LECTURE 4

BOGDAN ICHIM

Normal Simple Linear Model

Simple Linear Model: One makes the following assumptions

(1) **Linearity**: The conditional mean of Y given x is a linear function of x, i.e.

$$E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i.$$

(2) **Constant Error Variance**: The variability of Y is the same for all values of x, i.e.

$$\mathrm{var}(Y_i|X = x_i) = \sigma^2 \quad \Longleftrightarrow \quad \mathrm{var}(\epsilon_i) = \sigma^2.$$

(3) **Uncorelatedness of Errors**: The errors are not linearly associated, i.e.

$$\mathrm{cov}(Y_i, Y_j) = 0 \quad \Longleftrightarrow \quad \mathrm{cov}(\epsilon_i, \epsilon_j) = 0, \quad \forall i \neq j.$$

Normal Simple Linear Model: One can make the extra assumption that the errors are normally distributed, that is

(4) **Normality of Errors:** For a given value of x, Y has a normal distribution, i.e.

$$Y_i|X = x_i \sim N(\beta_o + \beta_1 x_i, \sigma^2) \quad \Longleftrightarrow \quad \epsilon_i \sim N(0, \sigma^2).$$

Remark 1. In general, if two random variables a, b are independent, then they are uncorrelated, that is independence implies $\mathrm{cov}(a, b) = 0$.

Please note that the converse does not stand, with the following exception: If two **normal** random variables are uncorrelated, then they are independent, i.e. $\mathrm{cov}(a, b) = 0 \Longleftrightarrow a$, b are independent. Therefore it is common to use an alternative of assumption 3

(3') **Independence of Errors**: Knowledge that one response Y_i is larger than expected does not give us information about whether a different Y_j is larger (or smaller) than expected.

Least Squares Estimation of β_0 and β_1

Our goal is to obtain coefficient estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that the linear model fits the data well, i.e. the resulting line is as *close* as possible to the data points. There are a number of ways of measuring *closeness*, so there are several possible estimators of β_0 and β_1 which could be used.

A standard approach is **least squares estimation** (i.e. using the Euclidean distance induced by the L_2 norm).

Let $(x_1, y_1), \ldots, (x_n, y_n)$ represent n observed sample pairs, each of which consists of a measurement of X and a measurement of Y.

Let
$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$
be the **estimated regression line**, i.e. $\widehat{y}_i$ will be the **prediction** for Y based on the i-th value of X.

Then
$$\widehat{\epsilon}_i = y_i - \widehat{y}_i$$
represents the i-th **residual**, i.e. the difference between the i-th observed value of Y and the i-th value of Y predicted by our linear model.

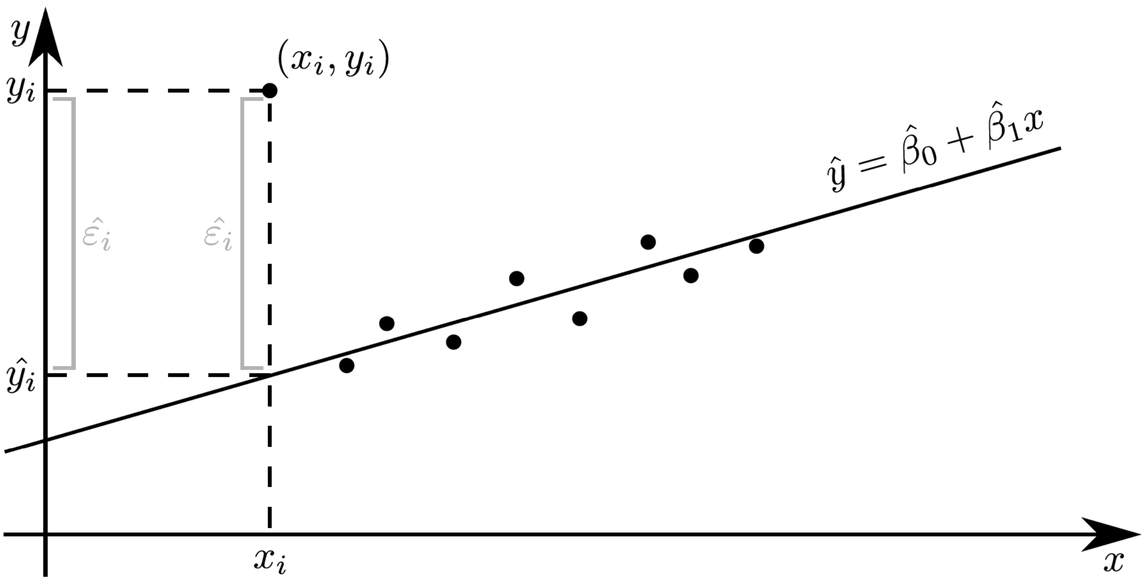

FIGURE 1. Estimated regression line and residuals

We define the **residual sum of squares** as
$$\text{RSS} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

The least squares approach finds $\widehat{\beta}_0$ and $\widehat{\beta}_1$ which minimize the RSS. We have to solve the optimization problem
$$\min_{\widehat{\beta}_0, \widehat{\beta}_1} \text{RSS} = \min_{\widehat{\beta}_0, \widehat{\beta}_1} \sum_{i=0}^{n} \widehat{\epsilon}_i^2 = \min_{\widehat{\beta}_0, \widehat{\beta}_1} \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

Remark 2. If $(\widehat{\beta}_0, \widehat{\beta}_1)$ is a solution to the optimization problem, then

$$1)\frac{\partial \operatorname{RSS}(\widehat{\beta}_0, \widehat{\beta}_1)}{\partial \widehat{\beta}_0} = 0;$$

$$2)\frac{\partial \operatorname{RSS}(\widehat{\beta}_0, \widehat{\beta}_1)}{\partial \widehat{\beta}_1} = 0.$$

We compute

$$\begin{cases} \dfrac{\partial \operatorname{RSS}(\widehat{\beta}_0, \widehat{\beta}_1)}{\partial \widehat{\beta}_0} = \displaystyle\sum_{i=1}^{n} 2(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \cdot (-1) \\ \dfrac{\partial \operatorname{RSS}(\widehat{\beta}_0, \widehat{\beta}_1)}{\partial \widehat{\beta}_1} = \displaystyle\sum_{i=1}^{n} 2(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \cdot (-x_i) \end{cases}$$

Together with Remark 2 we obtain a system of linear equations

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \displaystyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i & (1) \\ \widehat{\beta}_0 \displaystyle\sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i & (2) \end{cases}$$

In matrix notation we have

$$A \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = B,$$

where

$$A = \begin{pmatrix} n & \displaystyle\sum_{i=1}^{n} x_i \\ \displaystyle\sum_{i=1}^{n} x_i & \displaystyle\sum_{i=1}^{n} x_i^2 \end{pmatrix}, \quad B = \begin{pmatrix} \displaystyle\sum_{i=1}^{n} y_i \\ \displaystyle\sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Let

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

be the **sample means**.

Remark 3. We always have

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0,$$

which in turn implies

$$c \sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

for any constant c.

We compute

$$\begin{aligned}
\det\, A &= n\sum_{i=1}^{n} x_i^2 - n\overline{x}\sum_{i=1}^{n} x_i \\
&= n\sum_{i=1}^{n}(x_i^2 - x_i\overline{x}) \\
&= n\sum_{i=1}^{n} x_i(x_i - \overline{x}) \\
&= n(\sum_{i=1}^{n} x_i(x_i - \overline{x}) - \sum_{i=1}^{n}\overline{x}(x_i - \overline{x})) \\
&= n\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x}) \\
&= n\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&\Longrightarrow \boxed{\det\, A = n\sum_{i=1}^{n}(x_i - \overline{x})^2}
\end{aligned}$$

Remark 4. $\det\, A > 0 \Longrightarrow \exists!$ solution of the system.

Cramer's Rule:

$$\widehat{\beta}_1 = \frac{\det\, A_1}{\det\, A}, \text{ where } A_1 = \begin{pmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

We compute

$$\begin{aligned}
\det\, A_1 &= n\sum_{i=1}^{n} x_i y_i - n\overline{y}\sum_{i=1}^{n} x_i \\
&= n\sum_{i=1}^{n}(x_i y_i - x_i\overline{y}) = n\sum_{i=1}^{n} x_i(y_i - \overline{y}) \\
&= n(\sum_{i=1}^{n} x_i(y_i - \overline{y}) - \sum_{i=1}^{n}\overline{x}(y_i - \overline{y})) \\
&= n\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})
\end{aligned}$$

$$\xrightarrow{\text{Cramer}} \boxed{\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

Using equation (1) we further get

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

There is one remaining parameter to estimate, that is the error variance σ^2. The usual estimator is

$$\widehat{\sigma}^2 = \frac{\text{RSS}}{n-2}.$$

The estimator for the standard deviance of ϵ is called the **residual standard error**

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}.$$

Remark 5. $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2$ depend on the observation pairs used!!!

Interpretation of $\widehat{\beta}_1$

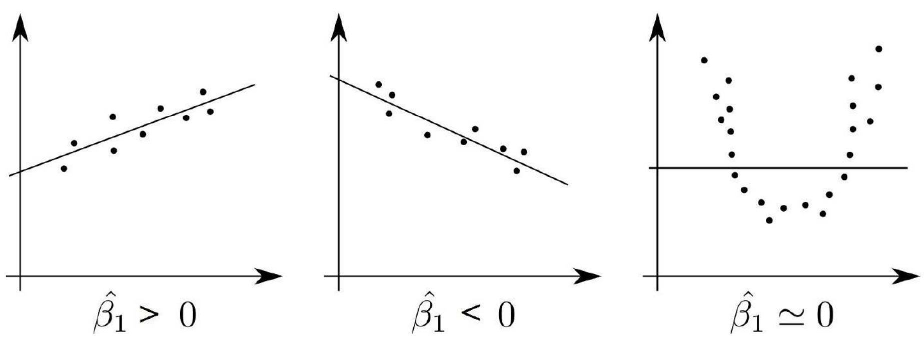

FIGURE 2. X and Y are not linearly related for $\widehat{\beta}_1 \simeq 0$

Suppose that $|\widehat{\beta}_1|$ is significantly different from 0.

- This does not mean that X and Y are linearly related.
- This does not mean that X and Y are casually related, that is, changes in X cause changes in Y or vice versa.