Correlation.

How do you study the relationship between two variables?

years of school + income
height of father + height of son
quiz scores + midterm score.

Visualize the relationship with a scatter diagram.

If we know something about the value of one variable, does that tell us something about the value of another variable?

Can we quantify correlation?
A scatter plot is shown with the dependent variable on the y-axis and the mid-term score on the x-axis. The plot suggests a negative correlation, where higher cumulative quiz scores are associated with lower mid-term scores. Some points are marked with positive correlation notes, indicating that people who do better on quizzes tend to do worse on the mid-term.
knowing the value of the independent variable helps a lot in predicting the dependent variable.
weak association
no association (independent variables)
How do we measure the strength of the association between the two variables?

Numerical summary of a cloud of points.

**univariate data.**
- Summarize by mean + SD

![Normal distribution curve](image)

**bivariate data.**
- 2 sets of mean, SD (4 numbers)
- This is not enough.
- Need something else to tell us about the shape of the scatter.

**correlation coefficient** - $r$
- Measures how tightly clustered the points are about the diagonal line.
- $r$ lies between -1 and +1

\[ \begin{array}{c}
\text{If } r > 0, \text{ then } \\
\text{positive correlation} \\
\text{If } r < 0, \text{ then } \\
\text{negative correlation} \\
\text{If } r = 0, \text{ then } \\
\text{no correlation} \\
\end{array} \]

"football shaped"
all have same mean x, mean y, SD x, SD y.
$r = -1$ - knowing one variable, I can predict the other one exactly.

$r = 0.95$

positive correlation.

$r = 1$ - all the points lie on the line.
Computing the correlation coefficient

1. Convert each variable to standard units
2. Compute the mean of the products

\[ r = \text{mean of } (x \text{ in standard units } \times y \text{ in standard units}) \]

Example:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( x ) in standard units</th>
<th>( y ) in standard units</th>
<th>product</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>-1.5</td>
<td>-0.5</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>-0.5</td>
<td>0.5</td>
<td>-0.25</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.5</td>
<td>-1.5</td>
<td>-0.75</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>1.5</td>
<td>1.5</td>
<td>2.25</td>
</tr>
</tbody>
</table>

Mean \( x \) = 4
SD\( x \) = 2

Mean \( y \) = 7
SD\( y \) = 4

to convert to standard units
subtract mean then divide by SD
\[ z = \frac{x - \text{mean}}{\text{SD}} \]
\[ r = \text{average of products} \]

\[ = \frac{0.75 \times (-0.25) + 0 + (-0.75) + 2.25}{5} \]

\[ = 0.4. \]

(not particularly strong correlation)

---

Features of the correlation coefficient

1. It's a pure number.
2. It has no units. (inches, pounds, dollars, etc.)
   because of the conversion to standard units.
3. Rescaling the variables + adding offset
   has no effect on the correlation coefficient.
4. Doesn't apply to non-linear association.
5. Outliers - \( r \) is sensitive to outliers.
   
   a few extreme values can cause
   a large change in value of the
   correlation coefficient
$r = 0.63$

Moderate degree of correlation.
(Any linear transformation)

Going which death change the correlation coefficient.

August Temperatures vs. Elevation in Northern California
R = 0.83

doesn't apply.
non-linear and correlation coefficient.

best fit straight line.
scatter around approach mean. Line.

Correlation coefficient.
Association is not Causation.

Confounding factor - age
Regression.

Regression method describes how one variable depends on another.

E.g. If I go to a place at 4000' elevation, what temperature do I expect to find? (in August)

If I go to many places all at 4000', what range of temperatures do I expect?

<table>
<thead>
<tr>
<th>Elevation</th>
<th>Mean 3524'</th>
<th>SD 1839'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>Mean 70.3°</td>
<td>SD 6.5°</td>
</tr>
</tbody>
</table>

\[ r = -0.76 \] (i.e. slope down)

Estimate average value of temperature given a value for elevation.

- Use the regression line.
regression line for $y$ (temp) on $x$ (altitude) estimates the average value of $y$ corresponding to each value of $x$.

Associated with an increase of 4 SD$_x$ in $x$, there is an increase of $r \times$ SD$_y$ in $y$.

temp vs alt.

as increase in temp change is $r \times$ SD$_{temp}$
alhlude by 1839' (1 SD of alhlude)

$-0.76 \times 6.5$

$= -4.95 \degree F.$

midterm vs quiz $r = 0.64$

1 SD increase in quiz score is associated with

0.64 SD$_m$

increase in midterm score.
August Temperatures vs Elevation in Northern California
How to predict the value of \( x \) at a particular altitude for a particular quiz score:

- Use the average value of \( y \) for the specified value of \( x \) given by the regression method.

  - Work out how far the \( x \) value is from the mean in terms of \( SD_x \)

  - Predict \( y \) by

    \[ \text{mean } y = r \times SD_y \times \text{no. of } SD_x \text{ computed above} \]

    \[ \Rightarrow \text{in standard units.} \]

    (ie \( y \) in standard units in \( r \times x \) in standard units.)

**Example**

SAT scores | mean 550 | SD 80
---|---|---
1st year GPA | 2.6 | SD 0.6

\( r = 0.4 \).

Estimate 1st year GPA of a student with SAT score of 650
1) convert SAT of 650 to standard units

\[
\frac{650 - 550}{80} = 1.25
\]

This student is 1.25 SD\text{SAT} above mean

\(\Rightarrow\) predict that their 1st year GPA will be

\[0.4 \times 1.25 \text{ SD}_{\text{GPA}} \text{ above mean GPA}.
\]

\[= 0.5 \text{ SD}_{\text{GPA}} \text{ above mean GPA}.
\]

\(\Rightarrow\) predicted GPA

\[= 2.6 + \frac{0.5 \times 0.6}{\text{mean GPA}}
\]

\[= 2.9, \text{ an expected value of GPA for a student with 650 SAT}
\]

\[\text{Can only use the regression to make predictions for a population for which the sample is representative.}
\]