Simple Linear Regression
Ice Cream Sales vs. Temperature

- If you were asked to describe the pattern between ice cream sales and temperature, you might say “ice cream sales seem to increase as temperature increases.”

- Temperature is called the independent variable, and Sales, the dependent.

- Temperature is also known as the explanatory, or predictor, variable, and Sales, the response.
Parameters that Define Relationships

- **Direction**
  - Positive (direct)
  - Negative (indirect)
- **Form**
  - Linear
  - Non-linear
- **Strength** (weak, strong, moderate)
- **Caution**: Outliers
Simple Linear Regression

- **Linear Regression** is the study of linear, additive relationships between variables.
- With **simple** linear regression, we fit a line to data, thereby describing the linear relationship between exactly **two** variables.
The Formal Problem Statement

- Find the line that “best” fits the data
- More precisely: given a set of \((x,y)\) pairs, find a line such that the squared distance between each of the points and the line is minimized.
- This distance is called the residual. So the problem is to minimize the sum of the squared residuals.
Fitting the “best” line

- The errors would be much larger if we fit this line to our data
- This line that minimizes the sum of the squared residuals
The Regression Equation

The regression equation takes the following form: $y = a + bx$

- $y$ is ice cream sales in dollars
- $a$ is the $y$ intercept of the line (the values of $y$ when $x$ is zero)
- $b$ is the slope of the line
- $x$ is temperature in celsius
Linear Regression in R

- The regression equation for Ice Cream Sales versus Temperature is:

  \[ \text{Sales} = -159.47 + 30.09(\text{Temperature}) \]

- \( b = 30.09 \) is the slope. For a one degree increase in temperature, sales are predicted to increase by 30.09 dollars.

- \( a = -159.47 \) is the y intercept. This value has no particular meaning; it definitely does not mean that when temperature is zero, sales are predicted to be -159.47 dollars!

- Caution: It is dangerous to make predictions outside the range of measured x values.
The Problem

- Given $D = \{(x_i, y_i) \mid i = 1, \ldots, n\}$
- Let $y_i$ represent the $i$th actual value.
- Let $y_i^p$ represent the $i$th predicted value.
  - $y_i^p = b_0 + b_1 x_i$
- Then, the regression problem can be formalized as:
  - Find $b_0$ and $b_1$ that minimize $\sum_i (y_i - y_i^p)^2$
The Solution

• We are trying to minimize a function, so we use the tools of calculus to solve this problem.
  ○ Set the partial derivatives of this function equal to zero, and solve.

• After doing so, the solution to the problem is:
  ○ $b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$
  ○ $b_0 = \bar{y} - b_1 \bar{x}$

• Very interesting: $b_1 = c_{xy} / (s_{xx})^2$
  ○ $c_{xy}$ is sample correlation
  ○ $s_{xx}$ is sample variance

• But remember $r_{xy} = c_{xy} / s_{xx} s_{xy}$, so $b_1 = c_{xy} / (s_{xx})^2 = r_{xy} (s_{xy} / s_{xx})$
  ○ So: if we regress on the $z$-scores of the data (instead of the data values themselves), so that $s_{xy} = s_{xx} = 1$, the slope of the regression line will be the correlation!
The Slope of the Regression Line

Regression Line in Original Units

Regression Line in Standard Units

(average of x, average of y)
Interpreting the Regression Line

In inches: $y = a + bx$

- The intercept $a = -159.47$
- The slope is $b = 30.9$

Each point on the regression line is the result of multiplying temperature by $b$ and adding $a$
Interpreting the Regression Line (cont’d)

In standard units: $y = rx$

- The intercept is 0
- The slope is $r$

Each point on the regression line is the result of multiplying temperature in standard units by $r$ (and adding 0)