Correlation

**Goal:** Find cause and effect links between variables.
Correlation

**Goal:** Find cause and effect links between variables.

What can we conclude when two variables are highly correlated?

<table>
<thead>
<tr>
<th>Positive Correlation</th>
<th>Negative Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>High values of $x$ are associated with high values of $y$.</td>
<td>High values of $x$ are associated with low values of $y$.</td>
</tr>
</tbody>
</table>
Correlation

**Goal:** Find cause and effect links between variables.

What can we conclude when two variables are highly **correlated**?

- **Positive Correlation**
  - High values of $x$ are associated with high values of $y$.

- **Negative Correlation**
  - High values of $x$ are associated with low values of $y$.

The **correlation coefficient**, $R^2$ is a number between 0 and 1.
Correlation

**Goal:** Find cause and effect links between variables.

What can we conclude when two variables are highly **correlated**?

<table>
<thead>
<tr>
<th>Positive Correlation</th>
<th>Negative Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>High values of ( x ) are associated with high values of ( y ).</td>
<td>High values of ( x ) are associated with low values of ( y ).</td>
</tr>
</tbody>
</table>

The **correlation coefficient**, \( R^2 \) is a number between 0 and 1. Values near 1 show **strong correlation** (data lies almost on a line). Values near 0 show **weak correlation** (data doesn’t lie on a line).
Calculating the $R^2$ Statistic

To find $R^2$, you need data and its best fit linear regression.
Calculating the $R^2$ Statistic

To find $R^2$, you need data and its best fit linear regression. Calculate:

- **The error sum of squares:** $SSE = \sum_i [y_i - f(x_i)]^2$.

- **The total corrected sum of squares:** $SST = \sum_i [y_i - \bar{y}]^2$, where $\bar{y}$ is the average $y_i$ value.
Calculating $R^2$ Statistic

To find $R^2$, you need data and its best fit linear regression. Calculate:

- The **error sum of squares**: $SSE = \sum_i [y_i - f(x_i)]^2$.

  ★ $SSE$ is the variation between the data and the function. ★

  ★ Note: this as what “least squares” minimizes. ★

- The **total corrected sum of squares**: $SST = \sum_i [y_i - \bar{y}]^2$, where $\bar{y}$ is the average $y_i$ value.
Calculating the $R^2$ Statistic

To find $R^2$, you need data and its best fit linear regression. Calculate:

- The error sum of squares: $SSE = \sum_i [y_i - f(x_i)]^2$.
  
  $SSE$ is the variation between the data and the function. ★
  
  Note: this as what “least squares” minimizes. ★

- The total corrected sum of squares: $SST = \sum_i [y_i - \bar{y}]^2$,
  where $\bar{y}$ is the average $y_i$ value.
  
  $SST$ is the variation solely due to the data. ★
Calculating the $R^2$ Statistic

To find $R^2$, you need data and its best fit linear regression. Calculate:

- The **error sum of squares**: $SSE = \sum_i [y_i - f(x_i)]^2$.
  
  **SSE** is the variation between the data and the function. ★

  **Note:** this as what “least squares” minimizes. ★

- The **total corrected sum of squares**: $SST = \sum_i [y_i - \bar{y}]^2$, where $\bar{y}$ is the average $y_i$ value.
  
  **SST** is the variation solely due to the data. ★

- Now calculate $R^2 = 1 - \frac{SSE}{SST}$. 
Calculating the $R^2$ Statistic

To find $R^2$, you need data and its best fit linear regression. Calculate:

- The **error sum of squares**: $SSE = \sum_i [y_i - f(x_i)]^2$.
  - $SSE$ is the variation between the data and the function. ★
  - Note: this as what “least squares” minimizes. ★

- The **total corrected sum of squares**: $SST = \sum_i [y_i - \bar{y}]^2$,
  - where $\bar{y}$ is the average $y_i$ value.
  - $SST$ is the variation solely due to the data. ★

- Now calculate $R^2 = 1 - \frac{SSE}{SST}$.
  - $R^2$ is the proportion of variation explained by the function. ★
Calculating the $R^2$ Statistic

To find $R^2$, you need data and its best fit *linear* regression. Calculate:

- The **error sum of squares**: $SSE = \sum_i [y_i - f(x_i)]^2$.
  
  $SSE$ is the variation between the data and the function. ★

  Note: this is what “least squares” minimizes. ★

- The **total corrected sum of squares**: $SST = \sum_i [y_i - \bar{y}]^2$, where $\bar{y}$ is the average $y_i$ value.
  
  $SST$ is the variation solely due to the data. ★

- Now calculate $R^2 = 1 - \frac{SSE}{SST}$.
  
  $R^2$ is the proportion of variation explained by the function. ★

**Is my $R^2$ good?** Use a critical value table for $R$. (Note: not $R^2$.)

http://www.gifted.uconn.edu/siegle/research/correlation/corrchrt.htm
Calculating the $R^2$ Statistic

Example. (cont’d from notes p. 33) What is $R^2$ for the data set: 
{\((1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\)}?
Calculating the $R^2$ Statistic

Example. (cont’d from notes p. 33) What is $R^2$ for the data set: 
$\{(1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\}$?

You first need the regression line: $f(x) = -0.605027x + 4.20332$. 
Calculating the $R^2$ Statistic

Example. (cont’d from notes p. 33) What is $R^2$ for the data set: 
\{(1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\}?

You first need the regression line: \( f(x) = -0.605027x + 4.20332 \).

\[ \text{The error sum of squares: } \text{SSE} = \sum \left[ y_i - f(x_i) \right]^2. \]

\[ \text{SSE} = (3.6 - f(1.0))^2 + (2.9 - f(2.1))^2 + (2.2 - f(3.5))^2 + (1.7 - f(4.0))^2 \]
Calculating the $R^2$ Statistic

Example. (cont’d from notes p. 33) What is $R^2$ for the data set: 
\{(1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\}?

You first need the regression line: $f(x) = -0.605027x + 4.20332$.

- The error sum of squares: $SSE = \sum_i [y_i - f(x_i)]^2$.

\[
SSE = (3.6 - f(1.0))^2 + (2.9 - f(2.1))^2 + (2.2 - f(3.5))^2 + (1.7 - f(4.0))^2
\]
\[
= (.0017)^2 + (-0.033)^2 + (0.114)^2 + (-0.083)^2 = 0.0210
\]
Example. (cont’d from notes p. 33) What is $R^2$ for the data set: 
\{(1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\}?

You first need the regression line: $f(x) = -0.605027x + 4.20332$.

- **The error sum of squares**: $SSE = \sum (y_i - f(x_i))^2$.

\[
SSE = (3.6 - f(1.0))^2 + (2.9 - f(2.1))^2 + (2.2 - f(3.5))^2 + (1.7 - f(4.0))^2
= (.0017)^2 + (-0.033)^2 + (0.114)^2 + (-0.083)^2 = 0.0210
\]

- **The total corrected sum of squares**: $SST = \sum (y_i - \bar{y})^2$.

**First**, calculate $\bar{y} = (3.6 + 2.9 + 2.2 + 1.7)/4 = 2.6$
Calculating the $R^2$ Statistic

Example. (cont’d from notes p. 33) What is $R^2$ for the data set: 
\{(1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\}?

You first need the regression line: $f(x) = -0.605027x + 4.20332$.

▷ The error sum of squares: $SSE = \sum_i [y_i - f(x_i)]^2$.

$$SSE = (3.6 - f(1.0))^2 + (2.9 - f(2.1))^2 + (2.2 - f(3.5))^2 + (1.7 - f(4.0))^2$$
$$= (.0017)^2 + (-0.033)^2 + (0.114)^2 + (-0.083)^2 = 0.0210$$

▷ The total corrected sum of squares: $SST = \sum_i [y_i - \bar{y}]^2$.

First, calculate $\bar{y} = (3.6 + 2.9 + 2.2 + 1.7)/4 = 2.6$

$$SST = (3.6 - 2.6)^2 + (2.9 - 2.6)^2 + (2.2 - 2.6)^2 + (1.7 - 2.6)^2$$
$$= (1)^2 + (0.3)^2 + (-0.4)^2 + (-0.9)^2 = 2.06$$
Calculating the $R^2$ Statistic

Example. (cont’d from notes p. 33)  What is $R^2$ for the data set: \{(1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\}?

You first need the regression line: $f(x) = -0.605027x + 4.20332$.

$\textbf{The error sum of squares:} \quad SSE = \sum_i [y_i - f(x_i)]^2$.

$SSE = (3.6 - f(1.0))^2 + (2.9 - f(2.1))^2 + (2.2 - f(3.5))^2 + (1.7 - f(4.0))^2$

$= (.0017)^2 + (-0.033)^2 + (0.114)^2 + (-0.083)^2 = 0.0210$

$\textbf{The total corrected sum of squares:} \quad SST = \sum_i [y_i - \bar{y}]^2$.

First, calculate $\bar{y} = (3.6 + 2.9 + 2.2 + 1.7)/4 = 2.6$

$SST = (3.6 - 2.6)^2 + (2.9 - 2.6)^2 + (2.2 - 2.6)^2 + (1.7 - 2.6)^2$

$= (1)^2 + (0.3)^2 + (-0.4)^2 + (-0.9)^2 = 2.06$

$\textbf{Now calculate} \quad R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0.0210}{2.06} = 1 - .01 = 0.99.$
Another $R^2$ Calculation

Another $R^2$ Calculation

**Example.** Estimating weight from height.

Here is a list of heights and weights for ten students.

<table>
<thead>
<tr>
<th>ht.</th>
<th>wt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>160</td>
</tr>
<tr>
<td>70</td>
<td>160</td>
</tr>
<tr>
<td>71</td>
<td>150</td>
</tr>
<tr>
<td>68</td>
<td>120</td>
</tr>
<tr>
<td>68</td>
<td>175</td>
</tr>
<tr>
<td>76</td>
<td>190</td>
</tr>
<tr>
<td>73.5</td>
<td>205</td>
</tr>
<tr>
<td>75.5</td>
<td>215</td>
</tr>
<tr>
<td>73</td>
<td>185</td>
</tr>
<tr>
<td>72</td>
<td>170</td>
</tr>
</tbody>
</table>
Another $R^2$ Calculation

**Example.** Estimating weight from height.

Here is a list of heights and weights for ten students.

We calculate the line of best fit:

$$\text{(weight)} = 7.07(\text{height}) - 333.$$
Another $R^2$ Calculation


Here is a list of heights and weights for ten students.

We calculate the line of best fit:

$$\text{(weight)} = 7.07\text{(height)} - 333.$$
Another $R^2$ Calculation

**Example.** Estimating weight from height.

Here is a list of heights and weights for ten students.

We calculate the line of best fit:

\[
(\text{weight}) = 7.07(\text{height}) - 333.
\]

Now find the correlation coefficient: ($\bar{w} = 173$)

<table>
<thead>
<tr>
<th>ht.</th>
<th>wt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>160</td>
</tr>
<tr>
<td>70</td>
<td>160</td>
</tr>
<tr>
<td>71</td>
<td>150</td>
</tr>
<tr>
<td>68</td>
<td>120</td>
</tr>
<tr>
<td>68</td>
<td>175</td>
</tr>
<tr>
<td>76</td>
<td>190</td>
</tr>
<tr>
<td>73.5</td>
<td>205</td>
</tr>
<tr>
<td>75.5</td>
<td>215</td>
</tr>
<tr>
<td>73</td>
<td>185</td>
</tr>
<tr>
<td>72</td>
<td>170</td>
</tr>
</tbody>
</table>
Another $R^2$ Calculation

**Example.** Estimating weight from height.

Here is a list of heights and weights for ten students.

We calculate the line of best fit:

$\text{(weight)} = 7.07 \times \text{(height)} - 333$.

Now find the correlation coefficient: ($\bar{w} = 173$)

$SSE = \sum_{i=1}^{10} [w_i - (7.07 \times h_i - 333)]^2 \approx 2808$

$SST = \sum_{i=1}^{10} [w_i - 173]^2 = 6910$

So $R^2 = 1 - (2808/6910) = 0.59$
Another $R^2$ Calculation


Here is a list of heights and weights for ten students.

We calculate the line of best fit:

\[
\text{(weight)} = 7.07(\text{height}) - 333.
\]

Now find the correlation coefficient: \( \bar{w} = 173 \)

\[
SSE = \sum_{i=1}^{10} [w_i - (7.07 h_i - 333)]^2 \approx 2808
\]

\[
SST = \sum_{i=1}^{10} [w_i - 173]^2 = 6910
\]

So $R^2 = 1 - (2808/6910) = 0.59$, a good correlation.
Another $R^2$ Calculation


Here is a list of heights and weights for ten students.

We calculate the line of best fit:

$\text{(weight)} = 7.07(\text{height}) - 333.$

Now find the correlation coefficient: ($\overline{w} = 173$)

$SSE = \sum_{i=1}^{10} [w_i - (7.07 h_i - 333)]^2 \approx 2808$

$SST = \sum_{i=1}^{10} [w_i - 173]^2 = 6910$

So $R^2 = 1 - (2808/6910) = 0.59$, a good correlation.

We can introduce another variable to see if the fit improves.
## Multiple Linear Regression

Add waist measurements to the data!

<table>
<thead>
<tr>
<th>ht.</th>
<th>wst.</th>
<th>wt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>34</td>
<td>160</td>
</tr>
<tr>
<td>70</td>
<td>32</td>
<td>160</td>
</tr>
<tr>
<td>71</td>
<td>31</td>
<td>150</td>
</tr>
<tr>
<td>68</td>
<td>29</td>
<td>120</td>
</tr>
<tr>
<td>68</td>
<td>34</td>
<td>175</td>
</tr>
<tr>
<td>76</td>
<td>34</td>
<td>190</td>
</tr>
<tr>
<td>73.5</td>
<td>38</td>
<td>205</td>
</tr>
<tr>
<td>75.5</td>
<td>34</td>
<td>215</td>
</tr>
<tr>
<td>73</td>
<td>36</td>
<td>185</td>
</tr>
<tr>
<td>72</td>
<td>32</td>
<td>170</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

Add waist measurements to the data!

We wish to calculate a linear relationship such as:

\[(\text{weight}) = a (\text{height}) + b (\text{waist}) + c.\]

<table>
<thead>
<tr>
<th>ht</th>
<th>wst</th>
<th>wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>34</td>
<td>160</td>
</tr>
<tr>
<td>70</td>
<td>32</td>
<td>160</td>
</tr>
<tr>
<td>71</td>
<td>31</td>
<td>150</td>
</tr>
<tr>
<td>68</td>
<td>29</td>
<td>120</td>
</tr>
<tr>
<td>68</td>
<td>34</td>
<td>175</td>
</tr>
<tr>
<td>76</td>
<td>34</td>
<td>190</td>
</tr>
<tr>
<td>73.5</td>
<td>38</td>
<td>205</td>
</tr>
<tr>
<td>75.5</td>
<td>34</td>
<td>215</td>
</tr>
<tr>
<td>73</td>
<td>36</td>
<td>185</td>
</tr>
<tr>
<td>72</td>
<td>32</td>
<td>170</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

Add waist measurements to the data!

We wish to calculate a *linear* relationship such as:

\[(\text{weight}) = a(\text{height}) + b(\text{waist}) + c.\]

Do a regression to find the *best-fit plane*:

Use the least-squares criterion. Minimize:

\[\text{SSE} = \sum_{(h_i, ws_i, wt_i)} [wt_i - (a \cdot h_i + b \cdot ws_i + c)]^2.\]

<table>
<thead>
<tr>
<th>ht.</th>
<th>wst.</th>
<th>wt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>34</td>
<td>160</td>
</tr>
<tr>
<td>70</td>
<td>32</td>
<td>160</td>
</tr>
<tr>
<td>71</td>
<td>31</td>
<td>150</td>
</tr>
<tr>
<td>68</td>
<td>29</td>
<td>120</td>
</tr>
<tr>
<td>68</td>
<td>34</td>
<td>175</td>
</tr>
<tr>
<td>76</td>
<td>34</td>
<td>190</td>
</tr>
<tr>
<td>73.5</td>
<td>38</td>
<td>205</td>
</tr>
<tr>
<td>75.5</td>
<td>34</td>
<td>215</td>
</tr>
<tr>
<td>73</td>
<td>36</td>
<td>185</td>
</tr>
<tr>
<td>72</td>
<td>32</td>
<td>170</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

Add waist measurements to the data!

We wish to calculate a *linear* relationship such as:

\[
(\text{weight}) = a \cdot (\text{height}) + b \cdot (\text{waist}) + c.
\]

Do a regression to find the *best-fit plane*:

Use the least-squares criterion. Minimize:

\[
SSE = \sum_{(h_i, ws_i, wt_i)} \left[ wt_i - (a \cdot h_i + b \cdot ws_i + c) \right]^2.
\]

This finds that the best fit plane is *(coeff sign)*

\[
(\text{weight}) = 4.59(\text{height}) + 6.35(\text{waist}) - 368.
\]
Visually, we might expect a plane to do a better job fitting the points than the line.
Multiple Linear Regression

Visually, we might expect a plane to do a better job fitting the points than the line.

► Now calculate $R^2$.

Calculate $SSE = \sum_{i=1}^{10} (w_i - f(h_i, ws_i))^2 \approx 955$

$SST$ does not change: (why?)

$\sum_{i=1}^{10} (w_i - 173)^2 = 6910$

<table>
<thead>
<tr>
<th>ht.</th>
<th>wst.</th>
<th>wt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>34</td>
<td>160</td>
</tr>
<tr>
<td>70</td>
<td>32</td>
<td>160</td>
</tr>
<tr>
<td>71</td>
<td>31</td>
<td>150</td>
</tr>
<tr>
<td>68</td>
<td>29</td>
<td>120</td>
</tr>
<tr>
<td>68</td>
<td>34</td>
<td>175</td>
</tr>
<tr>
<td>76</td>
<td>34</td>
<td>190</td>
</tr>
<tr>
<td>73.5</td>
<td>38</td>
<td>205</td>
</tr>
<tr>
<td>75.5</td>
<td>34</td>
<td>215</td>
</tr>
<tr>
<td>73</td>
<td>36</td>
<td>185</td>
</tr>
<tr>
<td>72</td>
<td>32</td>
<td>170</td>
</tr>
</tbody>
</table>
Visually, we might expect a plane to do a better job fitting the points than the line.

Now calculate $R^2$.

Calculate $SSE = \sum_{i=1}^{10} (w_i - f(h_i, ws_i))^2 \approx 955$

$SST$ does not change: (why?)

$\sum_{i=1}^{10} (w_i - 173)^2 = 6910$

So $R^2 = 1 - (955/6910) = 0.86$, an excellent correlation.
Multiple Linear Regression

Visually, we might expect a plane to do a better job fitting the points than the line.

► Now calculate $R^2$.

Calculate $SSE = \sum_{i=1}^{10} (w_i - f(h_i, ws_i))^2 \approx 955$

$SST$ does not change: (why?)

$\sum_{i=1}^{10} (w_i - 173)^2 = 6910$

So $R^2 = 1 - (955/6910) = 0.86$, an excellent correlation.

► When you introduce more variables, $SSE$ can only go down, so $R^2$ always increases.
Notes about the Correlation Coefficient

Example. Time and Distance (pp. 190)
Data collected to predict driving time from home to school.
Notes about the Correlation Coefficient

Example. Time and Distance (pp. 190)
Data collected to predict driving time from home to school.

Variables:
\( T = \text{driving time} \)
\( M = \text{miles driven} \)

Use a linear regression to find that
\( T = 1.89M + 8.05, \) with an \( R^2 = 0.867. \)
Notes about the Correlation Coefficient

Example. Time and Distance (pp. 190)
Data collected to predict driving time from home to school.

Variables:
\( T = \text{driving time} \quad S = \text{Last two digits of SSN} \)
\( M = \text{miles driven} \)

Use a linear regression to find that
\( T = 1.89M + 8.05 \), with an \( R^2 = 0.867 \).
Notes about the Correlation Coefficient

Example. Time and Distance (pp. 190)
Data collected to predict driving time from home to school.

Variables:
\[ T = \text{driving time} \quad S = \text{Last two digits of SSN.} \]
\[ M = \text{miles driven} \]

Use a linear regression to find that
\[ T = 1.89M + 8.05, \text{ with an } R^2 = 0.867. \]

Compare to a multiple linear regression of
\[ T = 1.7M + 0.0872S + 13.2, \text{ with an } R^2 = 0.883! \]
Notes about the Correlation Coefficient

Example. Time and Distance (pp. 190)
Data collected to predict driving time from home to school.

Variables:
\[ T = \text{driving time} \quad S = \text{Last two digits of SSN.} \]
\[ M = \text{miles driven} \]

Use a linear regression to find that
\[ T = 1.89M + 8.05, \text{ with an } R^2 = 0.867. \]

Compare to a multiple linear regression of
\[ T = 1.7M + 0.0872S + 13.2, \text{ with an } R^2 = 0.883! \]

\[ R^2 \text{ increases as the number of variables increase.} \]
\[ \text{This doesn’t mean that the fit is better!} \]
Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)
Does fluoride in the water cause cancer?
Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)
Does fluoride in the water cause cancer?

Variables:
\[ T = \log \text{ of years of fluoridation} \]
\[ C = \text{cancer mortality rate} \]
Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)
Does fluoride in the water cause cancer?

Variables:
\( T = \log \) of years of fluoridation
\( C = \) cancer mortality rate

Use a linear regression to find that
\[ C = 27.1T + 181, \text{ with an } R^2 = 0.047. \]
Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)
Does fluoride in the water cause cancer?

Variables:
$T = \log$ of years of fluoridation $A = \%$ of population over 65.
$C = \text{cancer mortality rate}$

Use a linear regression to find that
$C = 27.1T + 181$, with an $R^2 = 0.047$.  


Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)
Does fluoride in the water cause cancer?

Variables:
\( T = \log \) of years of fluoridation \( A = \% \) of population over 65.
\( C = \) cancer mortality rate

Use a linear regression to find that
\[ C = 27.1 T + 181, \text{ with an } R^2 = 0.047. \]

Compare to a multiple linear regression of
\[ C = 0.566 T + 10.6A + 85.8, \text{ with an } R^2 = 0.493. \]

► Be suspicious of a low \( R^2 \).
Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)
Does fluoride in the water cause cancer?

Variables:
\[ T = \text{log of years of fluoridation} \quad A = \% \text{ of population over 65.} \]
\[ C = \text{cancer mortality rate} \]

Use a linear regression to find that
\[ C = 27.1 T + 181, \text{ with an } R^2 = 0.047. \]

Compare to a multiple linear regression of
\[ C = 0.566 T + 10.6A + 85.8, \text{ with an } R^2 = 0.493. \]

- Be suspicious of a low \( R^2 \).
- Signs of coefficients tell positive/negative correlation.
Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)

Does fluoride in the water cause cancer?

Variables:

- $T = \log$ of years of fluoridation
- $A = \%$ of population over 65.
- $C =$ cancer mortality rate

Use a linear regression to find that

$$C = 27.1T + 181,$$

with an $R^2 = 0.047$.

Compare to a multiple linear regression of

$$C = 0.566T + 10.6A + 85.8,$$

with an $R^2 = 0.493$.

- Be suspicious of a low $R^2$.
- Signs of coefficients tell positive/negative correlation.
- Cannot determine relative influence of one variable in one model without some gauge on the magnitude of the data.
Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)

Does fluoride in the water cause cancer?

Variables:
\[ T = \log \text{of years of fluoridation} \quad A = \% \text{ of population over 65.} \]
\[ C = \text{cancer mortality rate} \]

Use a linear regression to find that
\[ C = 27.1T + 181, \text{ with an } R^2 = 0.047. \]

Compare to a multiple linear regression of
\[ C = 0.566T + 10.6A + 85.8, \text{ with an } R^2 = 0.493. \]

- Be suspicious of a low \( R^2 \).
- Signs of coefficients tell positive/negative correlation.
- Cannot determine relative influence of one variable in one model without some gauge on the magnitude of the data.
- **CAN** determine relative influence of one variable in two models.