Linear Regression

Summary

1. Regression is all about prediction.
2. We can calculate the strength of an association (the *correlation*) between two variables.
3. "Fundamental Theorem of Statistics": **Prediction = Reality + Error**

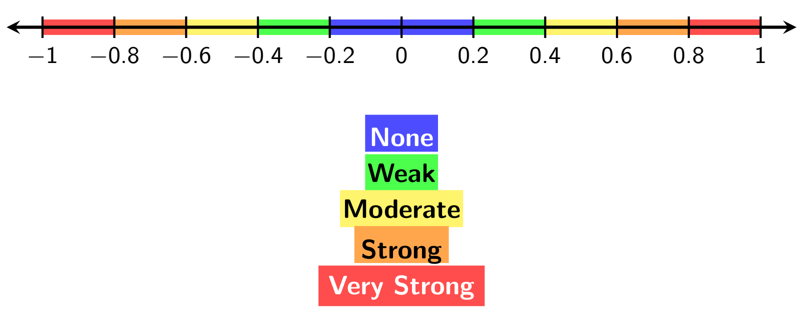
Linear Correlation Coefficient

The **linear correlation coefficient**, r, is a numerical value with $-1 \leq r \leq 1$ that measures the type of linear correlation of a bivariate data set.

- $r > 0$: positive linear correlation
- $r = 0$: no linear correlation
- $r < 0$: negative linear correlation

$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \cdot \sum(y-\overline{y})^2}}$$

- Closer r is to 1 (or -1) $\longrightarrow$ the more the data points "fall in line."
- Closer r is to 0 $\longrightarrow$ the more the data points resemble a "cloud"



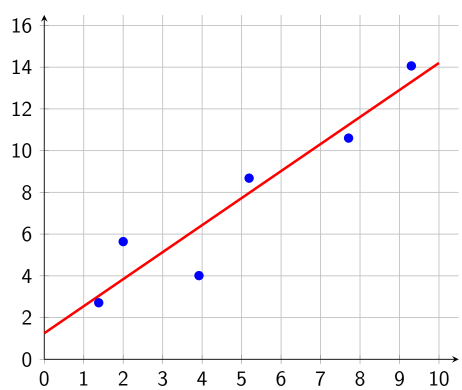
Note: These interpretations are not universal.

Example 1. Find and interpret the linear correlation coefficient.

x	y
7.6	19.1
9.2	22.9
3.3	10.3
1.1	6.6
3.7	10.6
3.9	11.3
4.6	12.9
2.3	8.6
5.1	15.2
5.3	15.1
2.5	13
3.4	11.2
3.1	10.6
1.7	6.8
3.7	13.7

Given our data set, we also want to be able to predict values.
To do this, we can create the **least squares regression equation**, (also called the *line of best fit*)
Which **minimizes** the total squared distance each data point is from the prediction line:

$$\hat{y} = mx + b$$



$$m = r\left(\frac{\sigma_y}{\sigma_x}\right) \quad \text{and} \quad b = \overline{y} - m\left(\overline{x}\right)$$

Example 2. Find the least squares regression equation for the data set from Example 1. Round your parameters to 3 decimal places.

Example 3. Use the regression equation from the previous example to predict the values of the response variable for each given explanatory variable.

(a) $x = 6$

(b) $x = 11$

Residual Error

Suppose we obtain an actual data point when $x = 6$ and observe that $y = 16$.

$$\text{Residual } (\epsilon) = \text{observed value} - \text{expected value}$$

We could then add the point $(6, 16)$ to our data set to get a more-accurate prediction equation.

Coefficient of Determination

The value r^2 is called the **coefficient of determination**

- Tells what percentage of the variability in the response (y) variable is due to the relationship between x and y.
- The rest may be due to such things as lurking variables or confounding.
- Since $-1 \leq r \leq 1 \longrightarrow 0 \leq r^2 \leq 1$

- But what does r^2 *actually* do???
  - Without our regression equation, the best predictor for response variables (y values) would be the average of the y-coordinates, $\overline{y}$.
  - The value of r^2 represents **how much of a decrease in prediction error we get** from using our regression equation rather than $\overline{y}$.
  - If r (and hence r^2) is close to 0, then the linear regression equation is not going to be a good predictor. So you could just use $\overline{y}$.

Example 4. Interpret the value of r^2 from example 1.