Lecture 3

Brennan Becerra

2024-01-16

Additional Data Sets

.csv $\rightarrow$ comma-separated variables

c1	**c2**
data	data
data	data

.data $\rightarrow$ Text-formatted data

$x \rightarrow$ scan("path to the file of data name")

Linear Regression

> **Linear Regression**
>
> A statistical methodology that utilizes the relation between two or more QUANTITATIVE variables so that one variable can be predicted from other(s).

Ex.

- Sales of product vs amount of advertising expenditure.
- The length of hospital stay of a surgical patient vs the severity of the surgical operation
- Dollar sales of product vs the number of units sold

$$Y = f(x)$$
$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

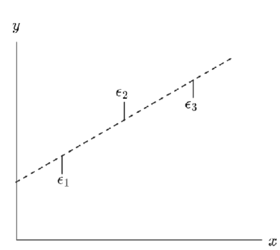
1. Mathematical Functions

$$Y = f(x)$$
$$X: \text{units sold} \qquad Y: \text{dollar sales}$$

Each unit is $2.

Units sold	**Dollar sales**
25	$50
75	$150
100	$200

2. Statistical Function. $Y = f(x) + \epsilon$ where ϵ is the error.



Note: $Y = f(x) + \epsilon$

- Y : Response or Dependent variable
- x : Explanatory or Independent or predictors

Note: If $f(\dot{})$ is linear, we call the model *Linear Regression Model*. If we only have one predictor,

$$\underbrace{Y = \beta_0 + \beta_1 x + \epsilon}_{\text{Simple Linear Regresiion Model}}$$

If we have more than one independent variables, $x_1, x_2, \ldots, x_p$, we write

$$\underbrace{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}_{\text{Multiple Linear Regression Model}}$$

Simple Linear Regression Model

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ ("Bi-variate Data")
Model: $Y = \beta_0 + \beta_1 x + \epsilon, i = 1, 2, \ldots, n$
Goal: Estimate β_0, β_1 from $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
$\implies$ Estimated values of $\beta_0, \beta_1 : \hat{\beta}_0, \hat{\beta}_1$
Assess the estimated values of $\hat{\beta}_0, \hat{\beta}_1$ through test of hypothesis.

1. x_i's: Given constants
2. β_0, β_1 : Unknown Constants
3. ϵ_i : Random variables with mean 0 and variance of σ^2

$$\underbrace{E(\epsilon_i)}_{\text{Expected Value}} = 0, \underbrace{\text{Var}(\epsilon_i)}_{\text{Variance}} = \sigma^2$$

$$\underbrace{Y_i}_{r.v} = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Unknown Constant}} + \underbrace{\epsilon_i}_{r.v}$$

4. $E(Y_i) = Ee(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + E(\epsilon_i)$
   $= \beta_0 + \beta_1 x_i$; *Regression function.* From $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we try to estimate

   $$E(Y_i) = \beta_0 + \beta_1 x_i \rightarrow E(\hat{Y}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

   Note: *Sir. Francis Galton,* 19th century.

Estimation of β_0, β_1.

$$\begin{cases} \epsilon_i \sim N(0, \sigma^2) : \text{Maximum Likelihood Estimation(MLE)} \\ \epsilon_i \sim (0, \sigma^2) : \text{Least Squares Estimation(LSE)} \end{cases}$$

Least Squares Estimation

$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i), i = 1, 2, 3, \ldots, n$
$Q = \sum_{i=1}^{n} \epsilon_i^2$

We want to minimize this with respect to β_0, β_1.

$$\begin{aligned} Q &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

And we can say

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= 2 \cdot \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot -1 = 0 \\ \frac{\partial Q}{\partial \beta_1} &= 2 \cdot \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot -x_i = 0 \end{aligned}$$

Which yields

$$\begin{aligned} \sum_{i=1}^{n} y_i - \hat{n}\beta_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \end{aligned}$$

Normal Equations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta_1 \overline{x}},$$

where $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ and $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

we can rewrite this as

$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \vec{\beta} = \begin{pmatrix} \beta_0 \\ b \\ \beta_1 \end{pmatrix}$$

such that

$$Y = \vec{x}\vec{\beta} + \vec{\epsilon}.$$