Correlation: basic properties.

- $-1 \leq r_{xy} \leq 1$ for all sets of paired data. The closer $r_{xy}$ is to $\pm1$, the stronger the linear relationship between the $x$-data and $y$-data. If $r_{xy} = \pm1$ then there is a perfect linear relationship between the two variables: $y_j = mx_j + b$, for all $(x_j, y_j)$.

- $r_{xy}$ does not identify nonlinear relationships. E.g., you can have $r_{xy} \approx 0$, even though $y \approx f(x)$ for some nonlinear function $f$. (The relationships between paired data with significant outliers are also not well-described by the correlation coefficient.)

Conclusion: If $r_{xy}$ is not too close to 0, then we may expect the $y$-values to be somewhat well approximated by a linear function of the corresponding $x$-values (or vice versa). The question is: which one?
Lines and linear functions: a refresher

- A straight line is the graph of a linear equation. These equations are most frequently written as

\[
(i) \ ax + by = c \quad \text{or} \quad (ii) \ y = mx + b.
\]

Version (ii) expresses \( y \) as a (linear) function of \( x \).

- The **slope** of a line is the ratio

\[
\frac{\text{rise}}{\text{run}} = \frac{\text{change in } y}{\text{change in } x} = \frac{\Delta y}{\Delta x}
\]

In equation (ii), the slope is given by \( m \).

- The slope is the amount by which \( y \) is changing for every *unit* change in \( x \). I.e.,

\[
\Delta y = m \cdot \Delta x.
\]
What (straight) line best describes the relationship seen in the data?

First Guess: The **SD-line**. This is the line that passes through the point of averages whose slope is given by $\pm SD_y/SD_x$, ($+SD_y/SD_x$ for positive relationships and $-SD_y/SD_x$ for negative ones).
The SD-line does...

• ...run through the center of the scatterplot, i.e., the points in the scatterplot are spread more-or-less symmetrically around the SD-line;

• ...describe the trend in the data as a whole. I.e., it gives the correct direction for the ‘cloud’ of points in the scatterplot.

The SD-line does not...

• ...do a good job of predicting $y$-values from given $x$-values.

Why not?

• The SD-line doesn’t take the correlation between the variables into account.
What we want:
A *formula* for the approximate $y$-value of an observation with a given $x$-value.

What we can reasonably hope to find:
A formula for the approximate *average* $y$-value for all observations with the same $x$-value.

The SD-line prediction:
For every 1 $SD_x$ increase in $x$, there is a 1 $SD_y$ increase in the average value of $y$... But this is usually *wrong*:
As the observations move away from the *point of averages*, the points on the SD-line tend to lie above or below the average $y$-value that we are trying to estimate, and the further we move from the point of averages, the bigger the errors become.
**Conclusion:** The SD-line goes up (or down) too steeply. I.e., its slope $\pm SD_y/SDx$ is too big (in absolute value). To better predict the average $y$-value for a given $x$-value, we use ...
The regression line.

- The regression line passes through the point of averages (like the SD-line).
- The slope of the regression line (for $y$ on $x$) is given by
  \[ r_{xy} \cdot \frac{SD_y}{SD_x}. \]
- In practical terms, the regression line predicts that for every $SD_x$ change in $x$, there is an approximate $r_{xy} \cdot SD_y$ change in the average value of the corresponding $y$s.

Paired data and the relationship between the two variables ($x$ and $y$) is summarized by the five statistics:

\[ \bar{x}, \quad SD_x, \quad \bar{y}, \quad SD_y \quad \text{and} \quad r_{xy}. \]
Example: Regression of weight on height for women in Great Britain in 1951.

<table>
<thead>
<tr>
<th>Height (inches)</th>
<th>Frequency</th>
<th>Weight (lbs)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>5</td>
<td>15</td>
<td>33</td>
</tr>
<tr>
<td>56</td>
<td>5</td>
<td>15</td>
<td>48</td>
</tr>
<tr>
<td>57</td>
<td>5</td>
<td>15</td>
<td>81</td>
</tr>
<tr>
<td>58</td>
<td>5</td>
<td>15</td>
<td>130</td>
</tr>
<tr>
<td>59</td>
<td>5</td>
<td>15</td>
<td>181</td>
</tr>
<tr>
<td>60</td>
<td>5</td>
<td>15</td>
<td>230</td>
</tr>
<tr>
<td>61</td>
<td>5</td>
<td>15</td>
<td>281</td>
</tr>
<tr>
<td>62</td>
<td>5</td>
<td>15</td>
<td>332</td>
</tr>
<tr>
<td>63</td>
<td>5</td>
<td>15</td>
<td>383</td>
</tr>
<tr>
<td>64</td>
<td>5</td>
<td>15</td>
<td>434</td>
</tr>
<tr>
<td>65</td>
<td>5</td>
<td>15</td>
<td>485</td>
</tr>
<tr>
<td>66</td>
<td>5</td>
<td>15</td>
<td>536</td>
</tr>
<tr>
<td>67</td>
<td>5</td>
<td>15</td>
<td>587</td>
</tr>
<tr>
<td>68</td>
<td>5</td>
<td>15</td>
<td>638</td>
</tr>
<tr>
<td>69</td>
<td>5</td>
<td>15</td>
<td>689</td>
</tr>
<tr>
<td>70</td>
<td>5</td>
<td>15</td>
<td>740</td>
</tr>
<tr>
<td>71</td>
<td>5</td>
<td>15</td>
<td>791</td>
</tr>
<tr>
<td>72</td>
<td>5</td>
<td>15</td>
<td>842</td>
</tr>
</tbody>
</table>

\[
\bar{h} \approx 63 \text{ inches}, \quad s_h \approx 2.7 \text{ inches}, \\
\bar{w} \approx 132 \text{ lbs}, \quad s_w \approx 22.5 \text{ lbs}. \\
r_{hw} \approx 0.32
\]

These numbers can be used to answer questions about average weights for women of different heights...
• How much do 5’6”-tall women weigh on average?

These women are \( \frac{3}{SD_h} = \frac{3}{2.7} \approx 1.11 \) standard deviations above average (height), so, on average they will about \( 0.32 \times 1.11 \approx 0.355 \) standard deviations above the average weight. I.e., the average weight for these women is about

\[
132 + 0.355 \times 22.5 \approx 140 \text{ lbs.}
\]

• How much does average weight go up when height increases by 1 inch?

1 inch represents \( \frac{1}{SD_h} = \frac{1}{2.7} \approx 0.37 \) standard deviations for height, so each additional inch of height adds about \( 0.32 \times 0.37 \approx 0.1184 \) \( SD_w \) to the average weight: this is about 2.66 lbs.
• How much do 5’6”-tall women weigh on average?

These women are $3/SD_h = 3/2.7 \approx 1.11$ standard deviations above average (height), so, on average they will about $0.32 \times 1.11 \approx 0.355$ standard deviations above the average weight. I.e., the average weight for these women is about

$$132 + 0.355 \times 22.5 \approx 140 \text{ lbs.}$$

• How much does average weight go up when height increases by 1 inch?

1 inch represents $1/SD_h = 1/2.7 \approx 0.37$ standard deviations for height, so each additional inch of height adds about $0.32 \times 0.37 \approx 0.1184$ $SD_w$ to the average weight: this is about 2.66 lbs more.

*How accurate are the predictions in this example..?*
• The green line is the regression line.
• The red asterisks are the average weights for each height class.
• The **graph of averages** is the graph that plots *average* \( y \) for each \( x \). In the example above, the graph of averages is represented by the red asterisks.
The regression equation

Since the regression line is a line, we can find its equation:

\[ \hat{y} = \beta_0 + \beta_1 x \]

Comment: \( \hat{y} \) is the predicted or estimated value of the average \( y \)-value for all observations whose \( x \)-value is \( x \).

- \( \beta_1 \) is the slope of the regression line:

\[ \beta_1 = \frac{r_{xy} \cdot SD_y}{SD_x} \]

- To find \( \beta_0 \), we remember that regression line passes through the point of averages, \((\bar{x}, \bar{y})\). Therefore \( \bar{y} = \beta_0 + \beta_1 \cdot \bar{x} \), so that

\[ \beta_0 = \bar{y} - \beta_1 \cdot \bar{x} . \]

\[ \Rightarrow \quad \text{First find } \beta_1, \text{ then use that to find } \beta_0. \]
In the *weight-on-height* example, we have

\[
\beta_1 = \frac{0.32 \cdot 22.5}{2.7} \approx 2.667 \quad \text{and} \quad \beta_0 = 132 - 2.667 \cdot 63 \approx -36,
\]

so the regression equation for weight \((w)\) on height \((h)\) is

\[
\hat{w} = -36 + 2.667 \cdot h.
\]

We can use this equation to answer the same questions as before.

- **How much do 5’6”-tall women weigh on average?**

  The average weight of women who are 66 inches tall is estimated by \(\hat{w}\):
  \[
  \bar{w}(66) \approx \hat{w}(66) = -36 + 2.667 \cdot 66 = 140 \text{ lbs}.
  \]

- **How much does average weight go up when height increases by 1 inch?**

  For every additional inch of height, *average* weight increases by about 2.667 lbs.
Using the regression line for individual predictions

Example: A study of education and income is done for men age 30 - 35. A representative sample of 10000 men in this age group is surveyed, and the following statistics are collected:

\[
\begin{align*}
\bar{E} &= 13 & SD_E &= 1.5 \\
\bar{I} &= 46 & SD_I &= 10 & r_{I,E} &= 0.45
\end{align*}
\]

where \( E \) = years of education and \( I \) = annual income, in $1000s.

- What is the best guess for the income of a man in this age range?

  ⇒ The average income for the whole group: $46,000.
What is the best guess for the income of a man in this age range, given that he has 16 years of education?

⇒ Once again, the best guess for the income of an individual is the average income in the class to which he belongs — i.e., the average income for men (in the given age group) with 16 years of education.

We can use the regression equation to estimate this:

\[ \beta_1 = 0.45 \cdot \frac{10}{1.5} = 3, \quad \beta_0 = 46 - 3 \cdot 13 = 7, \]

so the regression equation is \( \hat{I} = 7 + 3 \cdot E \).

Using this, we estimate the average income for men with 16 years of education to be \( \bar{I}(16) \approx \hat{I}(16) = 7 + 3 \cdot 16 = 55 \), which is our best guess for the income of any individual in the group.
**Question:** How can we make our estimate of the y-value of a subject whose observed x-value is $x_0$ more precise? And more precise in what way?

- In data that is approximately normally distributed, roughly 68% of the data lies within 1 SD of the average value and 95% of the data lies within 2 standard deviations of the average value.

- The average y-value for all observations with x-value $x_0$ is

  $\bar{y}(x_0) \approx \hat{y}(x_0) = \beta_0 + \beta_1 x_0$.

- Applying the normal approximation, we can say that if we observe an x-value of $x_0$, it is reasonable to conclude that the corresponding y-value is likely be in the range

  $\bar{y}(x_0) \pm SD \approx \hat{y}(x_0) \pm SD,$

  or is very likely to be in the range

  $\bar{y}(x_0) \pm 2SD \approx \hat{y}(x_0) \pm 2SD.$
This leaves two important questions:

1. How do we know that the $y$-data for a given $x$-value has an approximately normal distribution?

   and

2. What do we use for the SD?
The R.M.S. error of the regression

If we compare $y_j$ to the predicted value $\hat{y}_j = \beta_0 + \beta_1 x_j$, then we will typically observe an error

$$\varepsilon_j = y_j - \hat{y}_j$$

which may be either positive or negative, depending on whether the prediction was too small or too big. These errors reflect the fact that the data does not lie exactly on any straight line.

The square-root of the average of the squares of these errors,

$$\sqrt{\frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2}$$

is called the R.M.S. error of the regression. This is roughly equal to the average distance of points in the scatterplot to the regression line. The R.M.S. error of regression is also called the standard error of regression, or SER,