Chapter 12
Linear Regression & Correlation

\[ f(x) = y = mx + b \quad \text{(Algebra)} \]
\[ \hat{y} = b_0 + b_1 x \quad \text{(Statistics)} \]

\[ r = \begin{cases} \text{strength \\ & of line} \end{cases} \quad \text{(chapter 3)} \]

\[ \begin{cases} \text{dependent variable} \\ \text{predicted variable} \end{cases} = y = \hat{y} = f(x) \]
\[ \begin{cases} \text{independent variable} \\ \text{predictor variable} \end{cases} = x \]
\[ \{y\text{-intercept}\} = b = b_0 = \text{INTERCEPT Excel function} \]
\[ \{\text{slope}\} = \left\{ \text{how much } y \text{ moves} \right\} = m = b_1 = \text{SLOPE Excel function} \]

\[ r = \begin{cases} \text{strength \\ & of linear equation} \end{cases} \]

\[ r^2 = \begin{cases} \text{coefficient of determination} \\ "\text{goodness of fit of equation}" \end{cases} \]

\[ r = \text{PEARSON Excel function} \]
\[ r^2 = \text{RSQ Excel function} \]
Chapter 12: Simple Linear Regression

1. What we already know about "Relationship between 2 quantitative variables" from math & chapter 3:
   - **Independent variable (X)**
     - predictor variable.
   - **Dependent variable (Y or f(x) or E(y) or y)**
     - variable that is predicted or estimated

   **Example:**
   - Collected data from past:

   **Scatter Diagram/chart/plot**
   - Graphical technique to show relationship between 2 quantitative variables

<table>
<thead>
<tr>
<th>X: Ad Spend per week</th>
<th>Y: Weekly Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>14,000</td>
<td>97,000</td>
</tr>
<tr>
<td>27,000</td>
<td>181,000</td>
</tr>
<tr>
<td>49,000</td>
<td>260,000</td>
</tr>
<tr>
<td>17,000</td>
<td>143,000</td>
</tr>
<tr>
<td>34,000</td>
<td>230,000</td>
</tr>
<tr>
<td>43,000</td>
<td>398,000</td>
</tr>
</tbody>
</table>
Coefficient of correlation (r) (Interval or Ratio) (p3)

1. Measure of the strength and direction of the linear relationship (between -1 and 1).
   - 0 = No correlation.
   - 0.5 or -0.5 = Moderate correlation.
   - Near -1 or 1 = Strong correlation.

Coefficient of determination (r^2)

A measure of the goodness of fit of the estimated regression equation. The proportion of the total variation in the dependent variable, y, that is explained, or accounted for, by the variation in the independent variable, x. It does not say anything about causation, that is, it does not say that x causes y.
The coefficient of correlation is given by:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)S_xS_y}$$

This formula involves:
- $n$: count of observed pairs
- $S_x$: standard deviation of $x$'s
- $S_y$: standard deviation of $y$'s

The equation is only true for independent variables.

**Example**
- $x$ values: 14,000, 17,000, 27,000, 34,000, 59,000, 113,000
- $y$ values: 9,700, 143,000, 185,000, 270,000, 260,000, 398,000
- Mean $\bar{x} = 29,150$ and $\bar{y} = 225,500$
- Standard deviations $S_x = 11,938.8$ and $S_y = 107,596.93$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$x-\bar{x}$</th>
<th>$y-\bar{y}$</th>
<th>$(x-\bar{x})(y-\bar{y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>14,000</td>
<td>9,700</td>
<td>-1,150</td>
<td>-28,800</td>
<td>33,060</td>
</tr>
<tr>
<td>17,000</td>
<td>143,000</td>
<td>-1,150</td>
<td>-26,600</td>
<td>30,420</td>
</tr>
<tr>
<td>27,000</td>
<td>185,000</td>
<td>-1,150</td>
<td>-39,300</td>
<td>45,350</td>
</tr>
<tr>
<td>34,000</td>
<td>270,000</td>
<td>-1,150</td>
<td>-8,100</td>
<td>9,315</td>
</tr>
<tr>
<td>59,000</td>
<td>260,000</td>
<td>-1,150</td>
<td>19,500</td>
<td>-22,025</td>
</tr>
<tr>
<td>113,000</td>
<td>398,000</td>
<td>-1,150</td>
<td>9,430</td>
<td>-10,830</td>
</tr>
</tbody>
</table>

$\sum = 0$ \quad $\sum = 0$  

Thus, the coefficient is $r = \frac{6,920,950,000}{5 \times 119,388 \times 107,596.93} = .936074653$. This means the strength is strong and the direction is direct, as $x$ increases, so does $y$.

In Excel: `=PEARSON(yrange, xrange)`
### Coefficient of Determination

\[
\{ \text{Coefficient of Determination} \} = r^2 = \left( \frac{\text{Influence of } X \text{ on } Y}{\text{Not Causation}} \right)^2 = 0.936034 \approx 0.876
\]

### Example 1:

Ad dollars spent & sales earned

- 0.936034 indicates that the relationship between Ad dollars spent & sales earned is very strong. In addition, the relationship is direct, which means that as Ad dollars increase, sales increase. However, "strong" is not numerically precise. But coefficient of determination is numerically precise.

We can say that 87.6% of the variation in \( y(f(x)) \) can be explained by the variation in \( x \) (not causation). Correlation does not mean causation!! (from textbook: As population of donkeys decreases, number of doctoral degrees increases.)

This is called "spurious correlations." Conclusion: correlation is great for building models we can use to make predictions, but correlation does not mean causation.
Estimated Simple Linear Regression Equation

An equation that expresses the linear relationship between 2 variables

\[ y = mx + b \quad (\text{Algebra}) \]
\[ \hat{y} = b_0 + b_1x \quad (\text{Statistics}) \]

\[ y = \hat{y} \quad \text{predicted variable = dependent variable} \]
\[ x = \bar{x} \quad \text{predictor variable = independent variable} \]
\[ m = b_1 \quad \text{slope = how much y moves for 1 unit of x} \]
\[ b_0 = b_0 \quad \text{y-intercept} \]
\[ \bar{x} = \text{sample mean} \]
\[ \bar{y} = \text{sample mean} \]

\[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

\[ b_0 = \bar{y} - b_1 \bar{x} \]
\[
\bar{x} = 29,150 \quad n = 6 \quad n - 1 = 5
\]

\[
\bar{y} = 225,500
\]

\[
\sum x = 11,938,80
\]

\[
\sum y = 1,075,969,93
\]

From bottom:

\[
y = \hat{b}_0 + \hat{b}_1 x
\]

\[
y = -20,406 + 8.44x
\]

Can use to estimate!!

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>((x - \bar{x})(y - \bar{y}))</th>
<th>((x - \bar{x})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14,000</td>
<td>97,000</td>
<td>((14,000 - 29,150)(97,000 - 225,500))</td>
<td>((14,000 - 29,150)^2)</td>
</tr>
<tr>
<td>27,000</td>
<td>187,000</td>
<td>((27,000 - 29,150)(187,000 - 225,500))</td>
<td>((27,000 - 29,150)^2)</td>
</tr>
<tr>
<td>79,900</td>
<td>260,000</td>
<td>((39,900 - 29,150)(260,000 - 225,500))</td>
<td>((39,900 - 29,150)^2)</td>
</tr>
<tr>
<td>17,000</td>
<td>149,000</td>
<td>((17,000 - 29,150)(149,000 - 225,500))</td>
<td>((17,000 - 29,150)^2)</td>
</tr>
<tr>
<td>34,000</td>
<td>270,000</td>
<td>((34,000 - 29,150)(270,000 - 225,500))</td>
<td>((34,000 - 29,150)^2)</td>
</tr>
<tr>
<td>73,000</td>
<td>348,000</td>
<td>((73,000 - 29,150)(348,000 - 225,500))</td>
<td>((73,000 - 29,150)^2)</td>
</tr>
</tbody>
</table>

Total:

\[
\sum (x - \bar{x})(y - \bar{y}) = 6,012,050,000
\]

\[
\sum (x - \bar{x})^2 = 7,126,675,000
\]

\[
\text{Slope} \; m = b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{6,012,050,000}{7,126,675,000} = 8.44
\]

\[
\text{Intercept} \; b_0 = \bar{y} - b_1 \cdot \bar{x} = 225,500 - 8.44 \times 29,150 = -20,406.25
\]
\[ y = -26,406.28 + 8.44X \]

Use to make predictions

If we spend $55,000 on weekly ads, we can expect the sales revenue to be:

\[ f(55,000) = y = 443,567.83 = f(55,000) \]