BME 1132
Probability and Biostatistics

Instructor: Ali AJDER, *Ph.D.*

Week-13

- Introduction
- Simple Linear Regression
- Correlation
- Least-squares line/regression
- Residuals

Introduction-1

Scientists and engineers often collect data in order to determine the nature of the **relationship between two quantities.**

For example, a biomedical engineer may want to investigate the relationship between sodium chloride (salt) consumption and blood pressure among elderly people (e.g., above 65 years old).

The sodium chloride intake level x and the systolic blood pressure y are recorded.

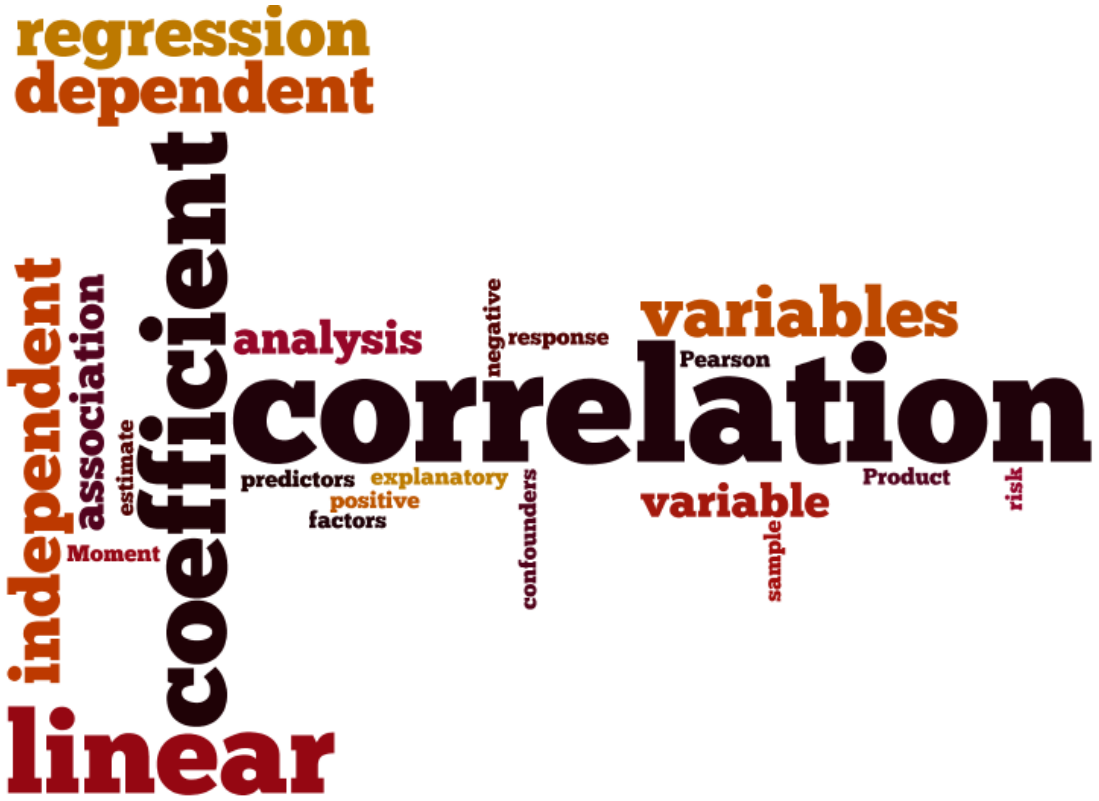
The experiment thus generates bivariate data; a collection of ordered pairs $(x_1, y_1), \ldots (x_n, y_n)$

If the ordered pairs of measurements fall approximately along a straight line when plotted, the data can be used to compute an equation for the line that “best fits” the data.

Introduction-2

The methods of correlation and simple linear regression are used to analyze bivariate data to:

- Determine **whether a straight-line fit** is appropriate;
- Compute the **equation** of the line if appropriate;
- Use that linear equation to draw **inferences** about the relationship between the two quantities.



Correlation

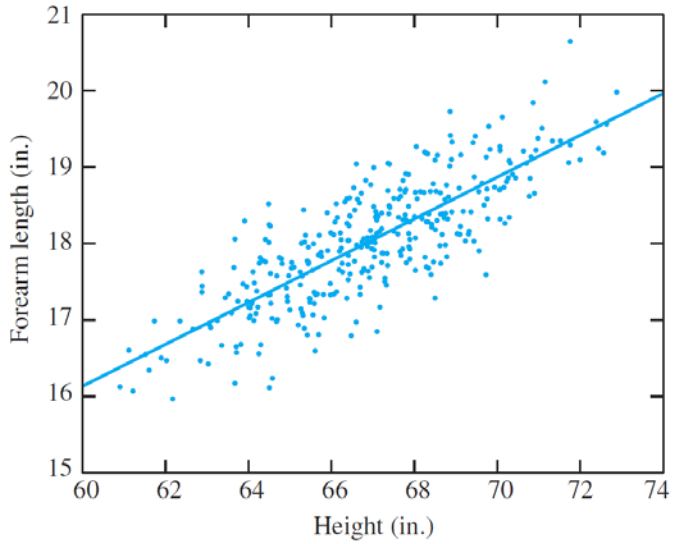
One of the earliest applications of statistics was to study the variation in physical characteristics in human populations.

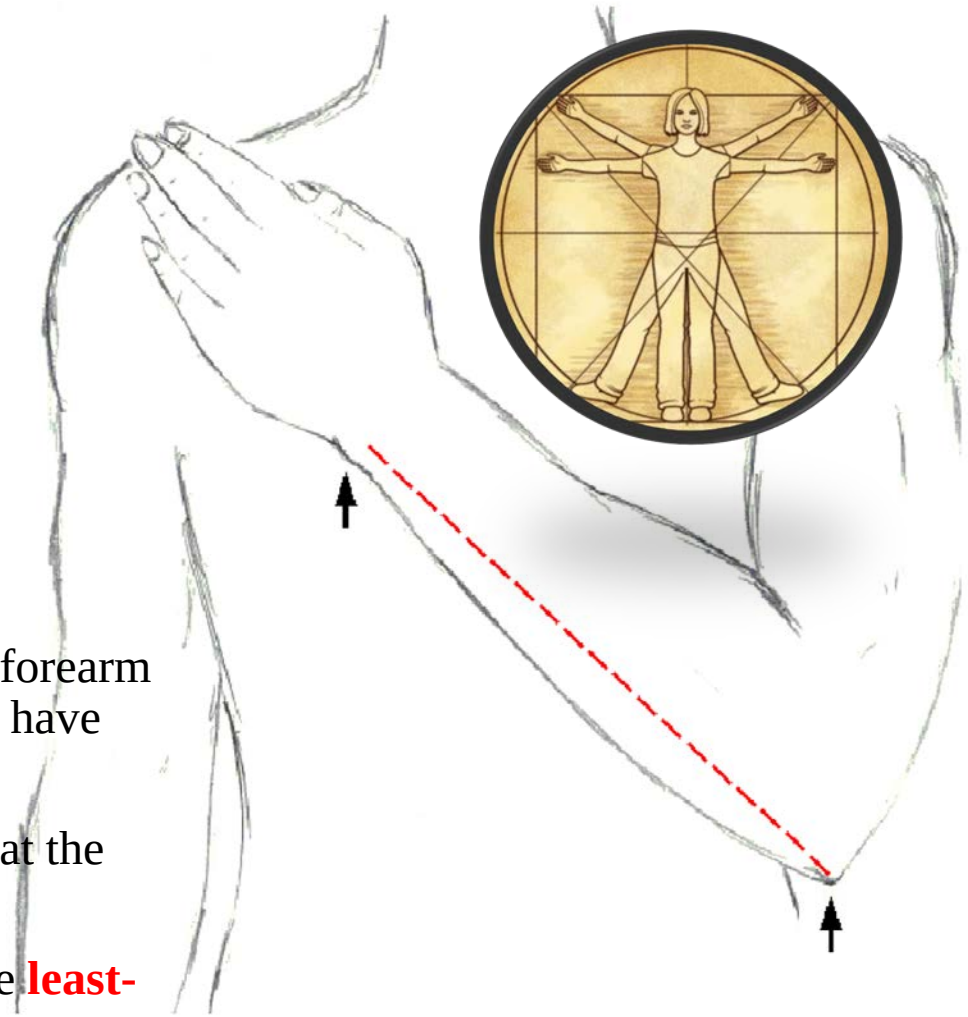
To this end, statisticians invented a quantity called the **correlation coefficient** is a way of describing how closely related two variables are.

The first published correlation coefficient was due to the English statistician Sir **Francis Galton**, who in 1888 measured the **heights** and **forearm lengths** of 348 adult men.

If we denote the height of the i_{th} man by x_i, and the length of his forearm by y_i, then Galton's data consist of 348 ordered pairs (x_i, y_i).

Example 1





This is a plot of **height versus forearm length** for men.

We say that there is a **positive association** between height and forearm length. This is because the plot indicates that taller men tend to have longer forearms.

The slope is roughly constant throughout the plot, indicating that the points are clustered around a straight line.

The line superimposed on the plot is a special line known as the **least-squares line**.

Correlation

When examining a scatterplot of bivariate data, we look at the,

- ✓ **direction** of the relationship (positive or negative),
- ✓ **strength** of the relationship,

and then we find a line that best fits the data.

The degree to which the points in a scatterplot tend to cluster around a line reflects the strength of the linear relationship between x and y.

The visual impression of a scatterplot can be misleading in this regard, because changing the scale of the axes can make the clustering appear tighter or looser.

For this reason, we define the **correlation coefficient**, which is a numerical measure of the strength of the linear relationship between two quantitative variables.

Note: In computing correlation, we can only use quantitative data.

Correlation Coefficient

The **correlation coefficient** is usually denoted by the letter $\boldsymbol{r}$.

There are several equivalent formulas for r.

Let $(x_1, y_1), \ldots (x_n, y_n)$ represent n points on a scatterplot.

Compute the **means** and the **standard deviations** of the x's and y's.

Then convert each x and y to standard units. That is, compute the $\boldsymbol{z-scores}$:

$$(x_i - \bar{x})/s_x$$

$$(y_i - \bar{y})/s_y$$

Computing r

The correlation coefficient is the average of the products of the $\boldsymbol{z-scores}$, except that we divide by $\boldsymbol{n-1}$ instead of n.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\left(x_i-\bar{x}\right)/_{s_x}\right)\left(\left(y_i-\bar{y}\right)/_{s_y}\right)$$

The following formula is easier for calculations by hand:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

Properties of r

- ✓ It is a mathematical fact that r is always between -1 and 1.
- ✓ **Positive** values of r indicate that the least squares line has a **positive slope**. The greater values of one variable are associated with greater values of the other.
- ✓ **Negative** values of r indicate that the least squares line has a **negative slope**. The greater values of one variable are associated with lesser values of the other.
- ✓ Values of r close to -1 or 1 indicate a **strong** linear relationship; values of r close to 0 indicate a **weak** linear relationship.
- ✓ When r is equal to -1 or 1, then all the points on the scatterplot lie **exactly on a straight line** (when there is a perfect linear relationship).

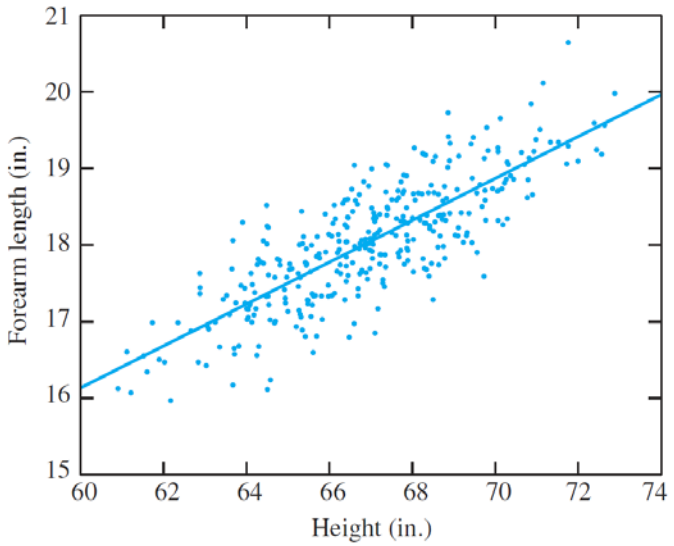
Technical Note: If the points lie exactly on a horizontal or a vertical line, the correlation coefficient is **undefined**, because one of the standard deviations is equal to zero.

More Comments

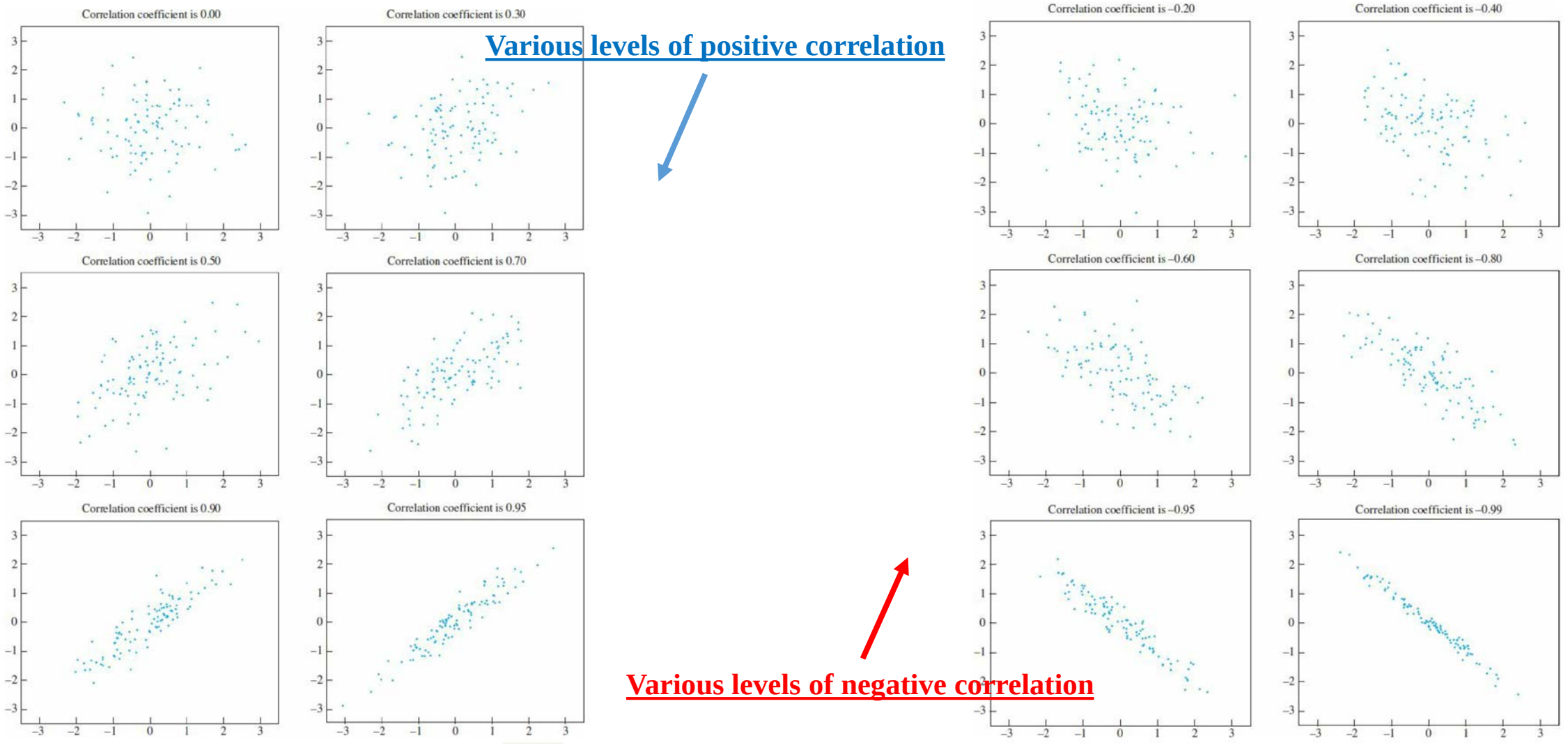
If $r \neq 0$, then x and y are said to be ***correlated***.

If $r = 0$, then x and y are ***uncorrelated***.

For the scatterplot of height versus forearm length, $r = 0.80$.



More Comments



Various levels of positive correlation

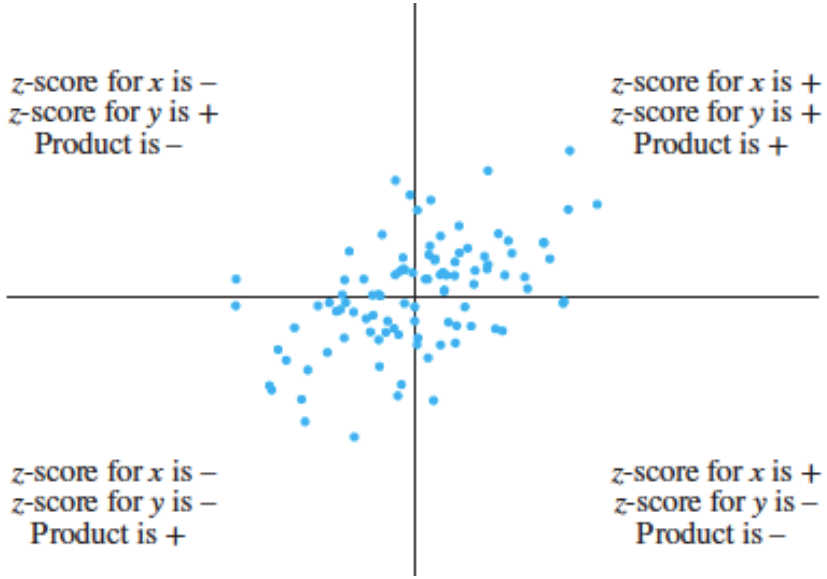
Various levels of negative correlation

How the Correlation Coefficient Works

Why does the formula for the correlation coefficient r measure the strength of the linear association between two variables?

In this scatterplot, the origin is placed at the point of averages.

More points in the first and third quadrants than in the second and fourth, so the correlation will be positive.



The correlation coefficient should only be used when the relationship between the x and y is linear.

Otherwise the results can be misleading.

For example, two variables that are related **quadratically** will have a **correlation coefficient of zero**—they are certainly related, but **not** linearly.

Correlation Is Not Causation

For children, **vocabulary size** is strongly correlated with **shoe size**.

However, learning new words **does not cause** feet to grow, nor do growing feet cause one's vocabulary to increase.

There is a third factor, namely **age**, that is **correlated with** both **shoe size** and **vocabulary**.

Older children tend to have both larger shoe sizes and larger vocabularies, and this results in a positive correlation between vocabulary and shoe size.

This phenomenon is known as **confounding**. *Confounding occurs when there is a third variable that is correlated with both of the variables of interest, resulting in a correlation between them.*

Simply because two variables are correlated with each other, we cannot assume that a change in one will tend to **cause** a change in the other.

Before we can conclude that two variables have a **causal relationship,** we must **rule out the possibility of confounding.**

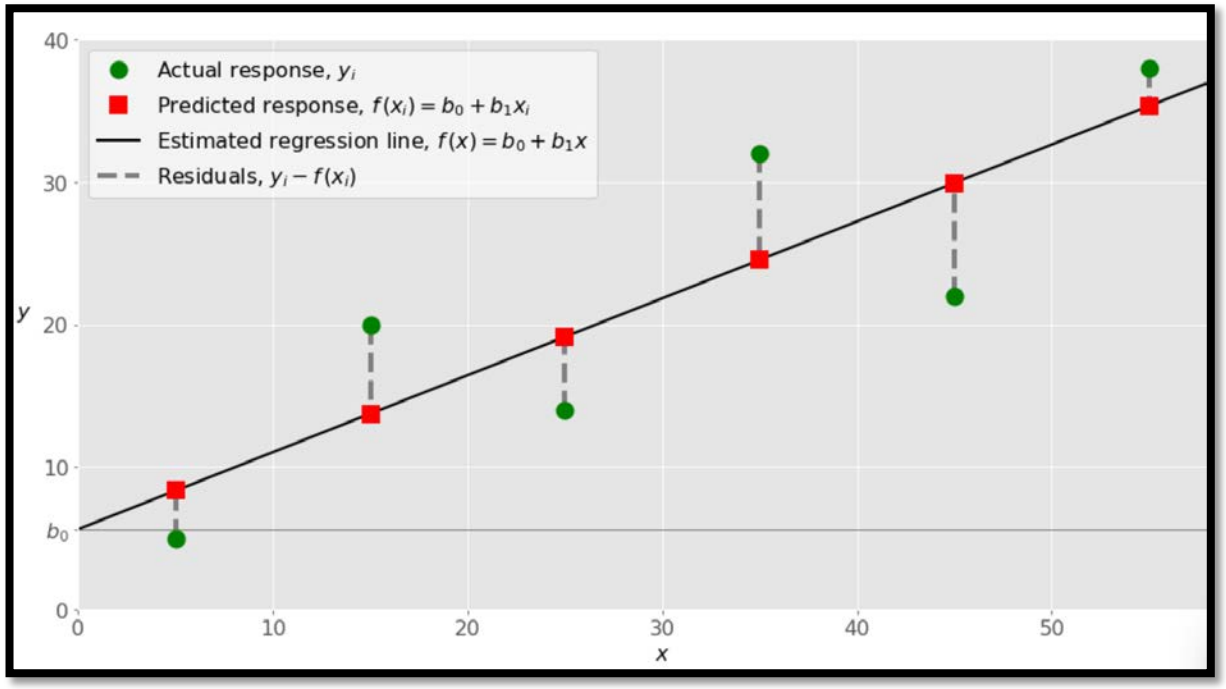
The Least-Squares Line

The estimated or predicted response, $f(x_i)$, for each observation $i = 1, \dots, n$ should be as close as possible to the corresponding actual response y_i.

The differences $y_i - f(x_i)$ for all observations $i = 1, \dots, n$, are called the **residuals**.

Regression is about determining the **best predicted weights**, that is the weights corresponding to the **smallest residuals**.

To get the best weights, we usually minimize the **sum of squared residuals** (SSR) for all observations



$i = 1, \dots, n$: $SSR = \sum_i \left(y_i - f(x_i)\right)^2$. This approach is called the **method of ordinary least squares**.

https://realpython.com/linear-regression-in-python/

Questions?