LECTURE 3

BOGDAN ICHIM

The Linear Model

(The Linear Regression Model)

The general formula for a **linear model** is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

where Y and ϵ are random variables. $X_1, X_2, \ldots, X_p$ are series of numbers (precisely known, determined).

Remark 1. We use the following terminology

- Y is called **the response variable** or **the prediction**;
- $X_1, \ldots, X_p$ are called **explanatory variables** or **predictors**;
- ϵ is called **the error term** or **the residual**.

Moreover, we distinguish the following

- **Computable Model**: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$;
- **Real Life Model**: Everything which is not explained by the computable model is quantified by the error term ϵ.

Simple Linear Model

(Simple Linear Regression)

In the case of a simple linear model we have a single predictor. Then the above formula becomes

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Let (x_i, y_i) for $i = \overline{1,n}$ represent n observation pairs (known from data). We assume that:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ \ \forall i = \overline{1,n}$$

where Y_i and ϵ_i are **families of random variables**.

Remark 2. In the above formula

- $\beta_0 + \beta_1 x_i$ is the **systematic (deterministic) part** of the model;
- ϵ_i is the **random part** of the model is a random variable;
- Y_i is the **response** and is a random variable;
- y_i is the **observed value (realization)** of the random variable Y_i.

We further assume that:

$$\begin{cases} E(\epsilon_i) = 0 \quad \text{(the mean of the random variable } \epsilon_i \text{ is 0)} \\ var(\epsilon_i) = \sigma^2 \quad \text{(the variance of the random variable } \epsilon_i \text{ is constant)} \end{cases}$$

and that the random variables $\epsilon_1, \dots, \epsilon_n$ are uncorrelated (i.e. $\text{cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$).

Equivalently, we can write the model as:

$$\begin{cases} E(Y_i|X = x_i) = \beta_0 + \beta_1 X \quad \text{(straight line relationship)} \\ \text{var}(Y_i|X = x_i) = \sigma^2 \quad \text{(constant variance)} \end{cases}$$

where, given the values $x_1, \dots, x_n$, the random variables $Y_1, \dots, Y_n$ are uncorrelated.

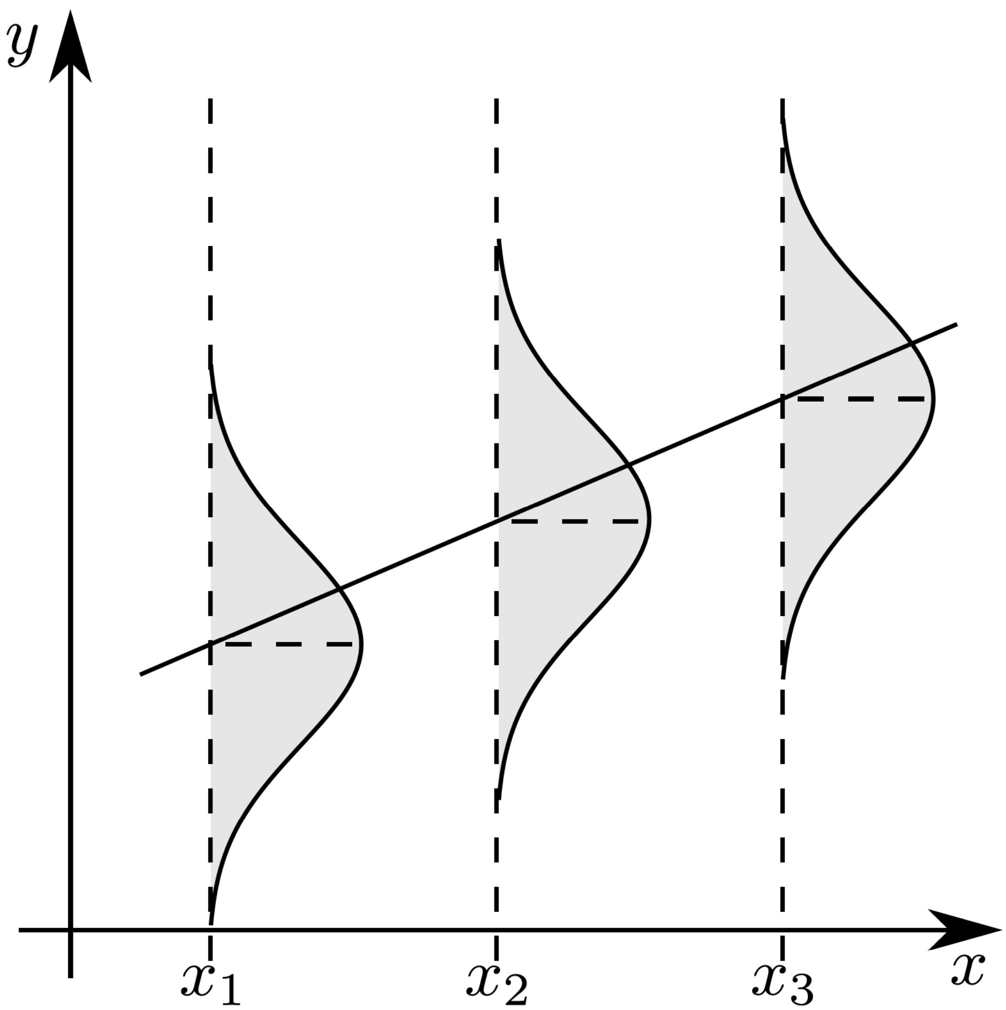

FIGURE 1. Conditional distributions of Y_i given $X = x_i$

Interpretation of the model parameters

Remark 3. We use the following terminology

- $\beta_0 = E(Y|X=0)$ is called **intercept** and represents the expected value (mean) of Y when $X=0$.
- $\beta_1 = E(Y|X=x+1) - E(Y|X=x)$ is called **gradient (or slope)** and represents the amount by which the mean of Y given $X=x$ increases when x increases by one unit.
- σ^2 is the **error variance** and represents the variability of the response in the vertical direction around the linear model line.

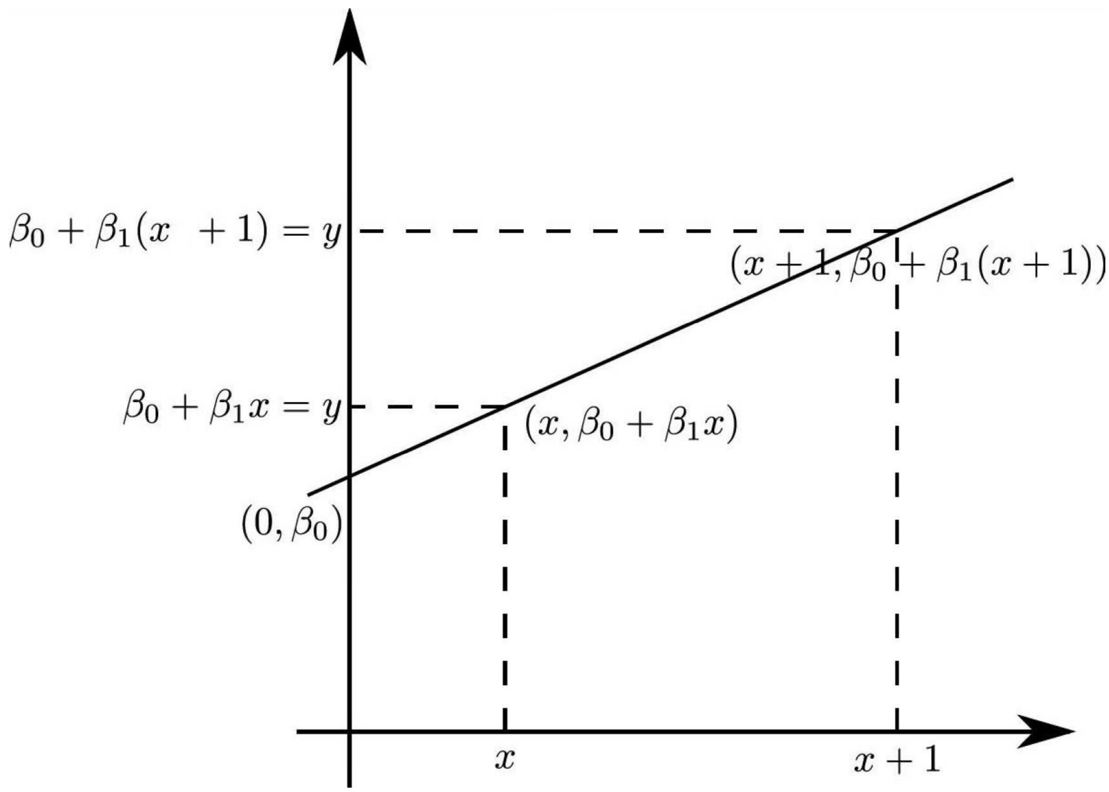

FIGURE 2

The **estimation** of the model parameters is done using n samples for which we have recorded both the levels of X and Y, that is, from the bidimensional series of data

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

Looking at Scatter Plots

Before fitting a linear model we should look at the scatter plot of Y against x.

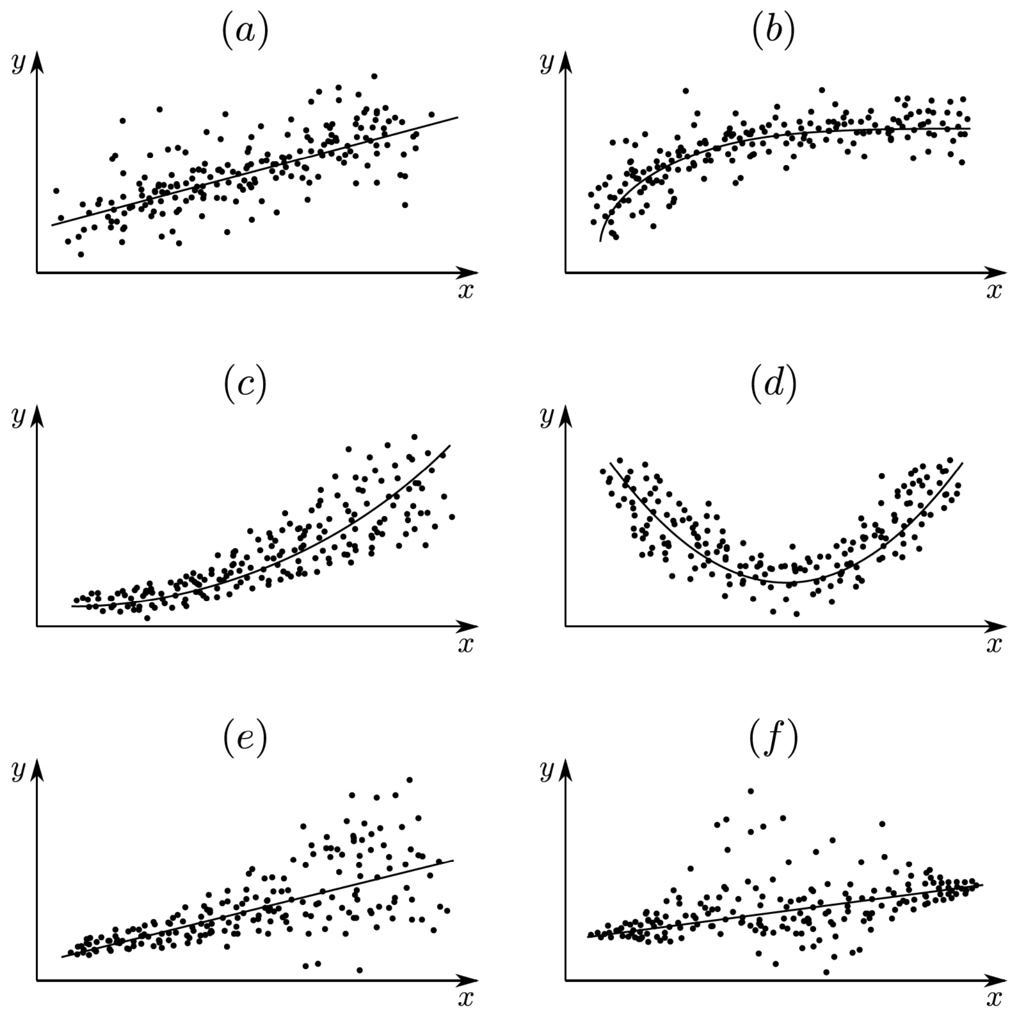

FIGURE 3. Scatter plots. (a) (approximately) linear relationship, (approximately) constant spread about line; (b) non-linear relationship, constant spread about curve; (c) non-linear relationship, increasing spread about curve; (d) quadratic relationship, constant spread about curve; (e) linear relationship, increasing spread about line; (f) linear relationship, non-constant spread about line.

Some questions to ask (and answer)

i. Is the relationship between Y and X approximately linear?

ii. Suppose that we draw a straight line (or a curve if the relationship is non-linear) through the data. Is the variability of Y (the vertical spread) around this line approximately constant?

iii. Are there any points which do not appear to fit in with the general pattern of the rest of the data (that is potential **outliers**)?

(A) See Figure 3 plot (a). The assumptions of linearity and constant error variance appear to be reasonable. Therefore, we just fit a simple linear regression model to these data.

(B) The assumption of linearity is not reasonable, but

(1) The relation between Y and x is monotonic (increasing);

(2) The variability in Y is approximately constant for all values of x.

See for example Figure 3 plot (b). One can try to transform x to straighten the scatter plot, because transforming x will not affect the vertical spread of the points.

Things to try in this case

$$x \to \log x, \quad x \to \sqrt{x}.$$

(C) The assumption of linearity is not reasonable, but

(1) The relation between Y and x is monotonic (increasing);

(2) The variability in Y increases as x increases.

See for example Figure 3 plot (c). One can try to transform Y to straighten the scatter plot. Transforming Y will affect the vertical spread of the points. We may be able to find a transformation of Y which both straightens the plots and makes the variability constant.

Things to try in this case

$$Y \to \log Y, \quad Y \to \sqrt{Y}.$$

(D) The assumption of linearity is not reasonable, and

(1) The relationship between Y and x is not monotonic;

(2) The variability in Y is approximately constant for all values of x.

See for example 3 plot (d). In such a case one can try to fit a model of the form

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X^i + \epsilon.$$

(E) See Figure 3 plot (e). The assumption of linearity is reasonable but the variability in Y increases as x increases. A transformation of Y may be able to make the variability of Y approximately constant, however it may produce a non-linear relationship. Transforming both Y and x might work.

(F) See Figure 3 plot (f). The assumption of linearity is reasonable but the variability in Y is small for extreme (small or large) values of x and large for middling values of x. The comments made above in case (E) also apply here.

Example 4. The Cobb-Douglas production function

$$Y = A \cdot L^{\beta} \cdot K^{\alpha}$$

where

Y = total production (the value of all goods produced in a year),
L = labor (the total number of person-hours worked in a year),
K = capital (machinery, equipment, buildings),
A = productivity.

The Cobb-Douglas production function can be estimated using the following **linear** expression

$$\ln Y = \ln A + \beta \ln L + \alpha \ln K + \epsilon.$$

Remark 5. In the original article, Cobb-Douglas have estimated $\alpha = 0.25$ and $\beta = 0.75$ such that $\alpha + \beta = 1$.