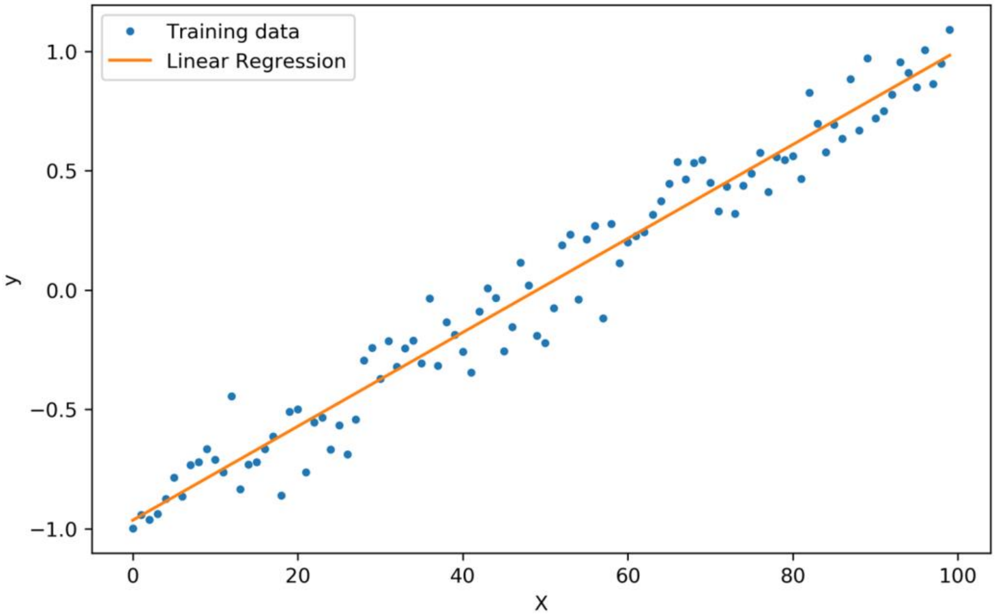
1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we predict future sales?
5. Is the relationship linear?
6. Is there synergy among the advertising media?

reference : ISLR

Linear Model : Linear Regression

Linear regression can be used to answer these questions.

Some questions need statistics, but you can answer practically.



Linear Model : Linear Regression

Linear regression

We assume a model with linearity for ‘**usefulness**’

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \boldsymbol{\varepsilon}$$

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

β : Coefficient, parameter, weights, 0 for slope, others for intercept
$\hat{\beta}$: estimate for β, (It is also called coefficient, parameter, ...)

Linear regression

We assume a model with linearity for '**usefulness**'

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \boldsymbol{\varepsilon}$$

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

These are steps 1&2 of the basic usage of sklearn

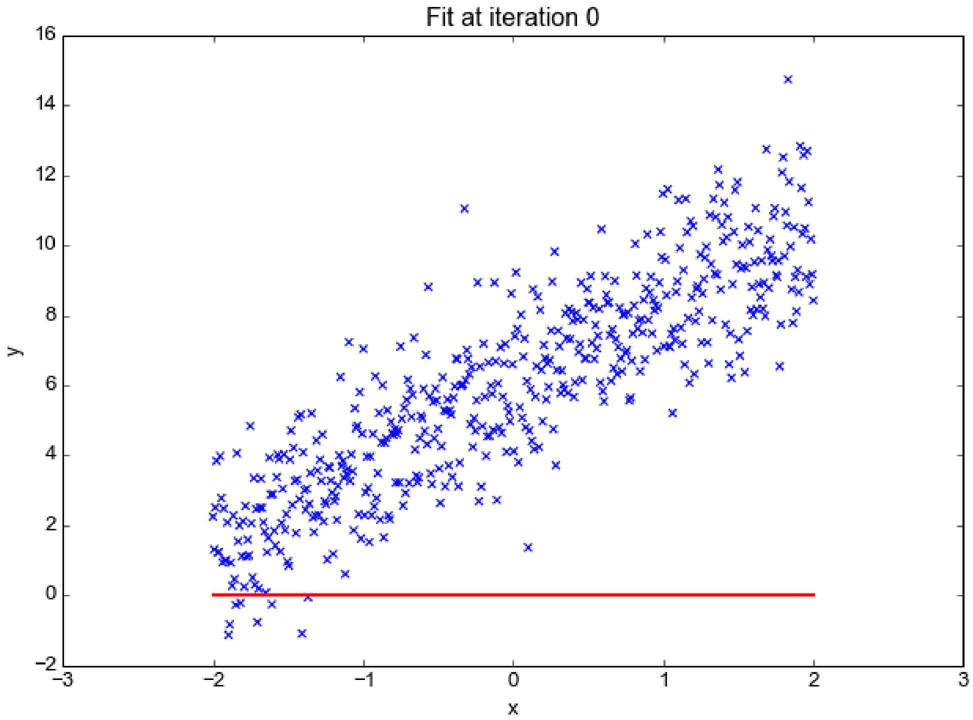
1. from sklearn.linear_model import LinearRegression
2. lr = LinearRegression( ) : Don't need to care about arguments

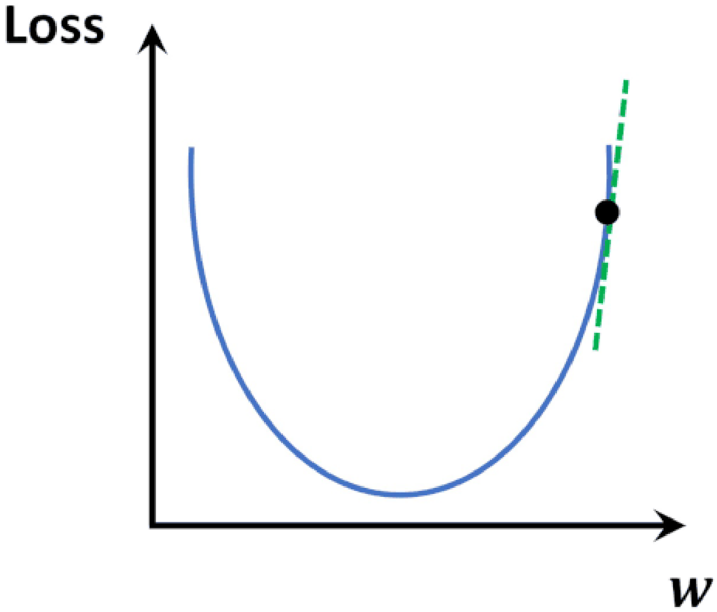
If you want to estimate beta, then

3. lr.fit(x, y)

Linear Model : Linear Regression

We will not focus ‘optimization’, but ‘interpretation’





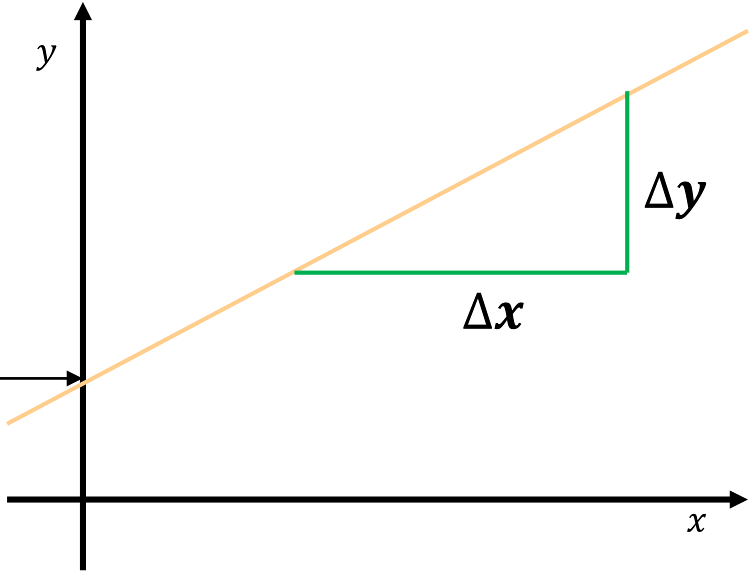
Linear Model : Linear Regression

We will not focus ‘optimization’, but ‘interpretation’

slope(a) $= \dfrac{\boldsymbol{\Delta y}}{\boldsymbol{\Delta x}}$

$$y = ax + b$$

intercept(b)



Interpretation of coefficients

Let’s assume that you are the owner of a food truck. Temperature in celsius, Humidiy in %, daily measured.

$$\text{Sales} = 10 * \text{Temperature} - 8 * \text{Humidity} + 15 * \text{Leaflets} + 120$$

Q1. What does 10 mean ?

Q2. In data, min(Humidity) is 10, max(Humidity) is 10.5.
What is the maximum effect of humidiy on sales?

Q3. ‘Leaflets’ is the number of leaflets you diligently distributed yesterday.
What is the difference between leaflets and temperature?

Q4. What does 120 mean?

Interpretation of coefficients

Let's assume that you are the owner of a food truck. Temperature in celsius, Humidiy in %, daily measured.

Sales = 10∗Temperature -8∗Humidity +15∗Leaflets + 120

Q1. What does 10 mean ?

→ For every 1 celsius increase in temperature, sales increase by 10

Q2. In data, min(Humidity) is 10, max(Humidity) is 10.5.
What is the maximum effect of humidiy on sales?

Q3. 'Leaflets' is the number of leaflets you diligently distributed yesterday.
What is the difference between leaflets and temperature?

Q4. What does 120 mean?

Interpretation of coefficients

Let's assume that you are the owner of a food truck. Temperature in celsius, Humidiy in %, daily measured.

$$\text{Sales} = 10 * \text{Temperature} - 8 * \text{Humidity} + 15 * \text{Leaflets} + 120$$

Q1. What does 10 mean ?

→ For every 1 celsius increase in temperature, sales increase by 10

Q2. In data, min(Humidity) is 10, max(Humidity) is 10.5.
What is the maximum effect of humidiy on sales?

→ 4 (= 8 x (10.5 – 10))

Q3. 'Leaflets' is the number of leaflets you diligently distributed yesterday.
What is the difference between leaflets and temperature?

Q4. What does 120 mean?

Linear Model : Linear Regression

Interpretation of coefficients

Let's assume that you are the owner of a food truck. Temperature in celsius, Humidiy in %, daily measured.

$$\text{Sales} = 10 * \text{Temperature} - 8 * \text{Humidity} + 15 * \text{Leaflets} + 120$$

Q1. What does 10 mean ?

→ For every 1 celsius increase in temperature, sales increase by 10

Q2. In data, min(Humidity) is 10, max(Humidity) is 10.5.
What is the maximum effect of humidiy on sales?

→ 4 (= 8 x (10.5 – 10))

Q3. 'Leaflets' is the number of leaflets you diligently distributed yesterday.
What is the difference between leaflets and temperature?

→ The temperature is hard to change with your efforts.

Q4. What does 120 mean?

Interpretation of coefficients

Let's assume that you are the owner of a food truck. Temperature in celsius, Humidiy in %, daily measured.

Sales = 10$*$Temperature -8$*$Humidity +15$*$Leaflets + 120

Q1. What does 10 mean ?

→ For every 1 celsius increase in temperature, sales increase by 10

Q2. In data, min(Humidity) is 10, max(Humidity) is 10.5.
What is the maximum effect of humidiy on sales?

→ 4 (= 8 x (10.5 – 10))

Q3. 'Leaflets' is the number of leaflets you diligently distributed yesterday.
What is the difference between leaflets and temperature?

→ The temperature is hard to change with your efforts.

Q4. What does 120 mean?

→ When all other variables have no effect, 120 is the basic sales volume

Linear Model : Linear Regression

Interpretation of coefficients

Let’s assume that you are the owner of a food truck. Temperature in celsius, Humidiy in %, daily measured.

$$\text{Sales} = \boxed{10} * \text{Temperature} \boxed{-8} * \text{Humidity} \boxed{+15} * \text{Leaflets} \boxed{+\ 120}$$

In sklearn

lr.coef_

lr.intercept_

Want to get ŷ? : lr.predict(x)

Basic metrics for regression

RMSE :

$$\sqrt{\frac{1}{n}\boxed{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}$$

SSE : Sum of Squared Error

Linear Model : Linear Regression

Basic metrics for regression

RMSE :

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

SSE : Sum of Squared Error

MSE : Mean Squared Error

Basic metrics for regression

RMSE : Root Mean Squared Error

$$\sqrt{\boxed{\frac{1}{n}\boxed{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}}$$

SSE : Sum of Squared Error

MSE : Mean Squared Error

Questions?