Scatter Plots and Regression

An example of perfect positive linear correlation.
Sample correlation coefficient: \( r = 1.0 \)
Equation of least-squares regression line:
\[
C = 0.55555F - 17.77777
\]
or
\[
C = \frac{5}{9}F - \frac{160}{9}
\]
A slope of \( \frac{5}{9} \) tells us that when the F temp increases 9°, the C temp increases 5° or C increases \((\frac{5}{9})^\circ\) when F increases 1°.

An example of perfect negative linear correlation.
Sample correlation coefficient: \( r = -1.0 \)
Equation of least-squares regression line:
\[
w = \frac{3}{2}n + 280
\]
or
\[
w = -1.5n + 280
\]
A slope of \( \frac{-3}{2} \) tells us that the weight loss rate is -3 pounds every two weeks or -1.5 pounds per week.
0.1262 is approximately 1/8. This tells us that on average, the student's GPA tends to increase one point for every additional 8 hours of study.

Sample correlation coefficient: $r = 0.92183$

Equation of least-squares regression line:

$$G = 0.126258821N + 0.96410307$$

An example of negative linear correlation.

Sample correlation coefficient: $r = -0.9215$

Equation of least-squares regression line:

$$S = -0.686792453Y + 11.011320755$$

An example of negative linear correlation.

Sample correlation coefficient: $r = -0.84281$

Equation of least-squares regression line:

$$D = -22.968767408N + 260.5634$$

A slope of -23 tells us that on average, the number of deaths from heart disease per 100,000 tends to drop 23 when the average consumption of red wine increases by one liter per year per person.
6. An example of two variable data with a correlation coefficient near zero. The x and y coordinates of each data point is a randomly selected real number from -9 to 9.

The sample correlation coefficient = 0.06805

The equation of the least-squares regression line is \( y = 0.0662x - 2.0614 \). In this case, it is a waste of time to find the equation of the regression line since \( r \) is here zero and therefore the regression equation has no predictive value.

7. An example of two variable data with no linear correlation, but near perfect quadratic correlation.

The equation of the least-squares regression line is \( y = -0.0295x - 3.7099 \) with a sample correlation coefficient of \( r = -.03377 \).

The equation of the quadratic least-squares regression equation is \( y = 0.194943x^2 - 0.029502x - 8.3884 \) with a sample correlation coefficient of \( r = 0.9732 \).

8. An example of two variable data with small linear correlation, but near perfect sinusoidal correlation.

The equation of the least-squares regression line is \( y = -0.322061x + 3.1872 \) with a sample correlation coefficient of \( r = -.20124 \).

The equation of the sinusoidal least-squares regression equation is \( y = 4.78562\sin(1.50383x - 0.022224) + 1.6097 \) with a sample correlation coefficient of \( r = 0.94711 \).
Properties of the Linear Correlation Coefficient \( r \)

1. For **linear** regression models, the correlation coefficient ranges from -1.0 to 1.0. A value of -1.0 or 1.0 indicates perfect correlation and a value near zero indicates little or no correlation between the variables.

2. For **nonlinear** regression models, the correlation coefficient ranges from 0.0 to 1.0. A value of 1.0 indicates perfect correlation and a value near zero indicates little or no correlation.

3. The conventional dictum that "correlation does not imply causation" means that correlation cannot be used to infer a causal relationship between variables. This dictum should not be taken to mean that high correlation cannot indicate a causal relationship between variables. However, the causes underlying the correlation, if any, may be indirect or unknown. Establishing a high correlation between two variables is does not automatically give you a license to claim a causal relationship between the variables (in either direction). A high correlation between two variables indicates a high degree of predictability in the sense that knowing the value of one of the variables allows you predict the value of the other variable with a fairly high degree of accuracy. When high correlation between variables is observed, do not rush to judgment and claim that one variable causes the other variable to change. Always consider other factors that might be contributing to the apparent relationship between variables.

A simple example: Hot weather may cause both a reduction in purchases of warm clothing and an increase in ice-cream purchases. Therefore warm clothing purchases are correlated with ice-cream purchases. A reduction in warm clothing purchases does not cause an increase in ice-cream purchases and increased ice-cream purchases do not cause a reduction in warm clothing purchases. Temperature is a lurking variable because it is the cause of change in ice cream sales and warm clothing purchases.

4. The significance of the correlation coefficient \( r \) can be determined from Table II, Appendix A of your textbook. If the absolute value of the correlation coefficient \( r \) is greater than the critical value corresponding to the sample size \( n \), we can say that there is a statistically significant linear relationship between the explanatory variable and the response variable. If \( |r| \) is less than the critical value, there is no statistically significant linear relationship between the variables.

5. The value of \( r \) does not change when the values of all variables are converted to another scale.

   **Example:** Suppose the value of \( x \) variable is in minutes and the value of the \( y \) variable is in inches. If all \( x \)-values are converted to seconds and all \( y \)-values are converted to cm, the value of \( r \) will not change, but the equation of the least-squares regression line will change.

6. The value of \( r \) **does not change** when the \( x \)-values and \( y \)-values are interchanged. However, the equation of the least-squares regression line **does change** when the \( x \)-values are interchanged with the \( y \)-values.

7. The **coefficient of determination** equals \( r^2 \) where \( r \) = the sample correlation coefficient. \( r^2 \) tells us the percent of the variation in the response variable (\( y \)-variable) that can be explained or predicted by using the equation of the regression line.

   Examples:

   If \( r = 1 \) or -1, then \( r^2 = 1 \) and therefore 100% of any change in the response variable can be explained or predicted by a change in the explanatory variable.

   If \( r \) is near 0, then \( r^2 \) is near 0 and very little or none of the change in the response variable can be explained or predicted by a change in the explanatory variable.

   If \( r = -0.805 \), then \( r^2 = 0.648025 \) and therefore about 65% of the change in the response variable can be predicted by a change in the explanatory variable. The remaining 35% of the variation in the response variable is unexplained and is due to chance or other unknown factors.
Example 1: Lab results from a chemistry experiment: \( x = \text{temp C}^0 \) and \( y = \text{amount of CuSO}_4 \) in grams.

\[
\begin{array}{c|c|c|c|c|c|c|c|c}
 x & 10 & 20 & 30 & 40 & 50 & 60 & 70 \\
 y & 17 & 21 & 25 & 28 & 33 & 40 & 49 \\
\hline
 10 & 17 & 170 & 100 & 289 \\
 20 & 21 & 420 & 400 & 441 \\
 30 & 25 & 750 & 900 & 625 \\
 40 & 28 & 1,120 & 1,600 & 784 \\
 50 & 33 & 1,650 & 2,500 & 1,089 \\
 60 & 40 & 2,400 & 3,600 & 1,600 \\
 70 & 49 & 3,430 & 4,900 & 2,401 \\
\hline
\end{array}
\]

\[
\begin{align*}
280 & \quad 213 & \quad 9,940 & \quad 14,000 & \quad 7,229 \\
\end{align*}
\]

\[n = 7, \quad \bar{x} = \frac{280}{7} = 40 \quad \text{and} \quad \bar{y} = \frac{213}{7} = 30.43\]

\[b = \frac{7 \cdot 9,940 - 280 \cdot 213}{7 \cdot 14,000 - 280^2} = 0.507\]

\[a = 30.43 - 0.507 \cdot 40 = 10.15\]

Equation of regression line: \( y = 0.507x + 10.15\)

\[
r = \frac{7 \cdot 9,940 - 280 \cdot 213}{\sqrt{7 \cdot 14,000 - 280^2} \sqrt{7 \cdot 7,229 - 213^2}} = 0.981
\]

a) Plot the x-y data pairs on the axes provided. Label the axes appropriately.

b) Find the equation of the least-squares regression line. ___________________________________________________________________

c) Predict the amount CuSO\(_4\) when C = 25\(^0\). _______________ , when C = 65\(^0\) _______________

d) Graph the least-squares regression line. Plot and label the point \((\bar{x}, \bar{y})\) which is always on the regression line.

e) Find the slope of the regression line and explain what it tells us. ___________________________________________________________________

_____________________________________________________________________________________

f) Find the correlation coefficient \(r\) of the data set and explain what it tells us. ___________________________________________________________________

_____________________________________________________________________________________

g) Find the coefficient of determination \(r^2\) and explain what it tells us. ___________________________________________________________________

_____________________________________________________________________________________
How to Calculate the Sample Correlation Coefficient $r$

Let $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ represent the coordinates of $n$ points on a scatterplot.

- $\bar{x}$ = the sample mean of the x-coordinates
- $\bar{y}$ = the sample mean of the y-coordinates
- $s_x$ = the sample standard deviation of the x-coordinates
- $s_y$ = the sample standard deviation of the y-coordinate

$x_i$ = x-coordinate of point number $i$ and the z-score of $x_i = \frac{x_i - \bar{x}}{s_x}$

$y_i$ = y-coordinate of point number $i$ and the z-score of $y_i = \frac{y_i - \bar{y}}{s_y}$

The correlation coefficient $r$ equals the average of the products of the z-scores, except that we divide by $n - 1$ instead of $n$.

**Formula 1:**

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

Formula 1 above is useful in that it helps us understand the basic idea behind the correlation coefficient, however, computer programs use the formula 2 below to calculate $r$ because it is much easier to program.

**Formula 2:**

$$r = \frac{n \left( \sum xy \right) - \left( \sum x \right) \left( \sum y \right)}{\sqrt{n \left( \sum x^2 \right) - \left( \sum x \right)^2} \sqrt{n \left( \sum y^2 \right) - \left( \sum y \right)^2}}$$

How to calculate the equation of the regression line $y = a + bx$ or $y = b_0 + b_1 x$ where $n$ = the number of x-y data pairs.

$b =$ slope of regression line $= \frac{n \left( \sum xy \right) - \left( \sum x \right) \left( \sum y \right)}{n \left( \sum x^2 \right) - \left( \sum x \right)^2}$

$a =$ the y-intercept of the regression line $= \bar{y} - b \bar{x}$.

The point $(\bar{x}, \bar{y})$ is always on the regression line.
Example 2: Cricket chirps and Temperature

Chirps in a minute: 882, 1,188, 1,104, 864, 1,200, 1,032, 960, 900
Temperature in °F: 69.7, 93.3, 84.3, 76.3, 88.6, 82.6, 71.6, 79.6

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>xy</th>
<th>x²</th>
<th>y²</th>
</tr>
</thead>
<tbody>
<tr>
<td>882</td>
<td>69.7</td>
<td>61,475.4</td>
<td>777,924</td>
<td>4,858.09</td>
</tr>
<tr>
<td>1,188</td>
<td>93.3</td>
<td>110,840.4</td>
<td>1,411,344</td>
<td>8,704.89</td>
</tr>
<tr>
<td>1,104</td>
<td>84.3</td>
<td>93,067.2</td>
<td>1,218,816</td>
<td>7,106.49</td>
</tr>
<tr>
<td>864</td>
<td>76.3</td>
<td>65,923.2</td>
<td>746,496</td>
<td>5,821.69</td>
</tr>
<tr>
<td>1,200</td>
<td>88.6</td>
<td>106,320.0</td>
<td>1,440,000</td>
<td>7,849.96</td>
</tr>
<tr>
<td>1,032</td>
<td>82.6</td>
<td>85,243.2</td>
<td>1,065,024</td>
<td>6,822.76</td>
</tr>
<tr>
<td>960</td>
<td>71.6</td>
<td>68,736.0</td>
<td>921,600</td>
<td>5,126.56</td>
</tr>
<tr>
<td>900</td>
<td>79.6</td>
<td>71,640.0</td>
<td>810,000</td>
<td>6,336.16</td>
</tr>
<tr>
<td>8,130</td>
<td>646.0</td>
<td>663,245.4</td>
<td>8,391,204</td>
<td>52,626.60</td>
</tr>
</tbody>
</table>

n = 8, \bar{x} = \frac{8,130}{8} = 1,016.25 \text{ and } \bar{y} = \frac{646}{8} = 80.75

b = \frac{8 \cdot 663,245.4 - 8,130 \cdot 646}{8 \cdot 8,391,204 - 8,130^2} = 0.05227

a = 80.75 - 0.05227 \cdot 1,016.25 = 27.63

Equation of regression line: \( y = 0.0523x + 27.63 \)

\( r = \frac{8 \cdot 663,246.4 - 8,130 \cdot 646}{\sqrt{8 \cdot 8,391,204 - 8,130^2 \cdot \sqrt{8 \cdot 52,626.60 - 646^2}}} = 0.8738 \)

a) Plot the x-y data pairs on the axes provided. Label the axes appropriately.

b) Find the equation of the least-squares regression line.

c) Predict the Fahrenheit temperature F when the number chirps = 900. chrips = 1,100 = 

d) Graph the least-squares regression line. Plot and label the point ( \( \bar{x}, \bar{y} \) ) which is always on the regression line.

e) Find the slope of the regression line and explain what it tells us.

f) Find the correlation coefficient \( r \) of the data set and explain what it tells us.

g) Find the coefficient of determination \( r^2 \) and explain what it tells us.
Example 3: In the table below, the explanatory variable x equals a person's credit score and the response variable y equals the interest rate charged for a 36-month auto loan.

<table>
<thead>
<tr>
<th>x</th>
<th>545</th>
<th>595</th>
<th>640</th>
<th>675</th>
<th>705</th>
<th>750</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>18.982%</td>
<td>17.967%</td>
<td>12.218%</td>
<td>8.612%</td>
<td>6.680%</td>
<td>5.150%</td>
</tr>
</tbody>
</table>

\[ y = 11.6015 \]
\[ x = 651.6667 \]

a) Plot the x-y data pairs on your graphing calculator.

b) Find the equation of the least-squares regression line.

\[ y = mx + b \]

\[ x = \bar{x} = 651.6667 \]
\[ y = \bar{y} = 11.6015 \]

\[ m = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \]
\[ b = \bar{y} - m\bar{x} \]

\[ y = mx + b \]

\[ m = \frac{-11.982}{-150} = 0.0799 \]
\[ b = 11.6015 - 0.0799 \times 651.6667 = -4.083 \]

\[ y = 0.0799x - 4.083 \]

\[ r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \]

\[ r = \frac{-11.982 \times -5.150}{\sqrt{150 \times 11.6015^2}} \approx 0.999 \]

\[ r^2 = 0.998 \]

\[ \bar{x} = 651.6667 \]
\[ \bar{y} = 11.6015 \]

\[ (x, y) \text{ is always on the regression line.} \]

\[ \text{Slope: } m = 0.0799 \]
\[ \text{Interpretation: Every increase in credit score by 1 point results in an increase of 0.0799 percentage points in the interest rate.} \]

\[ r \approx 0.999 \]
\[ r^2 \approx 0.998 \]
\[ \text{Interpretation: The linear relationship between credit score and interest rate is strong.} \]

\[ \text{Why would it not be a good idea to predict the interest rate for a credit score of 800?} \]
\[ \text{The model is likely to become less accurate for credit scores outside the range of the data set.} \]
Correlation Coefficient Critical Values for 0.05 and 0.01 Significance Levels

<table>
<thead>
<tr>
<th>Sample Size n</th>
<th>0.05</th>
<th>0.01</th>
<th>Sample Size n</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.950</td>
<td>0.990</td>
<td>21</td>
<td>0.433</td>
<td>0.549</td>
</tr>
<tr>
<td>5</td>
<td>0.878</td>
<td>0.959</td>
<td>22</td>
<td>0.423</td>
<td>0.537</td>
</tr>
<tr>
<td>6</td>
<td>0.811</td>
<td>0.917</td>
<td>23</td>
<td>0.413</td>
<td>0.526</td>
</tr>
<tr>
<td>7</td>
<td>0.754</td>
<td>0.875</td>
<td>24</td>
<td>0.404</td>
<td>0.515</td>
</tr>
<tr>
<td>8</td>
<td>0.707</td>
<td>0.834</td>
<td>25</td>
<td>0.396</td>
<td>0.505</td>
</tr>
<tr>
<td>9</td>
<td>0.666</td>
<td>0.798</td>
<td>26</td>
<td>0.388</td>
<td>0.496</td>
</tr>
<tr>
<td>10</td>
<td>0.632</td>
<td>0.765</td>
<td>27</td>
<td>0.381</td>
<td>0.487</td>
</tr>
<tr>
<td>11</td>
<td>0.602</td>
<td>0.735</td>
<td>28</td>
<td>0.374</td>
<td>0.479</td>
</tr>
<tr>
<td>12</td>
<td>0.576</td>
<td>0.708</td>
<td>29</td>
<td>0.367</td>
<td>0.471</td>
</tr>
<tr>
<td>13</td>
<td>0.553</td>
<td>0.684</td>
<td>30</td>
<td>0.361</td>
<td>0.463</td>
</tr>
<tr>
<td>14</td>
<td>0.532</td>
<td>0.661</td>
<td>35</td>
<td>0.335</td>
<td>0.430</td>
</tr>
<tr>
<td>15</td>
<td>0.514</td>
<td>0.641</td>
<td>40</td>
<td>0.312</td>
<td>0.403</td>
</tr>
<tr>
<td>16</td>
<td>0.497</td>
<td>0.623</td>
<td>45</td>
<td>0.294</td>
<td>0.380</td>
</tr>
<tr>
<td>17</td>
<td>0.482</td>
<td>0.606</td>
<td>50</td>
<td>0.279</td>
<td>0.361</td>
</tr>
<tr>
<td>18</td>
<td>0.468</td>
<td>0.590</td>
<td>60</td>
<td>0.254</td>
<td>0.330</td>
</tr>
<tr>
<td>19</td>
<td>0.456</td>
<td>0.575</td>
<td>70</td>
<td>0.236</td>
<td>0.305</td>
</tr>
<tr>
<td>20</td>
<td>0.444</td>
<td>0.561</td>
<td>80</td>
<td>0.220</td>
<td>0.286</td>
</tr>
</tbody>
</table>

Interpreting the correlation coefficient $r$ and the coefficient of determination $r^2$

For sample sizes $n \geq 4$, $r$ is statistically significant if $|r| >$ the critical value.

**Example 1: $n = 20$ and $r = 0.587$**

With $n = 20$ and $r = 0.587$, we can say there is a statically significant linear relationship between the explanatory variable and the response variable at the 0.01 level of significance. There is at most a 1% chance that this apparent relationship is due to chance or other unknown factors.

The coefficient of determination $r^2 = 0.3446$. This tells us about 34% of the variation or change in the response variable can be explained by variation or change in the explanatory variable. The remaining 66% of the variation in the response variable is unexplained and is due to chance or other unknown factors.

**Example 2: $n = 9$ and $r = -0.758$**

With $n = 9$ and $r = -0.758$, we can say there is a statically significant linear relationship between the explanatory variable and the response variable at the 0.05 level of significance. There is at most a 5% chance that this apparent relationship is due to chance or other unknown factors.

The coefficient of determination $r^2 = 0.5746$. This tells us about 57% of the variation or change in the response variable can be explained by variation or change in the explanatory variable. The remaining 43% of the variation in the response variable is unexplained and is due to chance or other unknown factors.

**Example 3: $n = 16$ and $r = 0.478$**

With $n = 9$ and $r = 0.478$, we can say there is no statistically significant linear relationship between the explanatory variable and the response variable at the 0.01 or 0.05 level of significance.