Chapter 7: Correlation and Simple Linear Regression

7.1: Correlation
BIVARIATE DATA: CORRELATION AND REGRESSION

Two variables of interest: X, Y.

GOAL: Quantify association between X and Y: correlation.

Predict value of Y from the value of X: regression.

EXAMPLES: (height, weight), (yrs. of education, salary), (hrs. of
studying, exam score), (SAT score, GPA), (chemical reaction time,
temperature), (rainfall, runoff volume), (demand, price), etc.

BIVARIATE DATA: PAIRS (X, Y): (x1, y1), (x2, y2), ..., (xn, yn).
(xi, yi) – i \(^{\text{th}}\) observation, values of X and Y on the i \(^{\text{th}}\) “subject”.

Correlation studies: study type and amount of association between X
and Y.

Regression studies: aim to predict Y from X by constructing a simple
equation relating Y to X.
CORRELATION

GRAPHICAL REPRESENTATION OF BIVARIATE DATA – SCATTER PLOT: plot observations \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) as points on the plane.

Types of association/relationship between vars: positive, negative, none.

Positive association: Two variables are positively associated if large values of one tend to be associated (occur) with large values of the other variable and small values of one tend to be associated with small values of the other variable.

Example: Height and weight are usually positively associated
CORRELATION, contd.

**Negative association**: Two variables are negatively associated if large values of one tend to be associated (occur) with small values of the other. The variables tend to “move in opposite directions”.

**No association**: If there is no association, the points in the scatter plot show no pattern.

E.g. High demand often occurs with low price.
PEARSON CORRELATION COEFFICIENT

Measure of strength of association is the correlation coefficient:
Population: $\rho$ (rho); Sample: $r_{xy}$ or $r$.

Sample correlation coefficient $r$ is a point estimator of the population correlation coefficient $\rho$.

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Sample means: $\bar{x}$ and $\bar{y}$.

Sample standard deviations: $s_x$ and $s_y$.

Sample correlation coefficient:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right).$$

Correlation coefficient measures strength of LINEAR association.
PROPERTIES OF THE SAMPLE CORRELATION COEFFICIENT $r_{xy}$.

- $r_{xy} > 0$ indicates positive association ($\rho_{xy} > 0$) between X and Y.
- $r_{xy} < 0$ indicates negative association ($\rho_{xy} < 0$) between X and Y.
- $r_{xy} \approx 0$ indicates no association ($\rho_{xy} \approx 0$) between X and Y.

- $-1 \leq r_{xy}$ (or $\rho_{xy}) \leq 1$, the closer $|r_{xy}|$ (or $|\rho_{xy}|$) to 1, the stronger the relationship between X and Y.

Computational formula for $r$: we can compute $r$ as

$$r = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}}$$

Best way to compute $r$ is using MINITAB ☺️
CORRELATION COEFFICIENT AND ASSOCIATION

Strong association

Moderate association

No association

Weak association

Almost perfect association, but not linear, $r$ small.

$r = 0.9$

$r = 0.75$

$r = 0.5$

$r = 0.28$

$r = -0.3$
CORRELATION, CONTD.

- Correlation does not imply CAUSATION!

- Watch out for hidden (lurking) variables.

Example. Study of fires. X=amount of damage, Y = # of firefighters. $r_{XY} \approx 0.85$. The more firefighters, the more damage?

Hidden variable: Size of the fire.

Example. There is a positive correlation between shoe size and vocabulary for a person.

Hidden variable: Age.
EXAMPLE

In a study of income and savings, data was collected from 10 households. Both savings and income are reported in thousands of \$ in the following table. Find the correlation coefficient between income and savings.

<table>
<thead>
<tr>
<th>income</th>
<th>savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.5</td>
</tr>
<tr>
<td>28</td>
<td>0.0</td>
</tr>
<tr>
<td>35</td>
<td>0.8</td>
</tr>
<tr>
<td>39</td>
<td>1.6</td>
</tr>
<tr>
<td>44</td>
<td>1.8</td>
</tr>
<tr>
<td>48</td>
<td>3.1</td>
</tr>
<tr>
<td>52</td>
<td>4.3</td>
</tr>
<tr>
<td>65</td>
<td>4.6</td>
</tr>
<tr>
<td>55</td>
<td>3.5</td>
</tr>
<tr>
<td>72</td>
<td>7.2</td>
</tr>
</tbody>
</table>

**Solution:** Summary statistics: \( X=\text{income}, \ Y=\text{savings} \)
\[ \Sigma x_i = 463, \ \Sigma x_i^2 = 23533, \ \Sigma y_i = 27.4, \ \Sigma y_i^2 = 120.04, \ \Sigma x_i y_i = 1564.4. \]

\[
\begin{align*}
    r &= \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}} \\
    &= \frac{10(1564.4) - (463)(27.4)}{\sqrt{10(23533) - (463)^2}\sqrt{10(120.04) - (27.4)^2}} \\
    &= 0.963.
\end{align*}
\]

There is a strong positive association between family income and savings.
Testing hypotheses about significant correlation

Test on significance level $\alpha$.

STEP 1. $H_0: \rho = 0$ vs $H_a: \rho \neq 0$ or $H_a: \rho > 0$ or $H_a: \rho < 0$

STEP 2. Test statistic: 
$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n-2}}}.$$ 

Under the $H_0$, the test statistic has t distr. with n-2 df.

STEP 3. Critical value? One-sided test $t_\alpha$, two-sided $t_{\alpha/2}$.

STEP 4. DECISION-critical/rejection region(s) depends on $H_a$.

- $H_a: \rho \neq 0$ Reject $H_0$ if $|t| > t_{\alpha/2}$;
- $H_a: \rho > 0$ Reject $H_0$ if $t > t_\alpha$;
- $H_a: \rho < 0$ Reject $H_0$ if $t < -t_\alpha$. 
Example

Income and savings. Test if there is a significant positive linear relationship between income and savings. Use significance level of 5%

STEP 1. Ho: $\rho = 0$ vs Ha: $\rho > 0$ (positive relationship)

STEP 2. Test statistic:

$$t = \frac{0.963}{\sqrt{1 - 0.963^2}} = 10.1$$

Under the Ho, the test statistic has t distr. with 8 df.

STEP 3. Critical value? One-sided test $t_{0.05} = 2.306$.

STEP 4. DECISION-critical/rejection region(s) depends on Ha.

Ha: $\rho > 0$ Reject Ho if $t > 2.306$; $t = 10.1 > 2.306$, so we reject Ho.

Answer: There is a significant positive correlation between income and savings.
Example: Eruptions of the Old Faithful Geyser

When Old Faithful geyser erupts in the Yellowstone NP, the length of the eruption time and the time interval until the next eruption are recorded. The data for 8 eruptions are in your book, page 483, Table 10-1. There seems to be a linear relationship between these two variables.

The sample correlation coefficient is $r = 0.926$. Test if this relationship is significantly different from zero. Use significance level of 0.01.

**STEP 1.** Ho: $\rho = 0$ \hspace{1cm} vs \hspace{1cm} Ha: $\rho \neq 0$

**STEP 2:** Test statistic:

$$ t = \frac{0.926}{\sqrt{\frac{1-0.926^2}{8-2}}} = 6.008 $$

**STEP 3.** Critical value? Two-sided test, df=6, $t_{0.005} = 3.707$.

**STEP 4.** $|t|=6.008 > 3.707$, so reject Ho. Thus, there is a significant linear relationship between the length of the eruption and the time after the eruption.
Correlation in MINITAB

Welcome to Minitab, press F1 for help.

Correlations: income, savings

Pearson correlation of income and savings = 0.963
P-Value = 0.000