Correlation

→ pairs of continuous observations.
→ Correlation exists between two variables when one of them is related to the other in some way.

e.g. height and weight of people, temperature and altitude, quiz scores and midterm score

★ Query 1: Do the two variables change together?

★ Query 2: Can changes in one variable predict the changes in the other?

★ Query 3: How do we measure the strength of the relationship between two quantitative variables?
• **Scatterplot**: a graph of the paired samples data.

![August Temperatures vs Elevation in Northern California](scatterplot.png)

**The linear correlation coefficient**, $r$, measures the linear association between two variables.

**Properties**

- $-1 \leq r \leq 1$
- It does not change if we change the scale of the measurement.
- It is sensitive to outliers.

*You need to understand the concept, but you don’t need to know the formula or how to test on it.*

→ If $r = 1$, there is a perfect positive linear relationship.
→ If $r = -1$, there is perfect negative linear relationship.
→ If $r = 0$, there is no relationship.
Correlation is not causation ! ! !

→ Coefficient of determination, $r^2$:

- Gives the proportion of the variation in variable 1 that is explained by the linear association between the two variables.
- $0 \leq r^2 \leq 1$
- 0 indicates no linear relationship while 1 indicates a perfect linear relationship.
Review of lines

- \( y = 1 + 2x \)
- slope = 2 = \( \frac{\Delta y}{\Delta x} \) for each one unit change in \( x \), \( y \) changes by 2 units.
- intercept = 1 : value of \( y \) when \( x = 0 \)

\[ y = b_0 + b_1 x, \]

where \( b_0 \) is the intercept and \( b_1 \) is the slope.
Linear Regression

- Fitting a line to data - to model the relationship between two quantitative variables.

⭐ Lots of lines can be fit to the data - which do we choose?

⭐ Fitted line = regression line = least squares line

⭐ Fitted values = predicted values = values predicted by the line for a particular value of $x$

†† Example: The fitted line is: $\hat{y} = 1 + \frac{1}{2}x$

The fitted values would be:

- $x = 1, \quad \hat{y} = 1 + \frac{1}{2} = \frac{3}{2}$
- $x = 2, \quad \hat{y} = 1 + \frac{1}{2}2 = 2$
- $x = 3, \quad \hat{y} = 1 + \frac{1}{2}3 = \frac{5}{2}$
- $x = 4, \quad \hat{y} = 1 + \frac{1}{2}4 = 3$
**Regression** is the predicting of $Y$ from $X$ assuming a linear relationship.

→ $X$ and $Y$ are **not** treated the same. We are predicting $Y$ from $X$!

* The regression line (least-squares line) is the one that minimizes the sum of squared errors in predicting $Y$ (sum of squared residuals):

\[ \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (b_0 + b_1 x_i - y_i)^2 \]

→ $b_0$ and $b_1$ are chosen to minimize

→ This is always goes through $(\bar{x}, \bar{y})$

Note that this does not minimize the distance (perpendicular) to the line.

• You don’t need to know how to compute $b_0$ and $b_1$ by hand.

• You will need to know how to interpret JMP output, compute predicted values, and do hypothesis tests with JMP.

→ Some data examples
Key Concepts!!!!!

- **Correlation**

- **Slope** and **Intercept**

- **Fitted** vs. **Predicted values**