Math 170S

Lecture Notes Section 6.5 [*][†]

Linear regression

Instructor: Swee Hong Chan

NOTE: The notes is a summary for materials discussed in the class and is not supposed to substitute the textbook. Please send me an email if you find typos.

---

[*]Version date: Sunday 11$^{\text{th}}$ October, 2020, 20:35.

[†]This notes is based on Hanbaek Lyu's and Liza Rebrova's notes from the previous quarter, and I would like to thank them for their generosity. "*Nanos gigantum humeris insidentes* (I am but a dwarf standing on the shoulders of giants)".

1 Linear regression example

We list the average age of people infected, and killed by COVID-19 in fictional countries:

1. Atlantis, average age: 25, fatality rate: 0.35;
2. Avalon, average age: 47, fatality rate: 0.57;
3. Lemuria, average age: 54, fatality rate: 0.64;
4. Shangri-la, average age: 72, fatality rate: 0.82;
5. Tartarus, average age: 90, fatality rate: 1.

Let $x_1, \ldots, x_5$ be the average age of the infectees in those 5 countries.

Let $y_1, \ldots, y_5$ be the corresponding fatality rate.

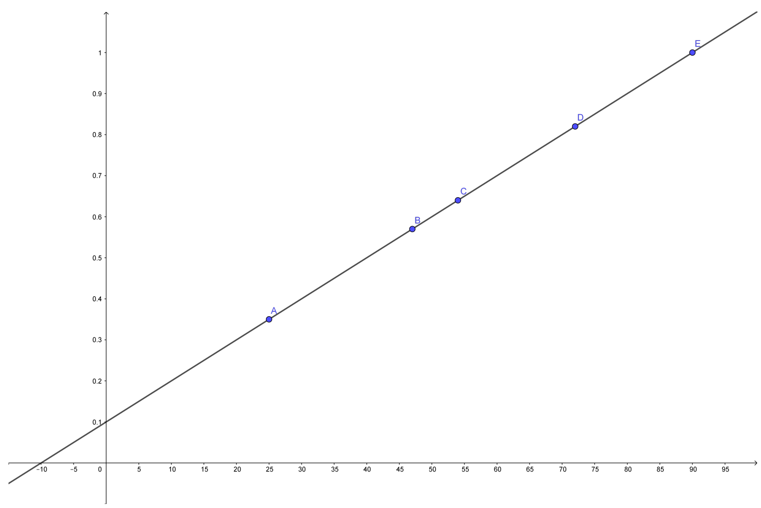

Figure 1: The plot of the average age of people infected by COVID-19 (the x-axes) and the corresponding fatality rate (y-axes).

From the figure, one predicts the relationship between x_i and y_i is

$$y_i = 0.1 + 0.01x_i.$$

We would like to do the same thing for general data.

2 Linear regression

Suppose that X and Y are two unknown **dependent** random variables, and we want to model their relationship. We do it as follows:

- We perform n experiments to get n random samples, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- The **linear regression** is the prediction that

$$Y = \alpha + \beta X + \epsilon, \tag{1}$$

  where α and β are constants, and ϵ is a normal $N(0, \sigma^2)$ random variable independent from X.

- This prediction means that we expect Y and X to have a(n almost) linear relationship.
- The normal random variable $\epsilon \sim N(0, \sigma^2)$ is the **Gaussian noise**, which emulates the (unpredictable yet unavoidable) error made by measurement tools.
- Finally, use MLE to guess $\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}^2$.

3 Linear regression: problems

- **Assumption** Unknown random variables X and Y that obeys a linear relationship.

- **Input:** Samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

- **Problem:** Find $\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}^2$ that best estimates

$$Y = \alpha + \beta X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

- **Method:** Use MLE to estimate $\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}^2$.

Computing the MLE for $\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}^2$ is not hard, but can be time-consuming. Therefore, we provide you the following formulas for $\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}^2$.

4 Formulas for $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\sigma}^2$

Theorem 1. *the MLEs $\widehat{\alpha}, \widehat{\beta}, \widehat{\sigma}^2$ are given by*

$$\widehat{\beta} = \frac{E - \frac{AB}{n}}{C - \frac{A^2}{n}}; \qquad \widehat{\alpha} = \frac{B}{n} - \widehat{\beta}\,\frac{A}{n};$$

$$\widehat{\sigma}^2 = \frac{D}{n} - \left(\frac{B}{n}\right)^2 - \widehat{\beta}\,\frac{E}{n} + \widehat{\beta}\,\frac{AB}{n^2},$$

where

$$A := \sum_{i=1}^{n} x_i = x_1 + \ldots + x_n;$$

$$B := \sum_{i=1}^{n} y_i = y_1 + \ldots + y_n;$$

$$C := \sum_{i=1}^{n} x_i^2 = x_1^2 + \ldots + x_n^2;$$

$$D := \sum_{i=1}^{n} y_i^2 = y_1^2 + \ldots + y_n^2;$$

$$E := \sum_{i=1}^{n} x_i y_i = x_1 y_1 + \ldots + x_n y_n.$$

Remark 2. Note that the constant α in our notes is presented as α_1 in the textbook, which explains why the formula for $\widehat{\alpha}$ in our theorem is off from the formula in textbook by a positive constant.

5 Example: linear regression

Let $x_1, \ldots, x_n$ be the midterm score of 10 students in a fictional statistics class:

$$70 \quad 74 \quad 72 \quad 68 \quad 58 \quad 54 \quad 82 \quad 64 \quad 80 \quad 61.$$

Let $y_1, \ldots, y_n$ be the final score of the 10 students:

$$77 \quad 94 \quad 88 \quad 80 \quad 71 \quad 76 \quad 88 \quad 80 \quad 90 \quad 69.$$

The key values A, B, C, D, E are given by

$$A = \sum_{i=1}^{n} x_i = 70 + 74 + 72 + 68 + 58 + 54 + 82 + 64 + 80 + 61$$

$$= 683;$$

$$B = \sum_{i=1}^{n} y_i = 77 + 94 + 88 + 80 + 71 + 76 + 88 + 80 + 90 + 69$$

$$= 813;$$

$$C = \sum_{i=1}^{n} x_i^2$$

$$= 70^2 + 74^2 + 72^2 + 68^2 + 58^2 + 54^2 + 82^2 + 64^2 + 80^2 + 61^2$$

$$= 47,405;$$

$$D = \sum_{i=1}^{n} y_i^2$$

$$= 77^2 + 94^2 + 88^2 + 80^2 + 71^2 + 76^2 + 88^2 + 80^2 + 90^2 + 69^2$$

$$= 66,731;$$

$$
\begin{aligned}
E &= \sum_{i=1}^{n} x_i y_i \\
&=(70)(77) + (74)(94) + (72)(88) + (68)(80) + (58)(71)+ \\
&\quad (54)(76) + (82)(88) + (64)(80) + (80)(90) + (61)(69) \\
&= 56,089;
\end{aligned}
$$

The MLEs are then given by

$$
\begin{aligned}
\widehat{\beta} &=\frac{E - \frac{AB}{n}}{C - \frac{A^2}{n}} \;=\; \frac{56,089 - (683)(813)/10}{47,405 - (683)(683)/10} \;=\; 0.742. \\
\widehat{\alpha} &=\frac{B}{n} - \widehat{\beta}\,\frac{A}{n} \;= 813/10 - (0.742)(683/10) \;=\; 30.6214. \\
\widehat{\sigma}^2 &=\frac{D}{n} - \left(\frac{B}{n}\right)^2 - \widehat{\beta}\,\frac{E}{n} + \widehat{\beta}\,\frac{AB}{n^2} \\
&=66,731/10 - (813/10)^2 - (0.742)(56,089)/10 + (0.742)(683) \\
&=21.77638.
\end{aligned}
$$