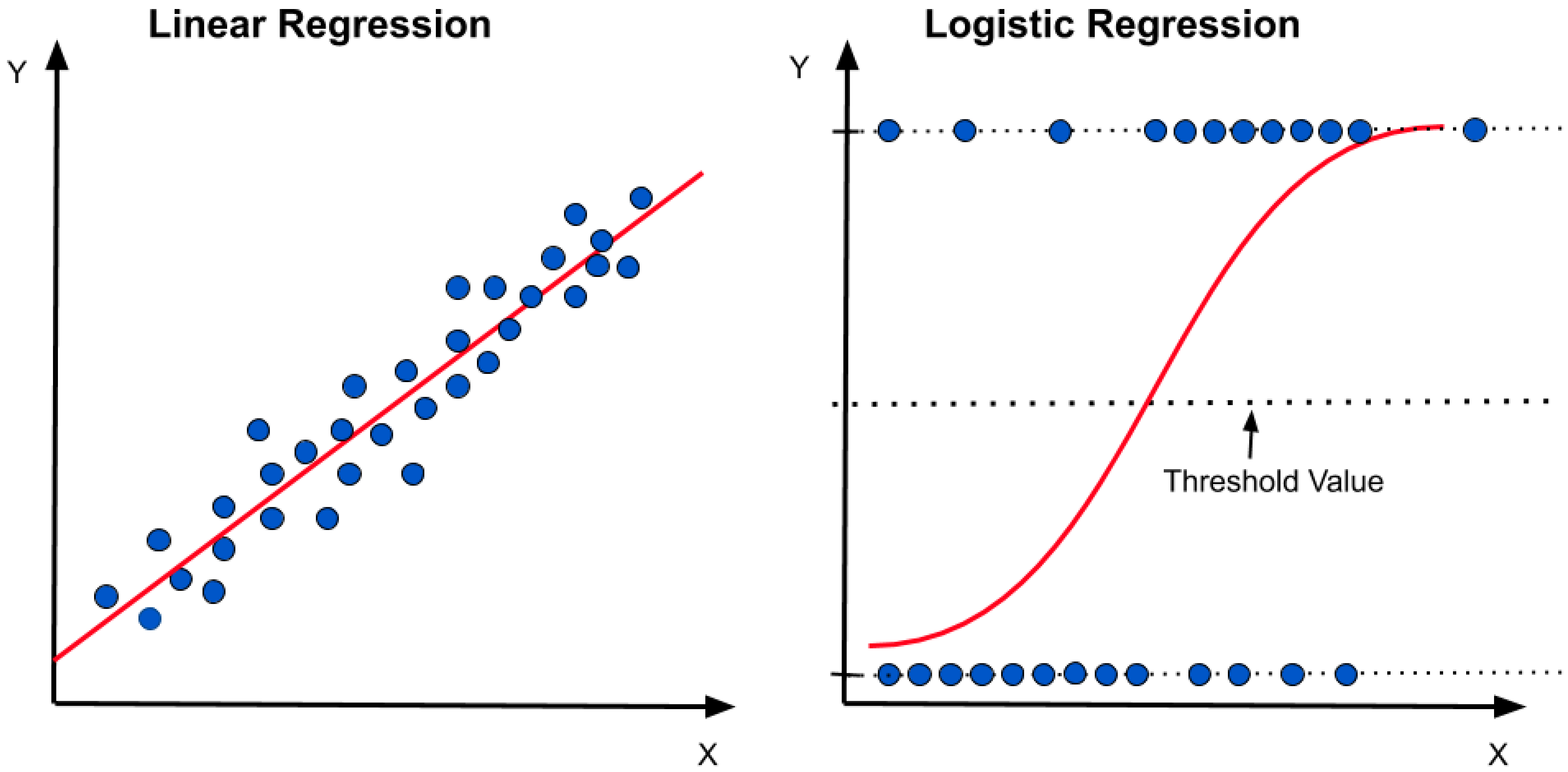
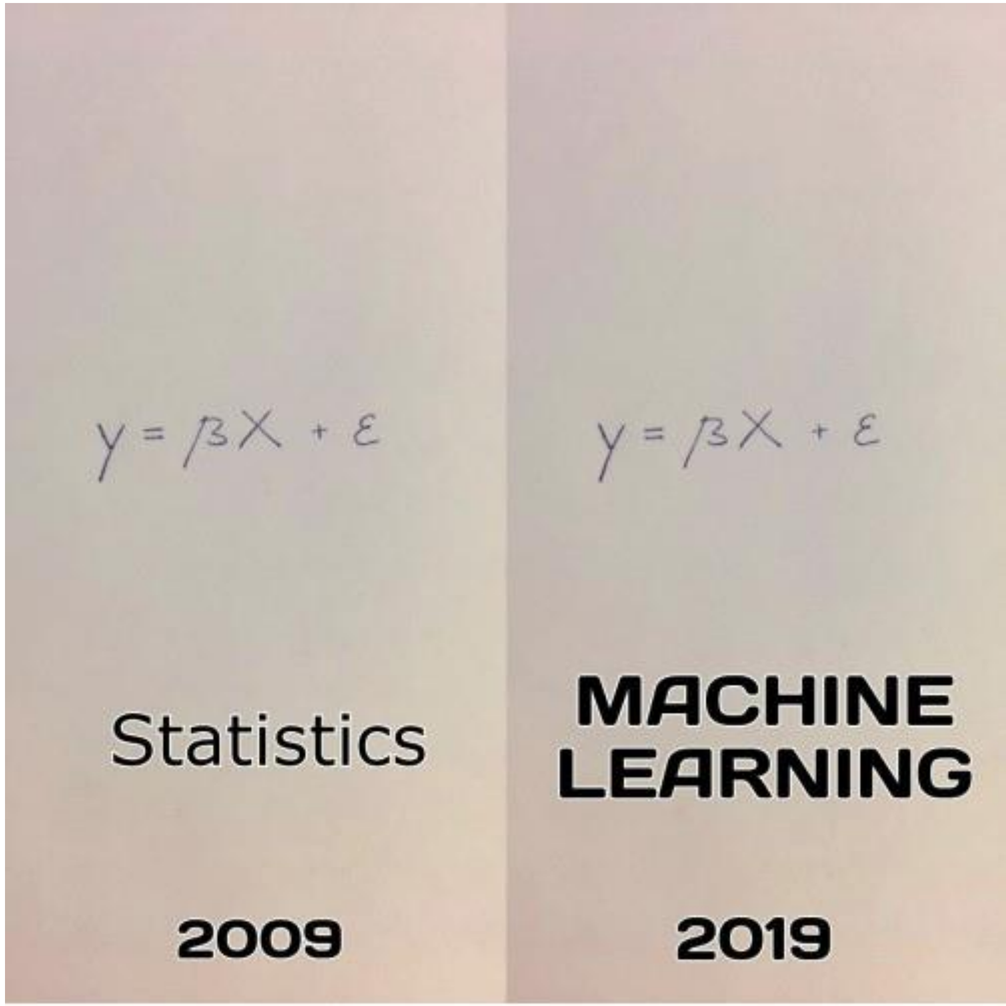
Simple Linear Regression

Ji Hun Kim

Applied Mathematics and Statistics at SUNY Korea



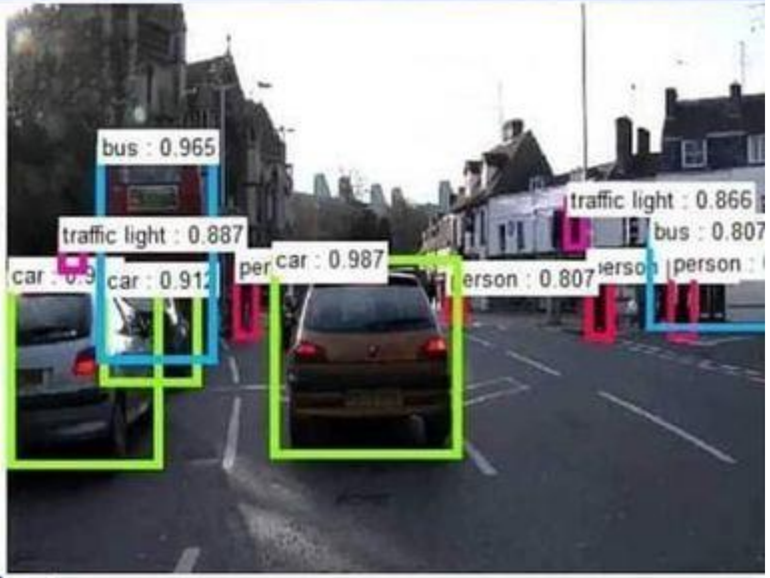
Linear Regression
Y
X
Logistic Regression
Y
Threshold Value
X



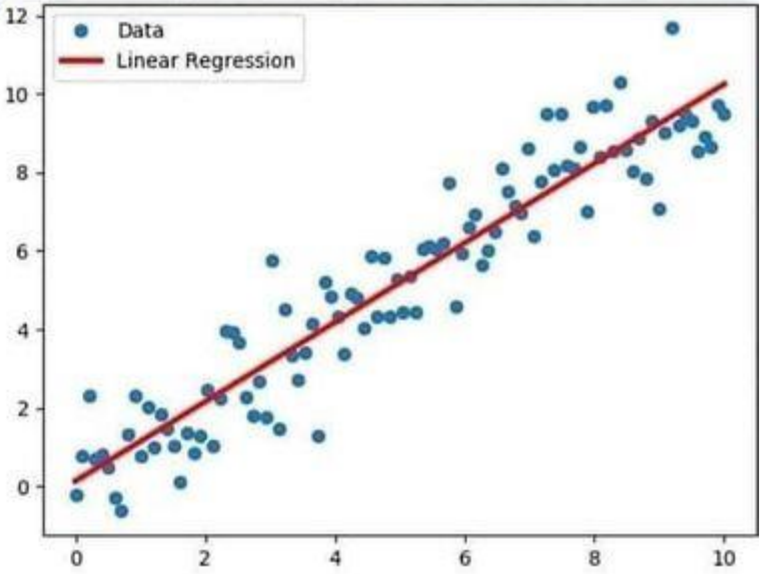
#10yearchallenge

Online Courses

What they promise you will learn

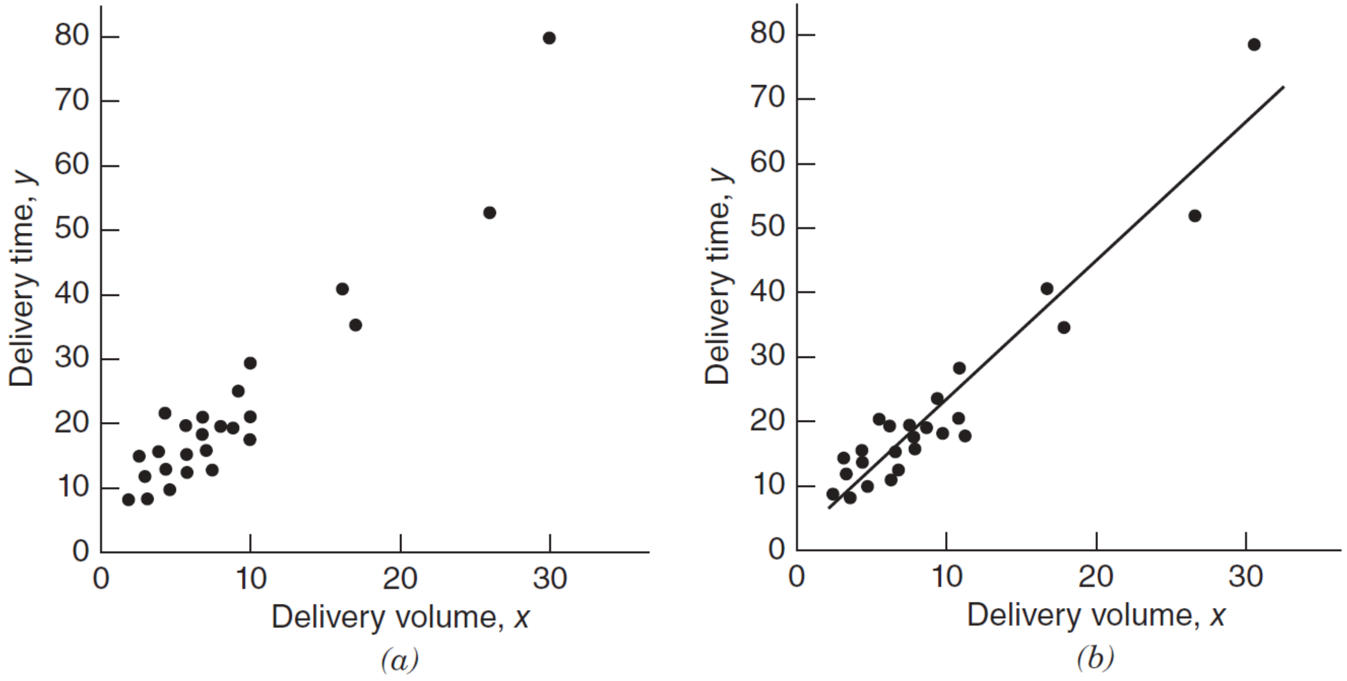


What you actually learn



Regression Model (Example)

- Simple Linear Regression



Regression Model (Example)

- Multiple Linear Regression

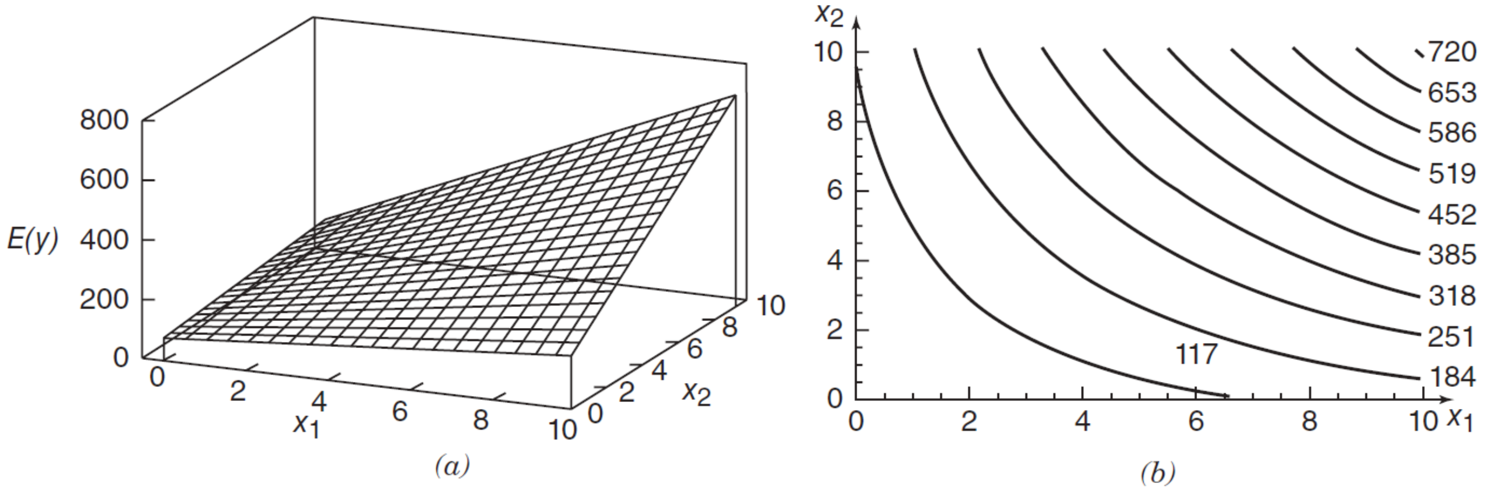


Figure 3.2 (a) Three-dimensional plot of regression model $E(y) = 50 + 10x_1 + 7x_2 + 5x_1x_2$. ($b$) The contour plot.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Regression Model (Example)

- Polynomial Regression

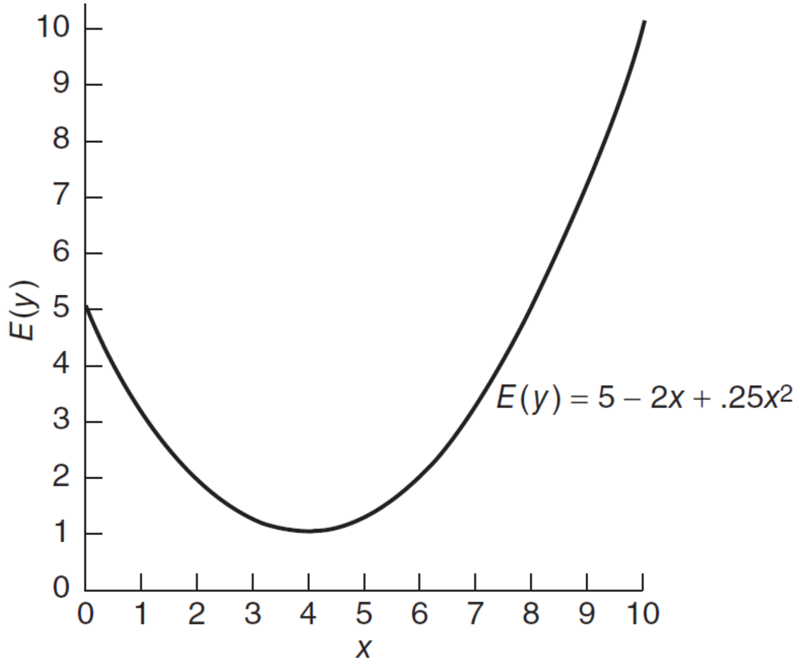


Figure 7.1 An example of a quadratic polynomial.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Regression

- Regression analysis is a statistical technique for investigating and modeling the **relationship between a dependent variable and one or more independent variables**.
  - Dependent variable is denoted by y.
  - Independent variables are denoted by $x_1, x_2, \dots x_n$.
- In this class, we are talking about Simple Linear Regression.
- We can**not** find out any **cause-and-effect relationship** between dependent variable and independent variable(s)

Terms

- X = independent / predictor variable(s) / feature / 독립변수
- Y = dependent / response variable / target / 종속변수
- β_0 = intercept; value of Y when $X = 0$. * β_0 and β_1 are also called regression coefficient.
- β_1 = slope; change in Y when X changes 1 unit.
- ε_i = random error * Assume that the errors follow normal distribution with mean 0 and unknown σ^2.
- e_i = residual
- n = number of observations
- i = i-th observation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Terms (Cont'd)

- When there is hat sign, it is our **estimated** value.
  - $\hat{y}_i$ = estimation
  - $\hat{\beta}_0$ = estimated intercept
  - $\hat{\beta}_1$ = estimated slope

- Finding a regression line means finding the parameters β_0 and β_1.

Simple Linear Regression

- A model with a single regressor x that has a relationship with a response y that is a **straight line**.
- Linear Model: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Expectation:

$$E(Y|X) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x \ (\because E(\epsilon) = 0)$$

- Variance

$$Var(Y|X) = Var(\beta_0 + \beta_1 x + \epsilon) = Var(\epsilon) = \sigma^2$$

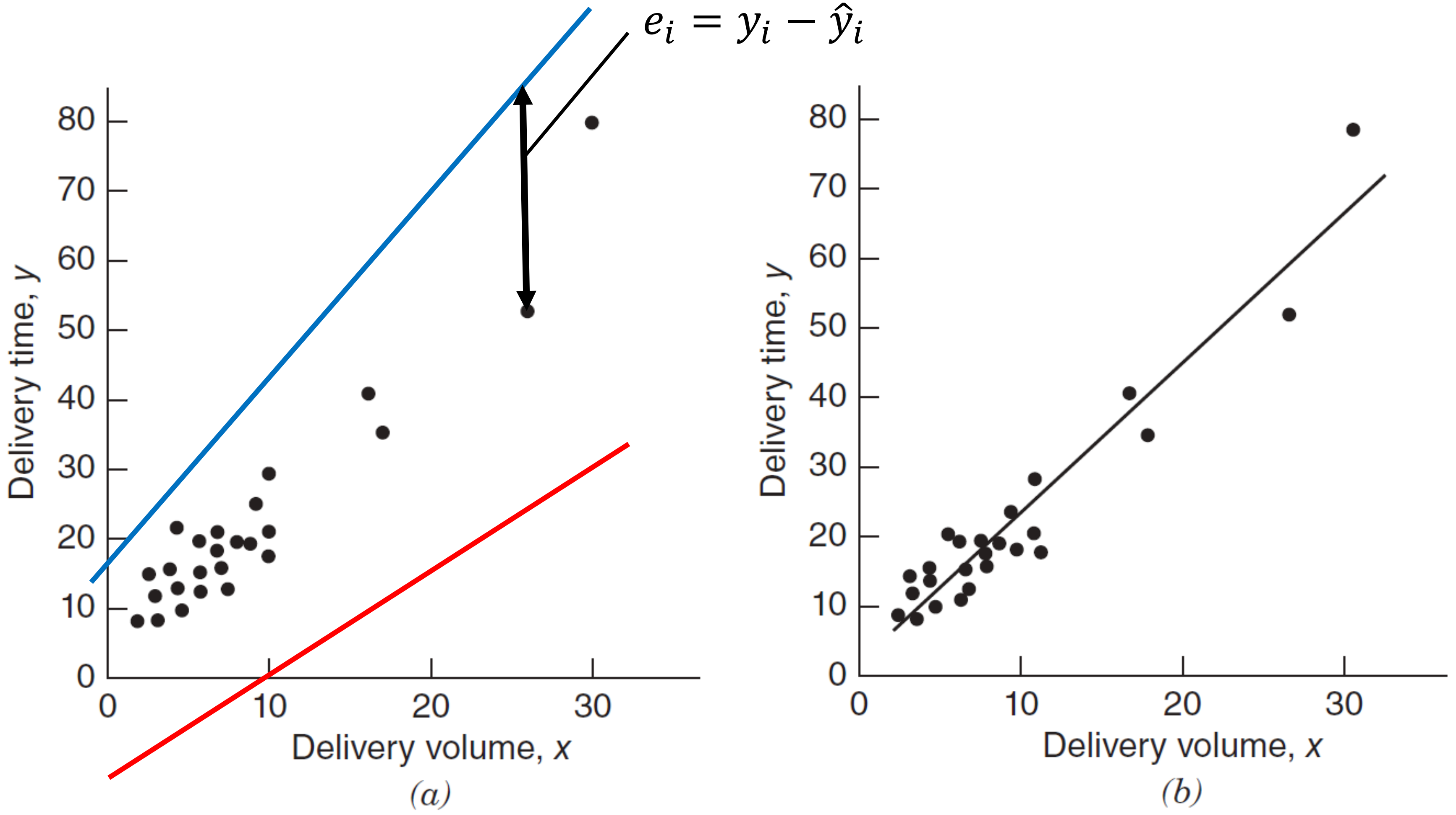
- When we calculate the expectation and variance, β_0, β_1 and x are regarded as **constants**.

Simple Linear Regression

- Given sample, we want to minimize $(y_i - \hat{y}_i)$^2 which is what we called residual sum of squares.

→**Least Squares Method**

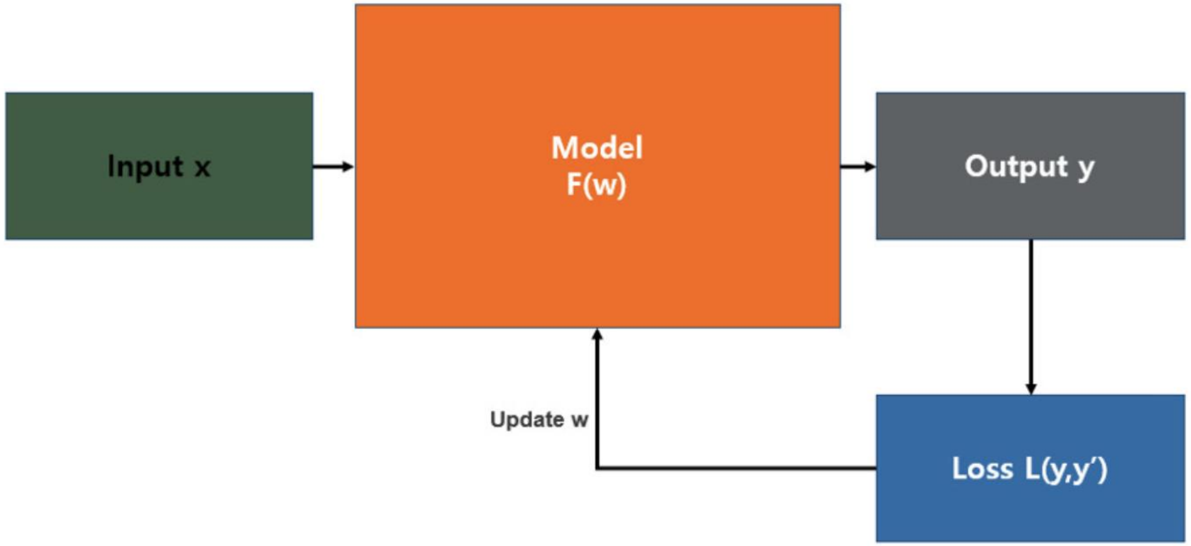
(최소제곱법)



Least Squares Method (최소제곱법)

- We find the **<u>cost function (or loss function)</u>** using residual sum of squares.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2$$



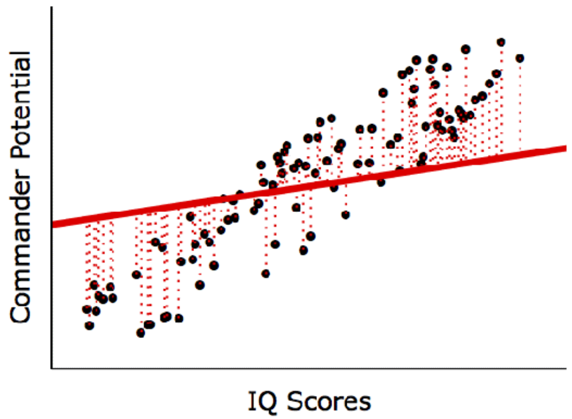
Least Squares Method (최소제곱법)

- Using partial derivative,
  - $\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$
  - $\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

- Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least-squares estimators of the intercept and the slope, respectively.
- Special case of **<u>gradient descent</u> (we will study later on)**.

Applications 1. Data Analysis (survey, experiments)



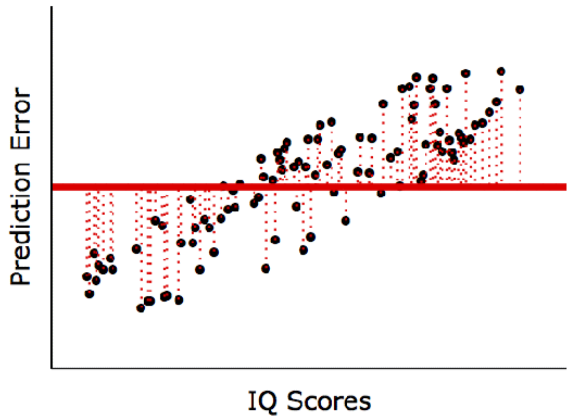


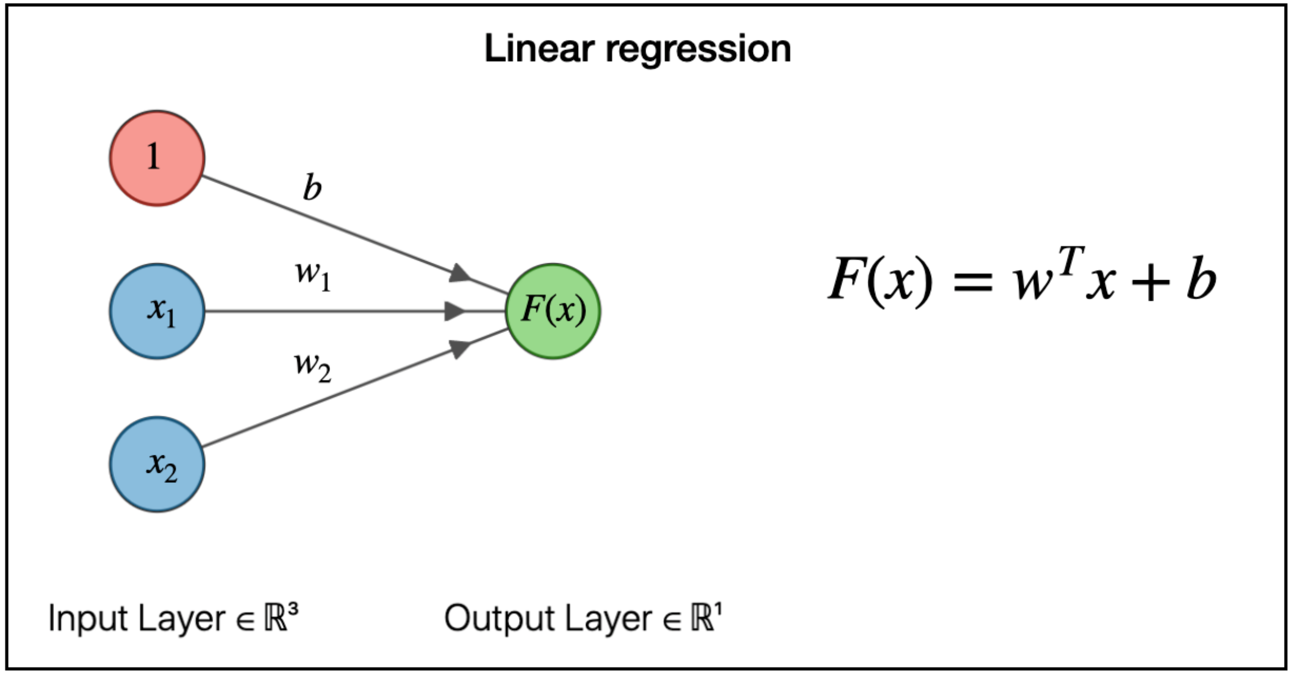
표 22. 스포츠중계시청에 미치는 영향요인

	비표준화 계수		표준화 계수	t	유의확률
	B	표준오차	베타		
상수항	1.634	0.295		5.536	0.000
배구토토 참여경험	0.175	0.048	0.244	3.678	0.000
동료친지와 즐김	0.179	0.064	0.180	2.771	0.006
취미와 여가생활	0.165	0.057	0.191	2.872	0.005
씨름관람경험	-0.191	0.048	-0.247	-3.962	0.000
주변사람의 긍정적시선	0.165	0.068	0.162	2.434	0.016
배당률에 의한 종목선택	0.115	0.053	0.138	2.155	0.032

종속변수: 문항 8-6

R^2=0.865, 수정된 R^2= 0.343, F값 변화량: 4.642(p-0.032)

Applications 2. Neural Network



$$F(x) = w_1 \cdot x_1 + w_2 \cdot x_2 + 1 \cdot b$$