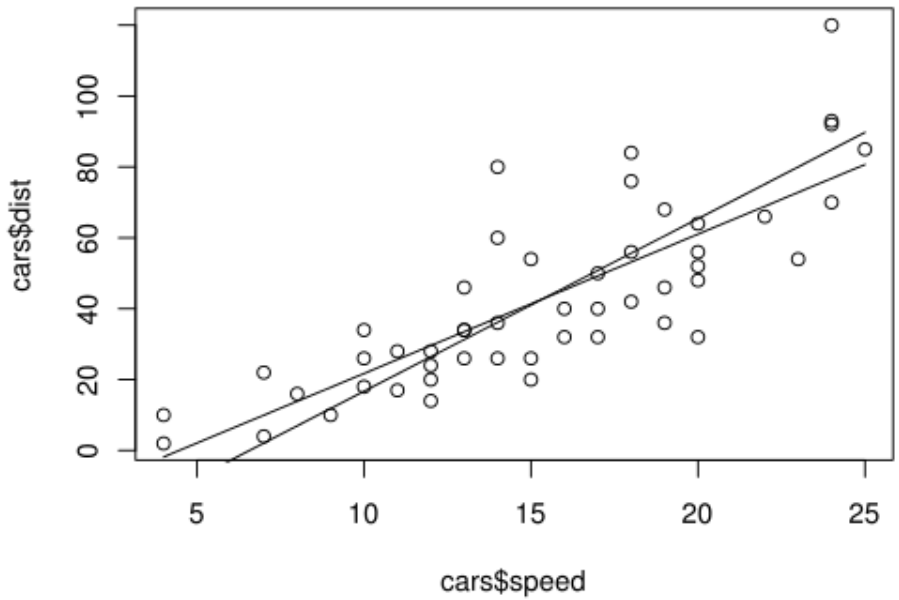
Math 207: Introduction to Statistics

Chapter 12: The Regression Line



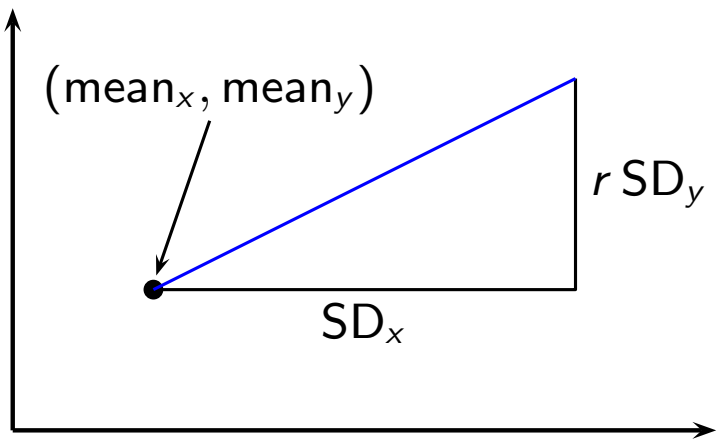
Dr. Ralph Wojtowicz

SHENANDOAH UNIVERSITY

The Regression Line

- For a given value of x,
  - the regression line estimates the average value for y
  - the point on the line is the predicted y for an individual
- For each increase of one SD in x, there is an increase of r SDs in y.

$$(y - \text{mean}_y) = r\,\frac{\text{SD}_y}{\text{SD}_x}\,(x - \text{mean}_x)$$



- The regression line is the line through the data that minimizes the RMS error.
- $\text{RMS}_{\text{reg}} = \text{SD}_y\sqrt{1 - r^2}$.
- In vertical strips, the y values approximately have a normal distribution with center on the line and with SD $= \text{RMS}_{\text{reg}}$.

The Correlation Coefficient

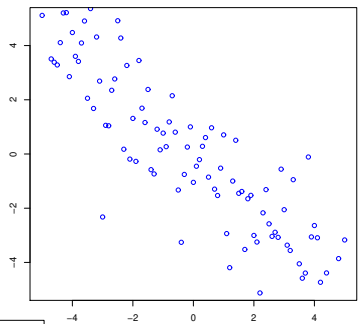
- Given lists $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, the correlation coefficient:
  - Is a measure of linear association between the lists
  - Is a measure of the clustering of the (x_i, y_i) points around a line
  - Is a number between -1 and 1
  - Is defined by:

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \text{mean}_x}{\text{SD}_x} \right) \left(\frac{y_i - \text{mean}_y}{\text{SD}_y} \right)$$

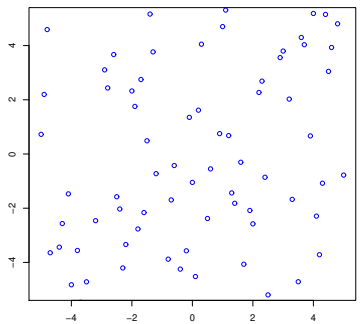
$= \text{average of the } x \text{ and } y \text{ values measured in standard units}$

- A positive correlation means that the cloud of (x_i, y_i) points slopes up
- A negative correlation means that the cloud of (x_i, y_i) points slopes down

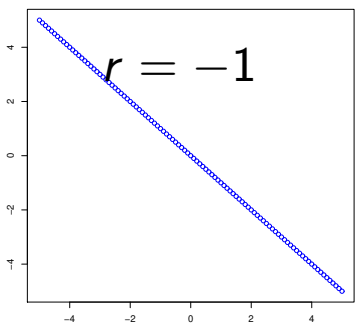
Sign and Magnitude of the Correlation Coefficient



$r = -0.9$

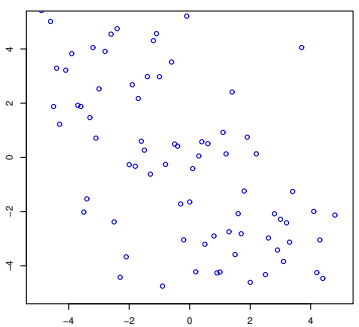


$r = 0.4$

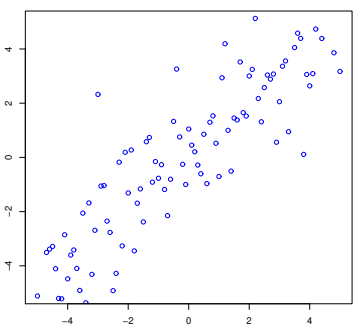


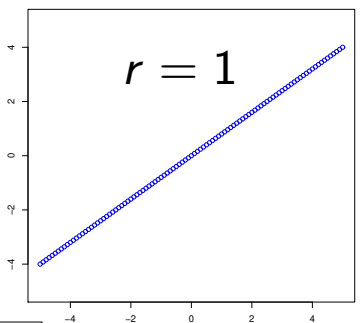
$r = -1$

$r = -0.7$



$r = 0.9$





$r = 1$

Example

See Exercise 9 on page 215.

$$\begin{aligned} \text{average income} &\approx \$90{,}000, & \text{SD} &\approx \$45{,}000, & \\ \text{average IQ} &\approx 100, & \text{SD} &\approx 15, & r \approx 0.50 \end{aligned}$$

Example

See Exercise 9 on page 215.

$$\begin{array}{rcl@{\qquad}rcl@{\qquad}l} \text{average income} & \approx & \$90{,}000, & \text{SD} & \approx & \$45{,}000, & \\ \text{average IQ} & \approx & 100, & \text{SD} & \approx & 15, & r \approx 0.50 \end{array}$$

- Find the regression line for predicting IQ from income.

$$(y - 100) = 0.5 \frac{15}{45,000} (x - 90,000) \quad \text{which is} \quad (y - 100) = 0.000167 (x - 90,000)$$

Example

See Exercise 9 on page 215.

$$\begin{aligned} \text{average income} &\approx \$90{,}000, & \text{SD} &\approx \$45{,}000, & \\ \text{average IQ} &\approx 100, & \text{SD} &\approx 15, & r \approx 0.50 \end{aligned}$$

- Find the regression line for predicting IQ from income.

$$(y - 100) = 0.5 \frac{15}{45,000} (x - 90,000) \quad \text{which is} \quad (y - 100) = 0.000167\,(x - 90,000)$$

- Predict IQ of an individual who makes \$120,000. Put a plus or minus on your estimate.
  $y = 100 + 0.000167\,(120,000 - 90,000) \approx 105$
  The $\pm$ is $\text{RMS}_{\text{reg}} = \text{SD}_y \sqrt{1 - r^2} = 15\sqrt{1 - 0.5^2} \approx 13$.

Example

See Exercise 9 on page 215.

$$\begin{array}{rcl@{\qquad}rcl@{\qquad}l} \text{average income} & \approx & \$90{,}000, & \text{SD} & \approx & \$45{,}000, & \\ \text{average IQ} & \approx & 100, & \text{SD} & \approx & 15, & r \approx 0.50 \end{array}$$

- Find the regression line for predicting IQ from income.

$$(y - 100) = 0.5\,\frac{15}{45,000}\,(x - 90,000) \quad \text{which is} \quad (y - 100) = 0.000167\,(x - 90,000)$$

- Predict IQ of an individual who makes \$120,000. Put a plus or minus on your estimate.
  $y = 100 + 0.000167\,(120,000 - 90,000) \approx 105$
  The $\pm$ is $\text{RMS}_{\text{reg}} = \text{SD}_y\sqrt{1 - r^2} = 15\sqrt{1 - 0.5^2} \approx 13$.
- About 95% of the individuals who made \$120,000 had IQs in what range? $105 \pm 2 \cdot 13$

Example

See Exercise 9 on page 215.

$$\begin{array}{rcl@{\qquad}rcl@{\qquad}l} \text{average income} & \approx & \$90{,}000, & \text{SD} & \approx & \$45{,}000, & \\ \text{average IQ} & \approx & 100, & \text{SD} & \approx & 15, & r \approx 0.50 \end{array}$$

- Find the regression line for predicting IQ from income.

$$(y - 100) = 0.5\,\frac{15}{45,000}\,(x - 90,000) \quad \text{which is} \quad (y - 100) = 0.000167\,(x - 90,000)$$

- Predict IQ of an individual who makes \$120,000. Put a plus or minus on your estimate.
  $y = 100 + 0.000167\,(120,000 - 90,000) \approx 105$
  The $\pm$ is $\text{RMS}_{\text{reg}} = \text{SD}_y\sqrt{1 - r^2} = 15\sqrt{1 - 0.5^2} \approx 13$.
- About 95% of the individuals who made \$120,000 had IQs in what range? $105 \pm 2 \cdot 13$
- Find the regression line for predicting income from IQ.

$$(y - 90,000) = 0.5\,\frac{45,000}{15}\,(x - 100) \quad \text{which is} \quad (y - 90,000) = 1500\,(x - 100)$$

Example

See Exercise 9 on page 215.

$$\begin{array}{rcl@{\qquad}rcl@{\qquad}l} \text{average income} & \approx & \$90{,}000, & \text{SD} & \approx & \$45{,}000, & \\ \text{average IQ} & \approx & 100, & \text{SD} & \approx & 15, & r \approx 0.50 \end{array}$$

- Find the regression line for predicting IQ from income.

$$(y - 100) = 0.5\,\frac{15}{45,000}\,(x - 90,000) \quad \text{which is} \quad (y - 100) = 0.000167\,(x - 90,000)$$

- Predict IQ of an individual who makes \$120,000. Put a plus or minus on your estimate.
  $y = 100 + 0.000167\,(120,000 - 90,000) \approx 105$
  The $\pm$ is $\text{RMS}_{\text{reg}} = \text{SD}_y\sqrt{1 - r^2} = 15\sqrt{1 - 0.5^2} \approx 13$.
- About 95% of the individuals who made \$120,000 had IQs in what range? $105 \pm 2 \cdot 13$
- Find the regression line for predicting income from IQ.

$$(y - 90,000) = 0.5\,\frac{45,000}{15}\,(x - 100) \quad \text{which is} \quad (y - 90,000) = 1500\,(x - 100)$$

- Predict the income of an individual whose IQ is 120. Put a plus or minus on your estimate.
  $y = 90,000 + 1500\,(120 - 100) = 120,000$
  The $\pm$ is $\text{RMS}_{\text{reg}} = \text{SD}_y\sqrt{1 - r^2} = 45,000\sqrt{1 - 0.5^2} \approx \$39,000$.