NNSE 784
Advanced Analytics Methods

Instructor: F Doyle (CESTM L210)

MW 4:30 – 5:50, NFN 203

Slide Set #14
Linear Regression

Lecture Outline

- Moving toward Machine Learning:
  - Regression – our entry point in this lecture
  - Classification
  - Clustering
- Simple Linear Regression
  - Objectives
  - Population regression equation
  - Sample regression line
  - SSR/SSE/SST
  - R-Squared (R^2)
- Python exercise using simple linear regression
- Multiple Linear Regression
- Python exercise using multiple linear regression

Portions of lecture are adapted from: https://www.youtube.com/watch?v=eYTumjgE2IY and IBM's "Python for Datascience"

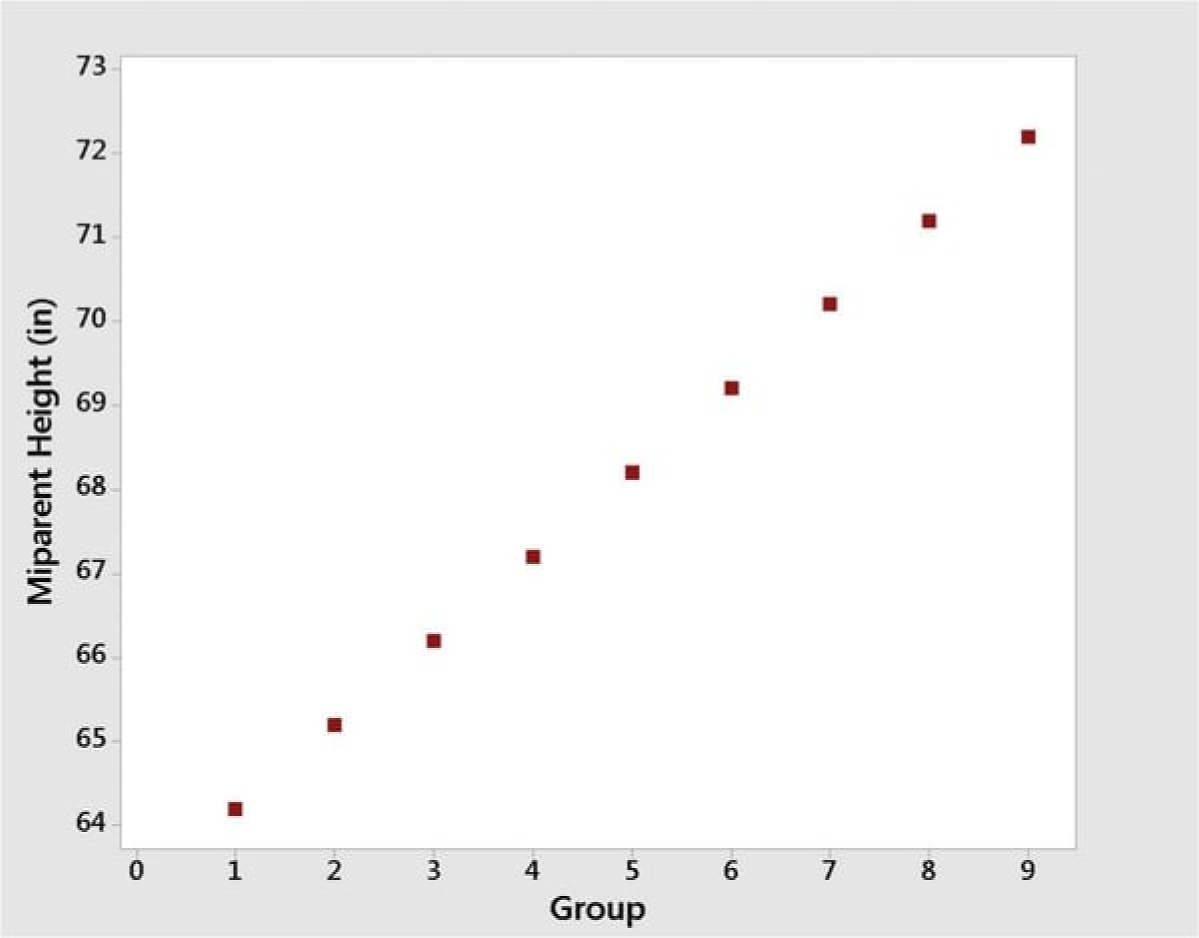
What is “Regression”?

- Process to model the relationship between one or more “input” variables and an “output” variable
- Provides an equation to predict values for the output value based by specifying values of the input variable(s)

Table of “Also Known As”

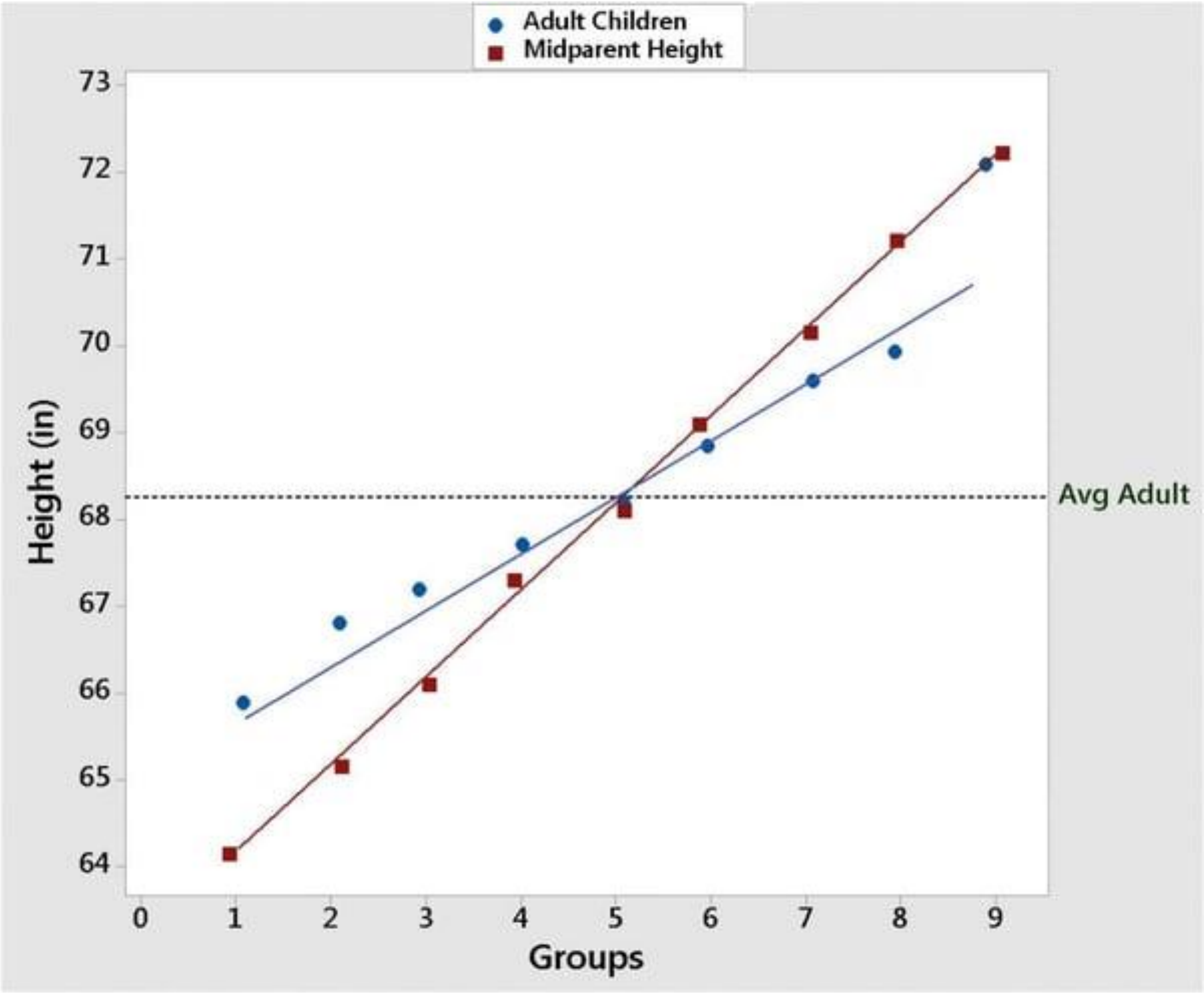
input	output
independent	dependent
regressor	regressand
predictor	predicted
explanatory	explained
features	target

Why is it Called Regression?



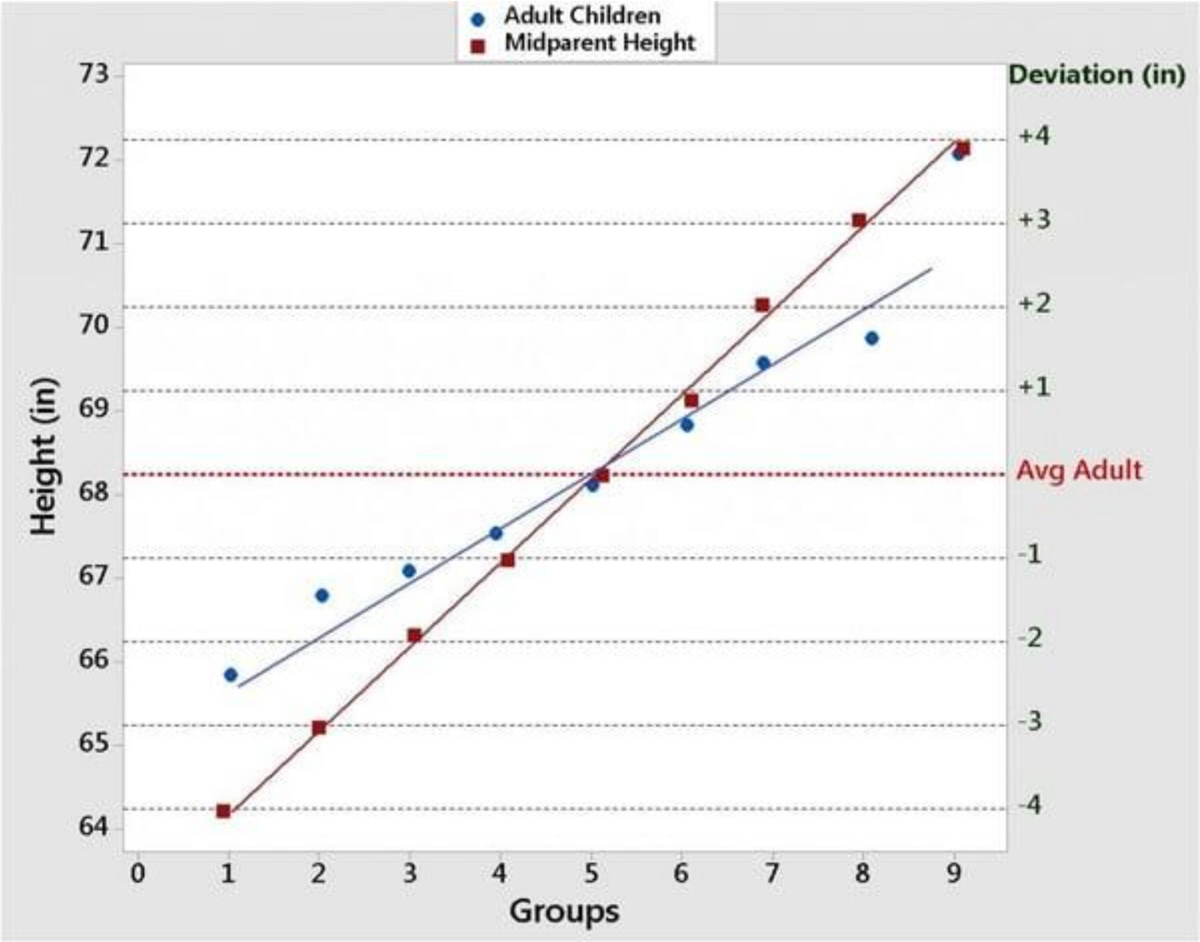
The term dates back to work done in the late 1800s by Sir Francis Galton. Galton was interested in hereditary and in one study he collected data on the heights of 205 sets of parents and their adult children. He calculated the average height of each set of parents and grouped them based on range of heights.

Why is it Called Regression?



Galton then looked at median heights of each group's adult children, fitted lines to both sets of data and plotted a reference line for average adult height.

Why is it Called Regression?



Galton concluded that as heights of parents deviated from the average height, their children tended to be less extreme. That is, the heights of the children "**regressed**" to the average (to 2/3 of the deviation). Galton published these findings in "*Regression towards Mediocrity in Hereditary Stature.*" in 1886

Why is it Called Regression?

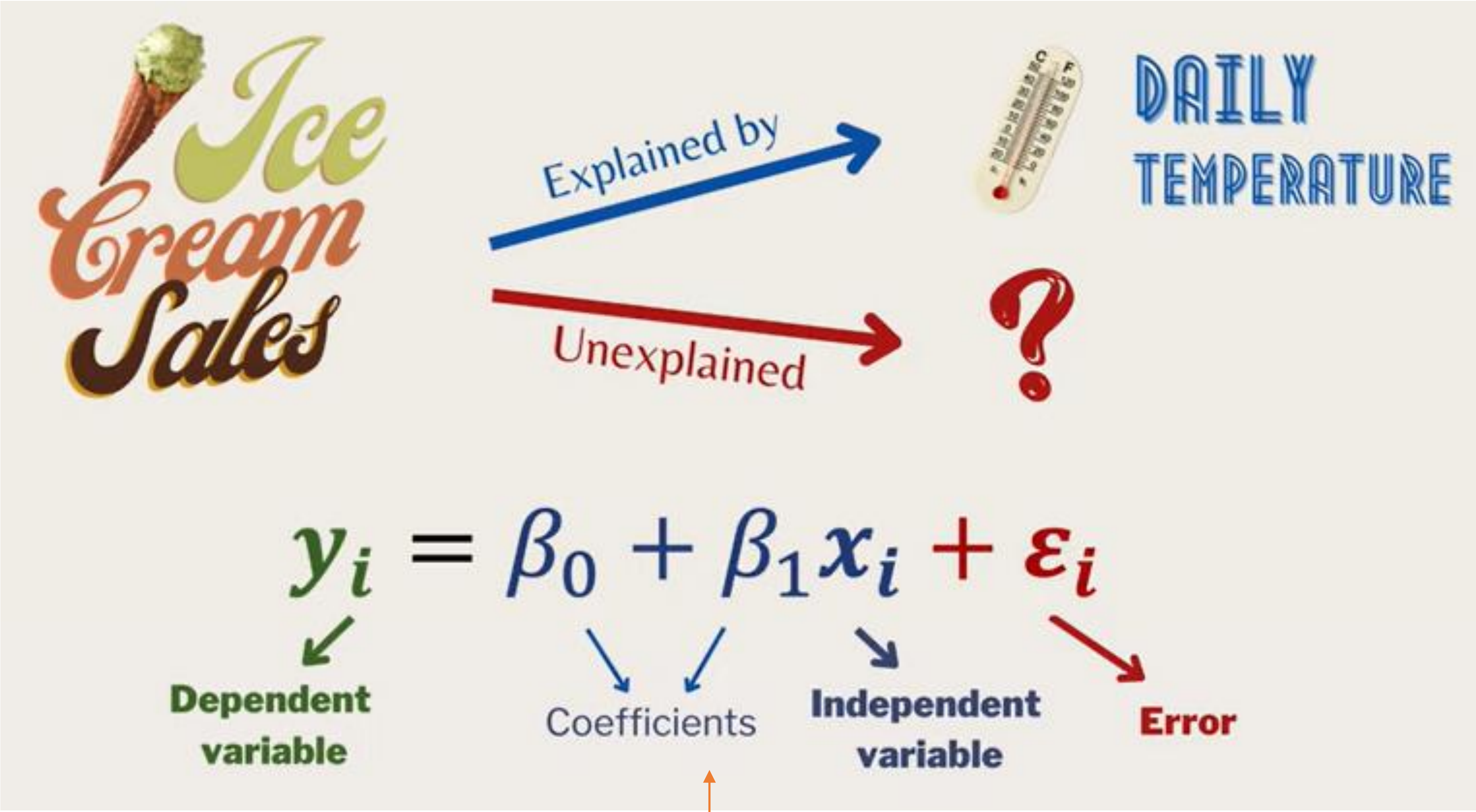
- Galton's work obviously bears little to no resemblance to what we currently think of as "Regression Analysis"
- However, as he and other statisticians built on the methodology that was used to:
  - Quantify correlation relationships
  - Fit lines to data values

  the term "regression" became associated with the statistical analysis that we now use for predicting dependent variable values

Objectives of Regression

- Regression is a means of exploring the variation in some quantity
- The variation is separated into **Explained** and **Unexplained** components

Population Regression Equation



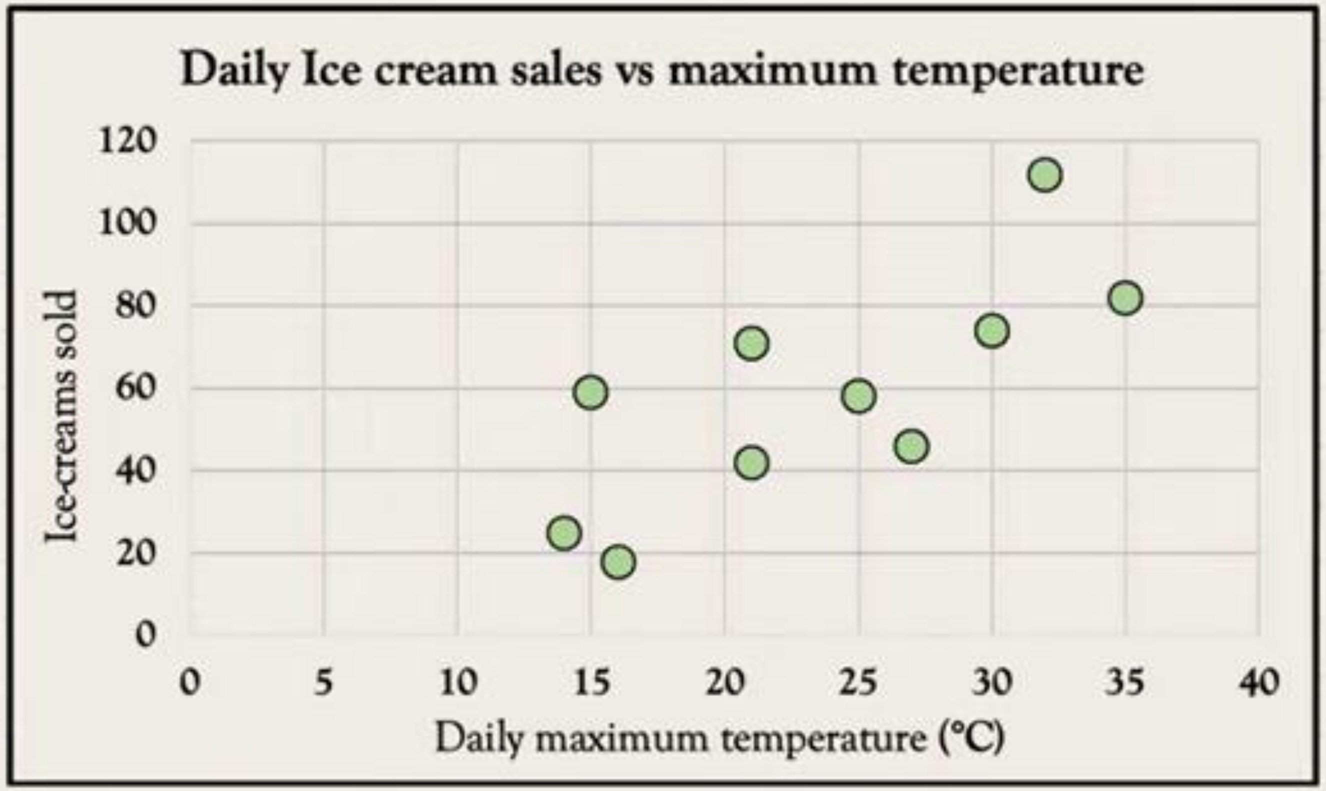
Note the linear form: y = mx + b

Sample Regression Line

The Sample Regression Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \underline{\text{or}} \quad \hat{y} = b_0 + b_1 x$$

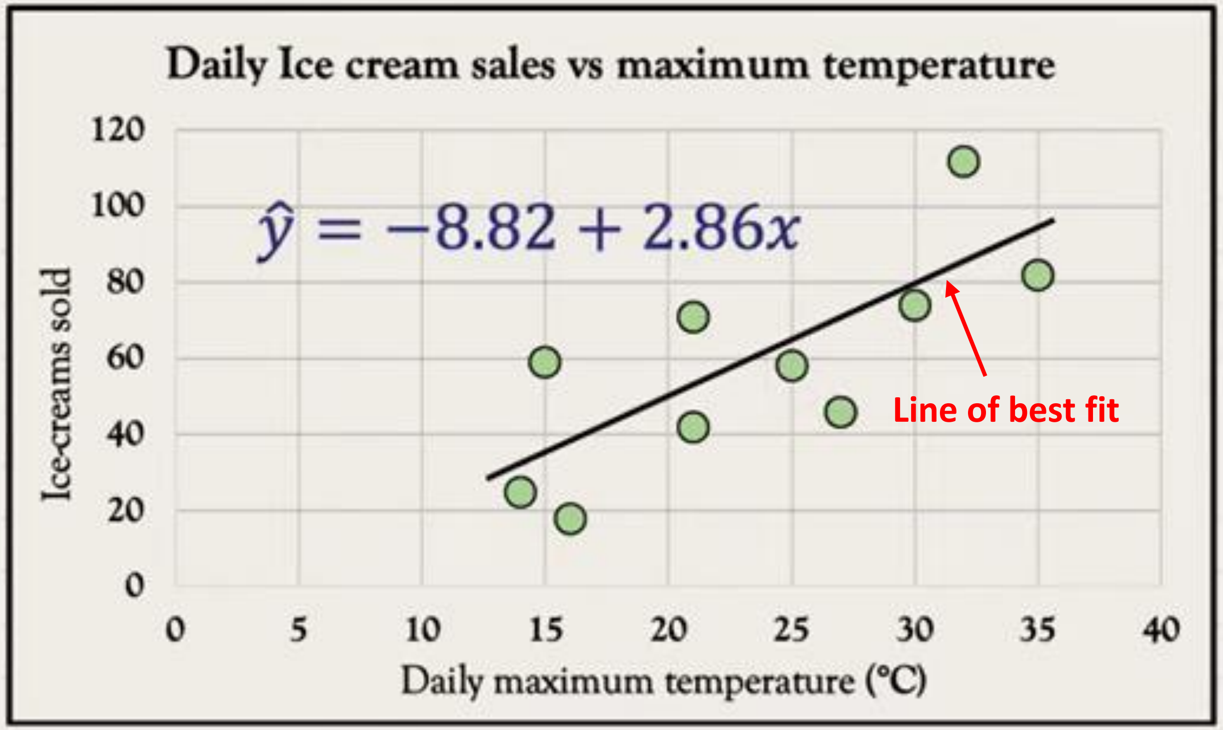
Date	Ice cream Sales	Temp (°C)
Sat 3 June	42	21
Sat 10 June	18	16
Sat 17 June	25	14
Sat 24 June	74	30
Sat 1 July	112	32
Sat 8 July	71	21
Sat 15 July	58	25
Sat 22 July	46	27
Sat 29 July	82	35
Sat 5 August	59	15



Sample Regression Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \underline{\textbf{or}} \quad \hat{y} = b_0 + b_1 x$$

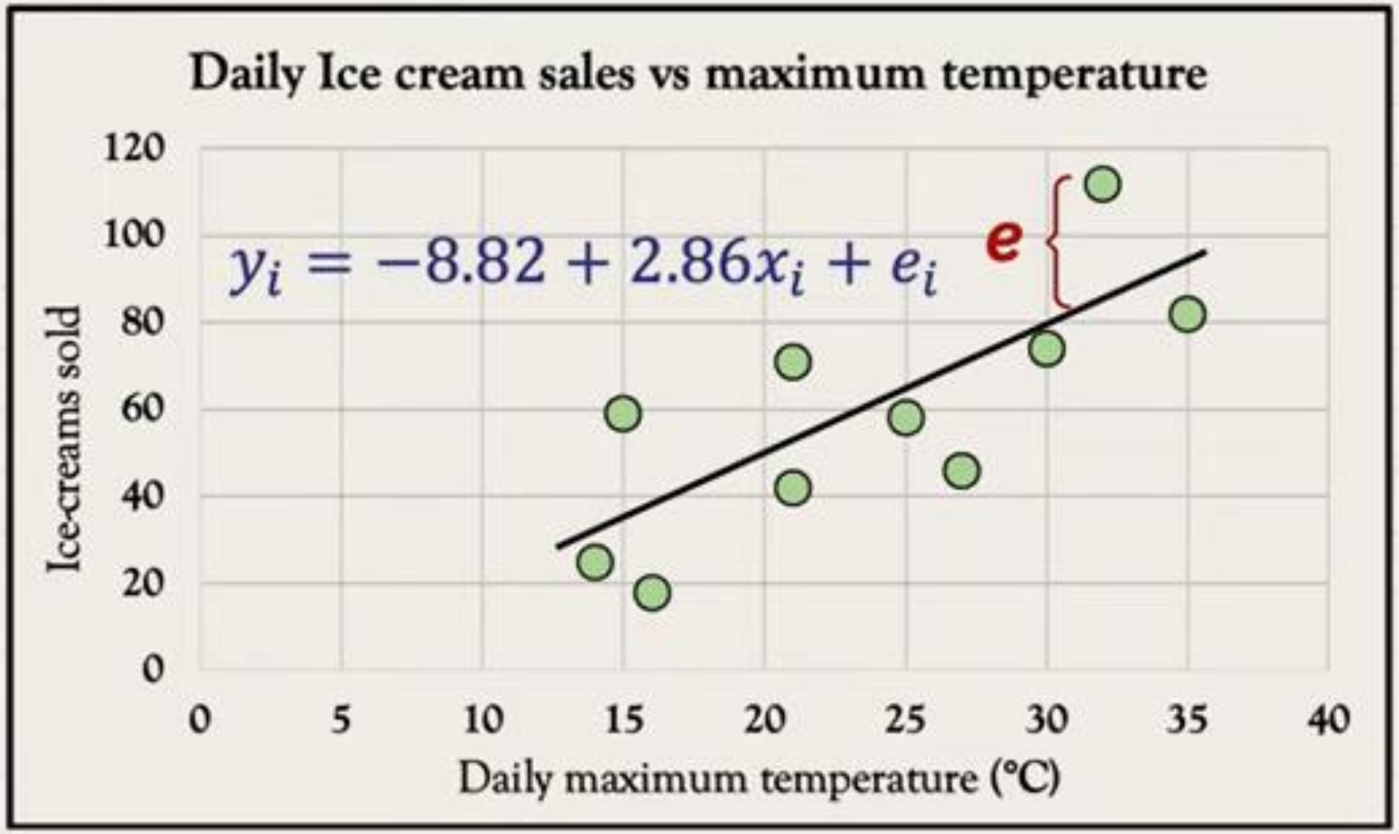
Date	Ice cream Sales	Temp (°C)
Sat 3 June	42	21
Sat 10 June	18	16
Sat 17 June	25	14
Sat 24 June	74	30
Sat 1 July	112	32
Sat 8 July	71	21
Sat 15 July	58	25
Sat 22 July	46	27
Sat 29 July	82	35
Sat 5 August	59	15



Sample Regression Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \underline{\textbf{or}} \quad \hat{y} = b_0 + b_1 x$$

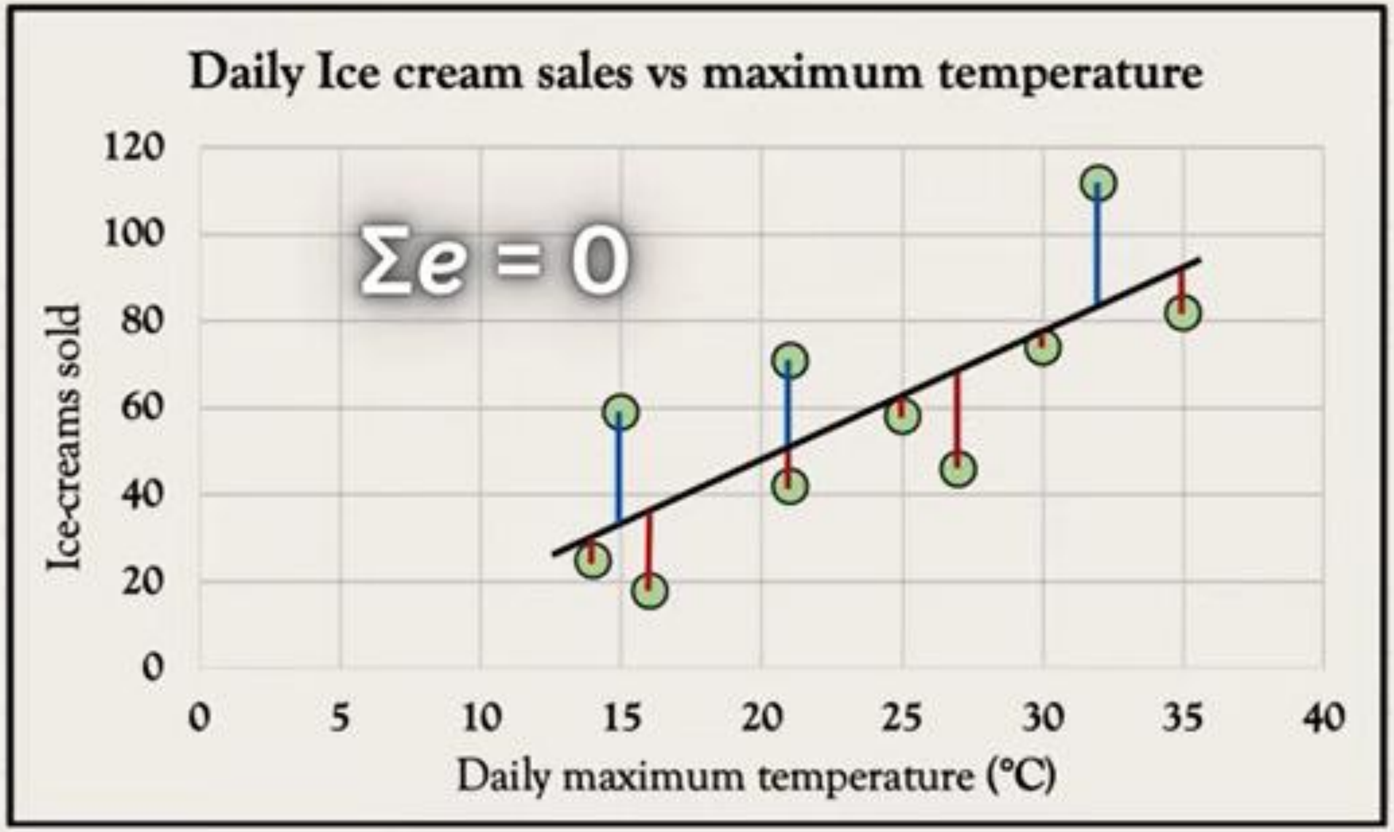
Date	Ice cream Sales	Temp (°C)
Sat 3 June	42	21
Sat 10 June	18	16
Sat 17 June	25	14
Sat 24 June	74	30
Sat 1 July	112	32
Sat 8 July	71	21
Sat 15 July	58	25
Sat 22 July	46	27
Sat 29 July	82	35
Sat 5 August	59	15



Sample Regression Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \underline{\textbf{or}} \quad \hat{y} = b_0 + b_1 x$$

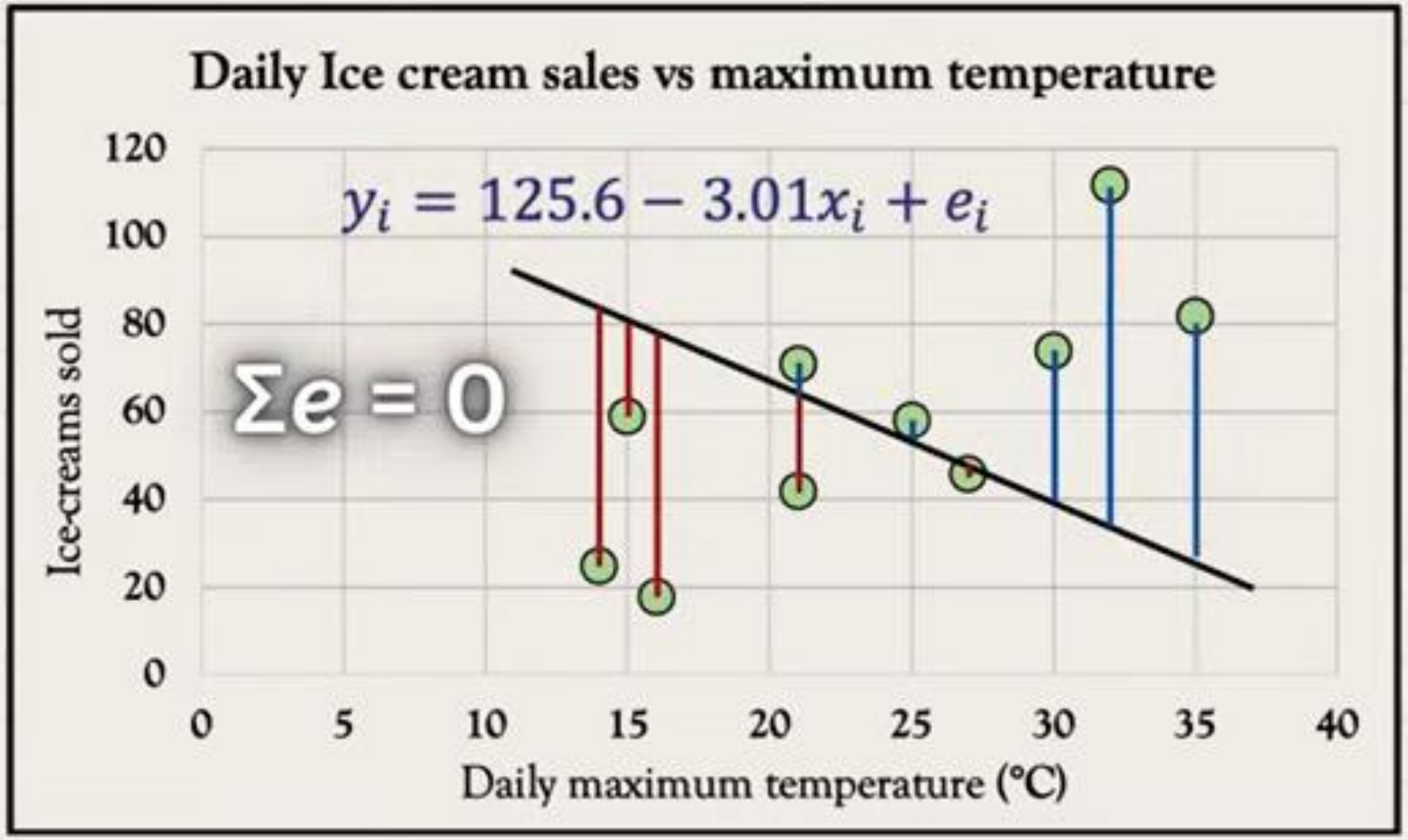
Date	Ice cream Sales	Temp (°C)
Sat 3 June	42	21
Sat 10 June	18	16
Sat 17 June	25	14
Sat 24 June	74	30
Sat 1 July	112	32
Sat 8 July	71	21
Sat 15 July	58	25
Sat 22 July	46	27
Sat 29 July	82	35
Sat 5 August	59	15



Sample Regression Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \underline{\text{or}} \quad \hat{y} = b_0 + b_1 x$$

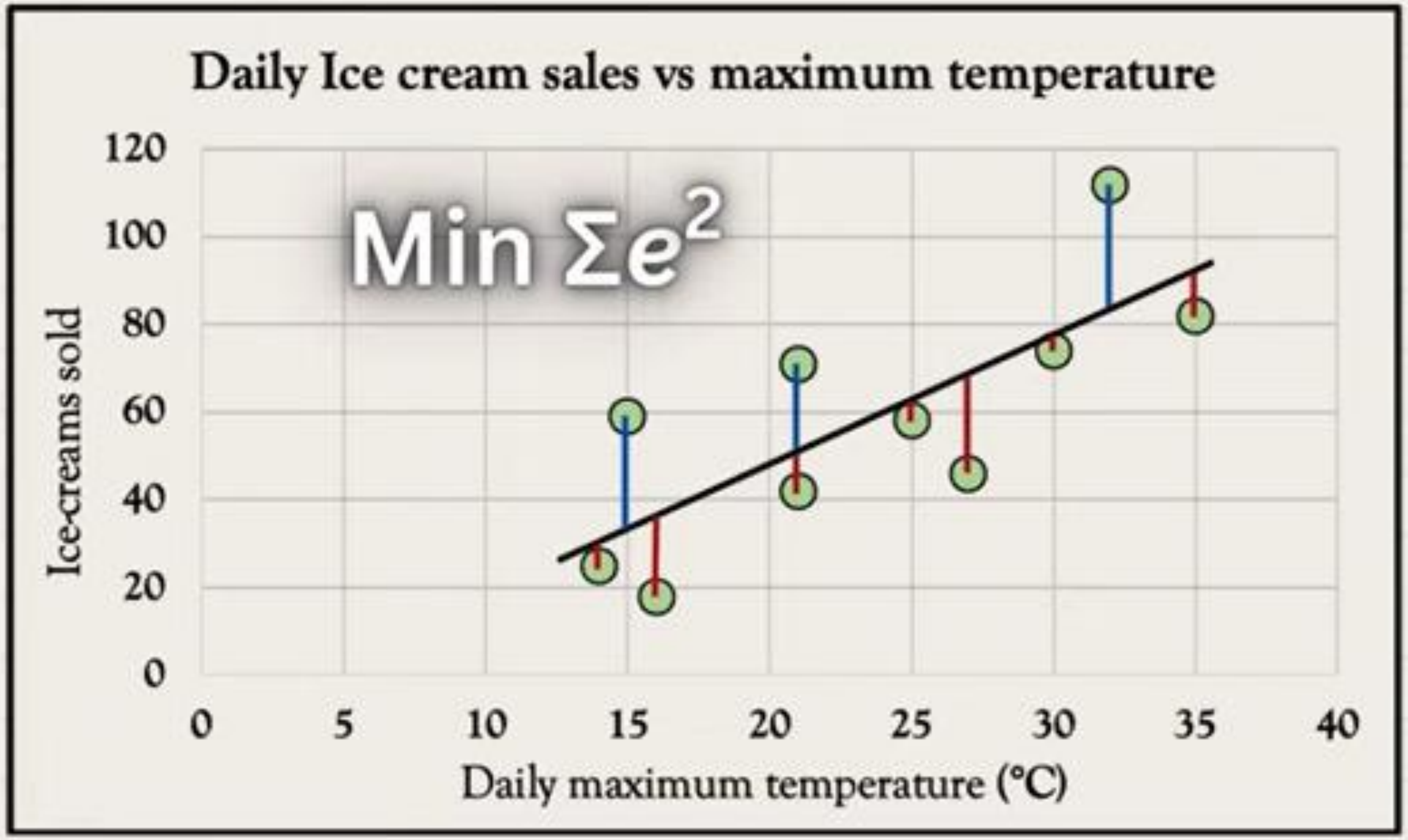
Date	Ice cream Sales	Temp (°C)
Sat 3 June	42	21
Sat 10 June	18	16
Sat 17 June	25	14
Sat 24 June	74	30
Sat 1 July	112	32
Sat 8 July	71	21
Sat 15 July	58	25
Sat 22 July	46	27
Sat 29 July	82	35
Sat 5 August	59	15



Sample Regression Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \underline{\text{or}} \quad \hat{y} = b_0 + b_1 x$$

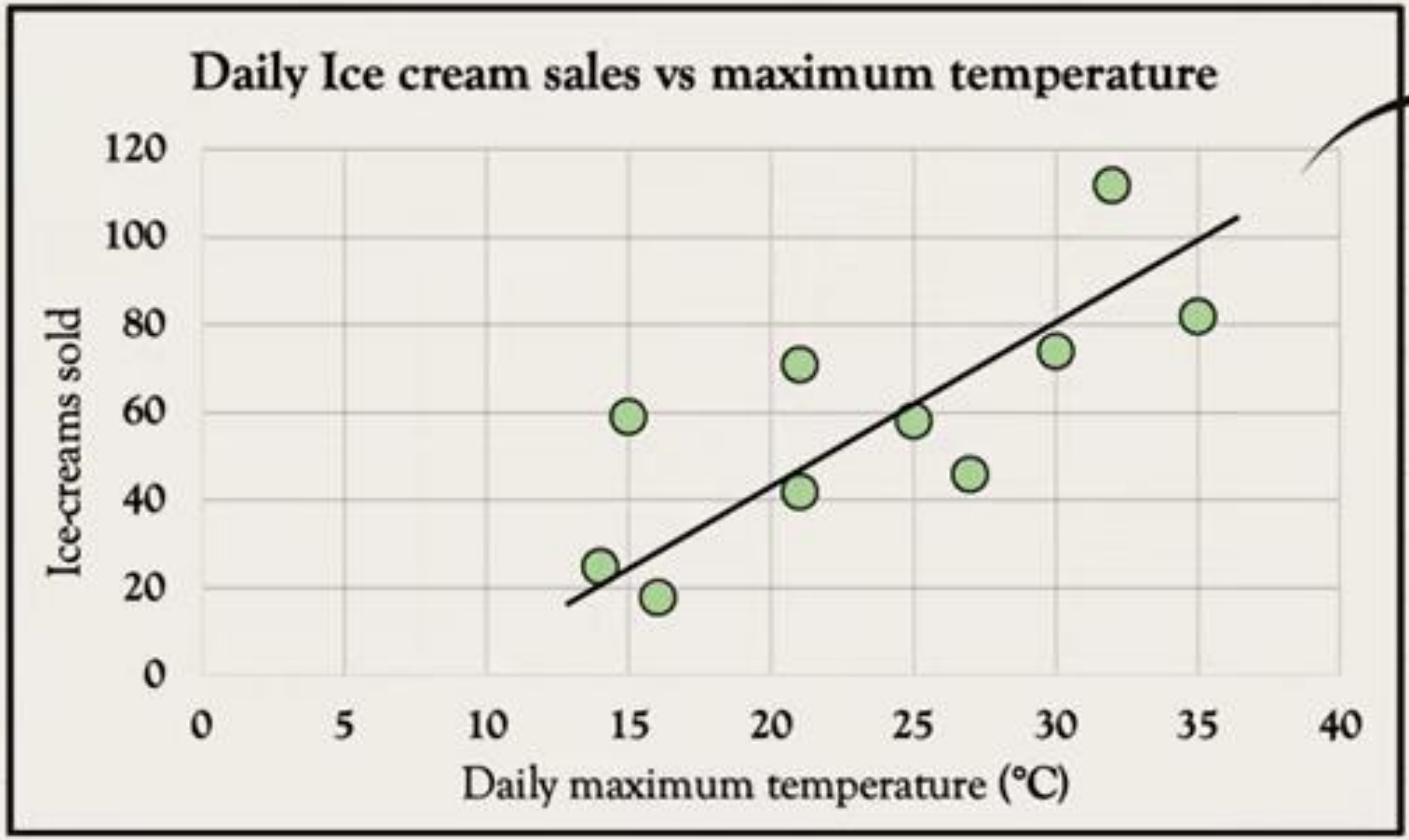
Date	Ice cream Sales	Temp (°C)
Sat 3 June	42	21
Sat 10 June	18	16
Sat 17 June	25	14
Sat 24 June	74	30
Sat 1 July	112	32
Sat 8 July	71	21
Sat 15 July	58	25
Sat 22 July	46	27
Sat 29 July	82	35
Sat 5 August	59	15



Ordinary Least Squares (OLS)

Sample Regression Line

$y_i = b_0 + b_1 x_i + e_i$ is an estimate of $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



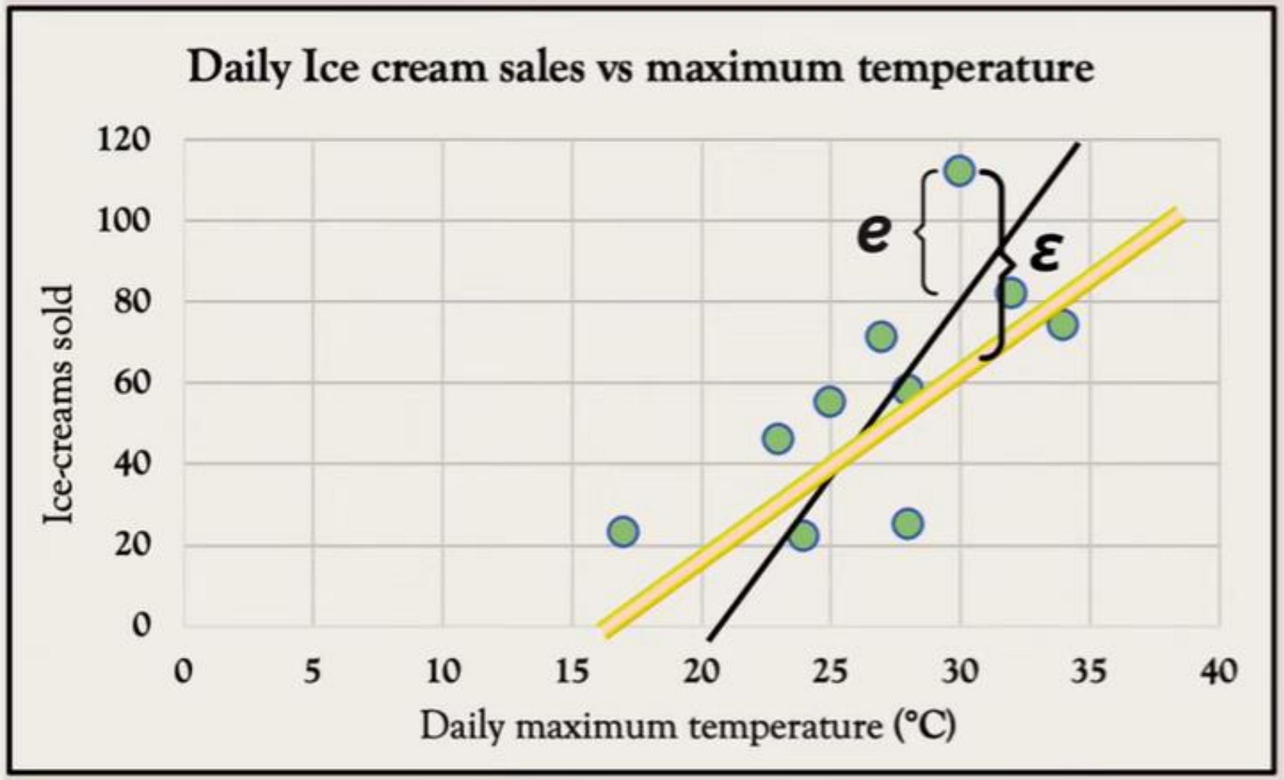
$$\boldsymbol{y_i = -8.82 + 2.86 x_i + e_i}$$

estimate of

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

Sample Regression Line

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



$$\boldsymbol{y_i = -52.4 + 4.08x_i + e_i}$$

estimate of

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

SSR/SSE/SST

Total variance in ice cream sales

- Variance Explained by the temperature
- Variance still unexplained

SST = Sum of Squares (Total)

- SSR = Sum of Squares due to Regression
- SSE = Sum of Squares due to Error

TSS = Total Sum of Squares

- ESS = Explained Sum of Squares
- RSS = Residual Sum of Squares

SSR/SSE/SST

SST = Sum of Squares (Total)

- SSR = Sum of Squares due to Regression
- SSE = Sum of Squares due to Error

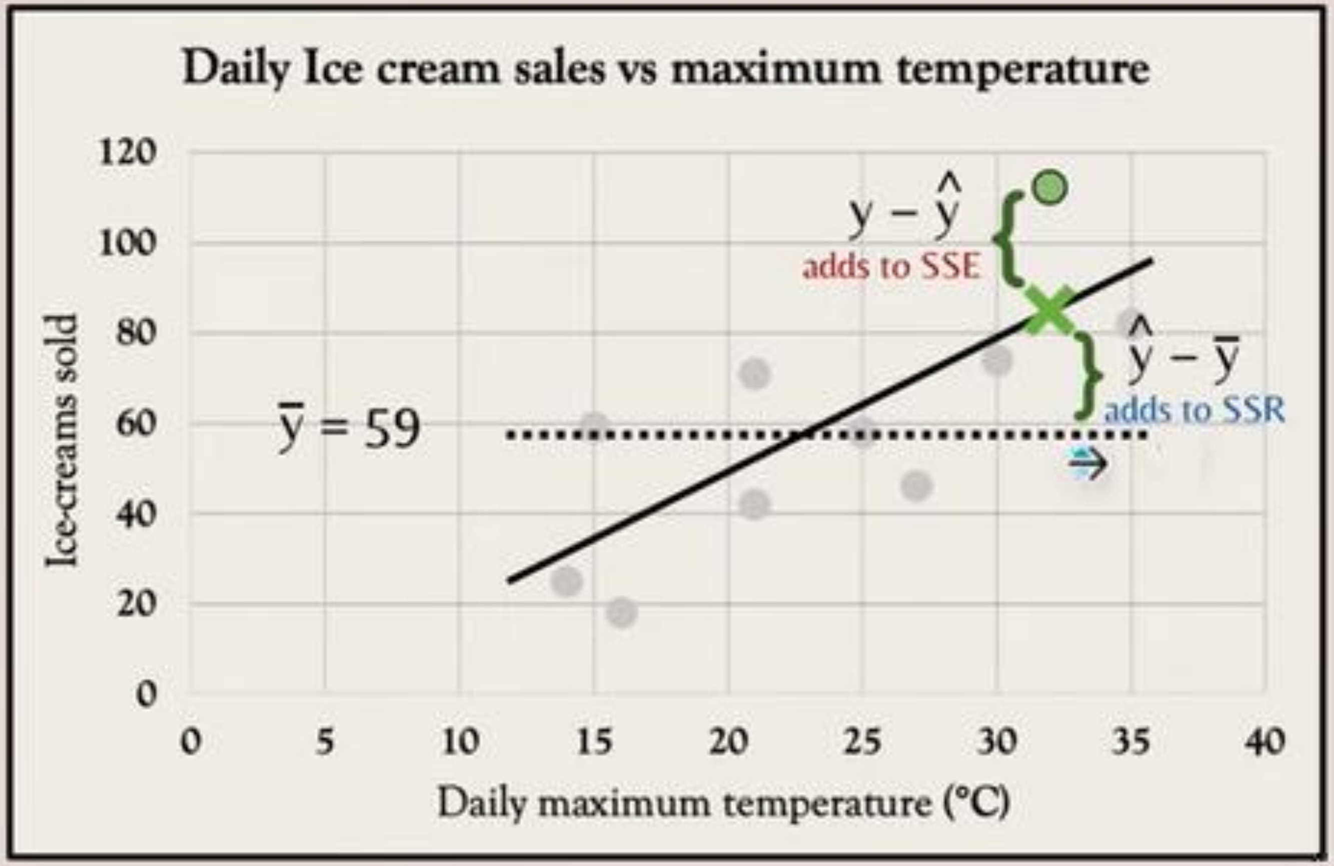
Date	Ice cream Sales	Temp (°C)
Sat 1 July	112	32

$$SST = \sum (y_i - \bar{y}_i)^2$$

$$\boldsymbol{SSR = \sum (\hat{y}_i - \bar{y}_i)^2}$$

$$\boldsymbol{SSE = \sum (y_i - \hat{y}_i)^2}$$

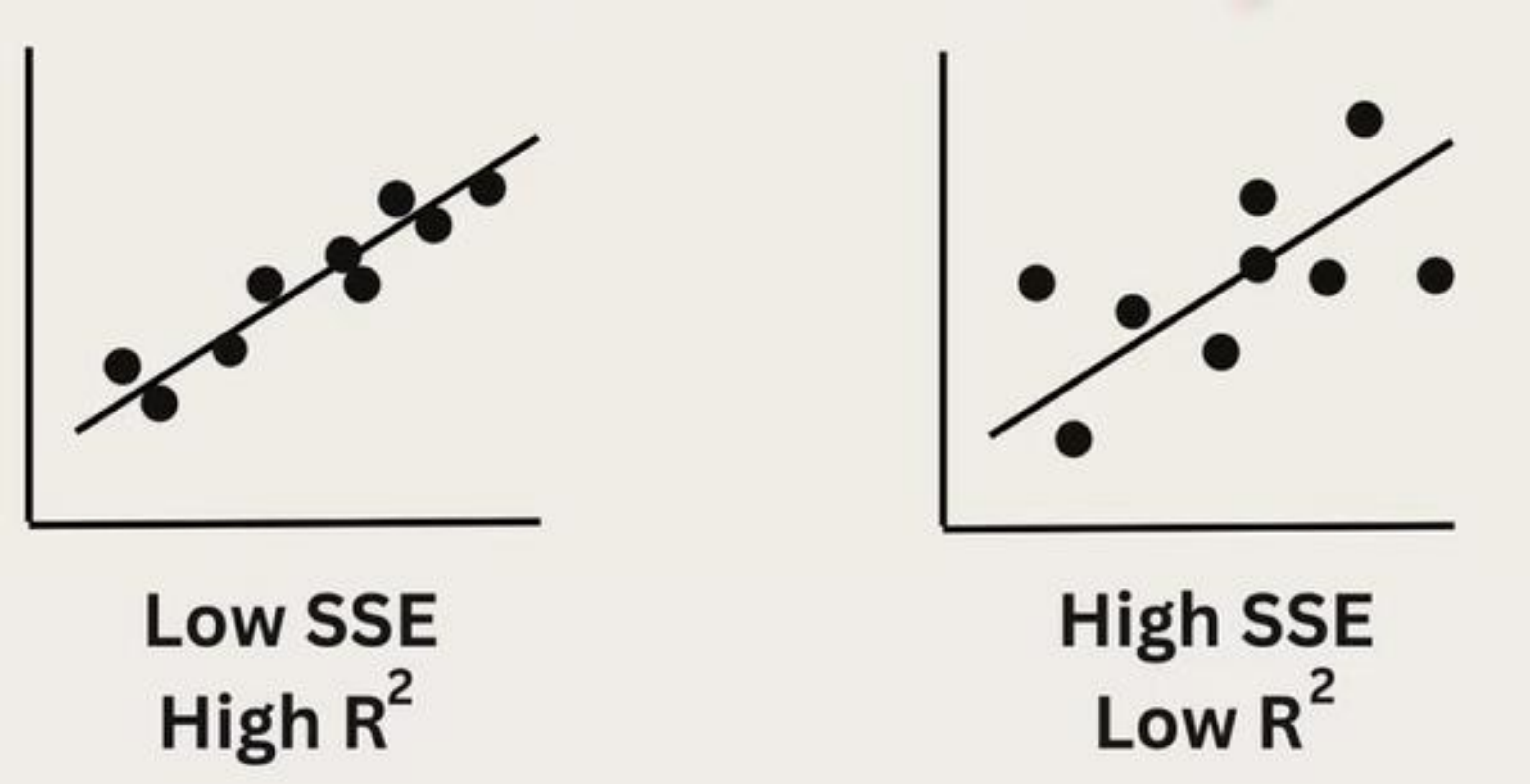
$$\boldsymbol{SST = SSR + SSE}$$



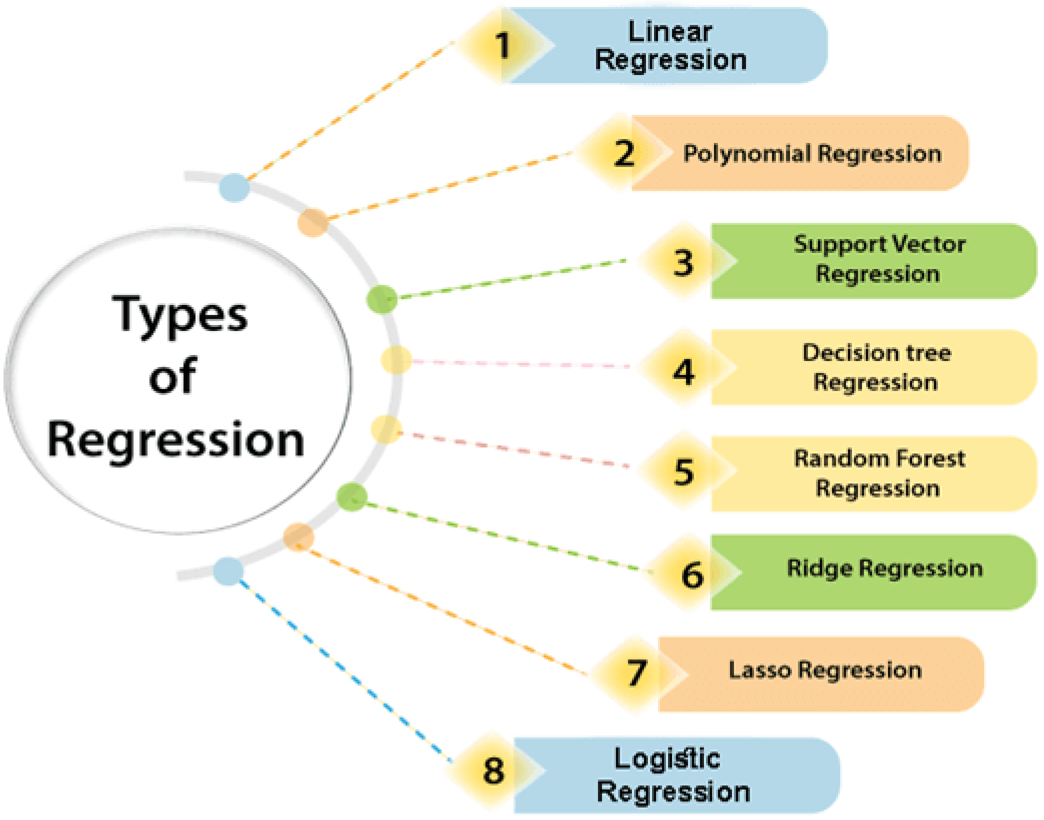
R-squared

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

R-squared is the proportion of the VARIATION in the Y-variable being explained by the Variation in the X-variable(s)



Other Types of Regression



A Note on the Use of “bitwise not” ~

- In the notebook where we went over simple linear regression, there was code that made a random Boolean mask as a numpy array the same length as our dataframe
- We used this mask against our original dataframe to obtain a training set
- We then applied a logical not operation using “~” to the mask to obtain the inverse and used it to derive a test set (all the rows that weren’t include din the training set)
- When we applied “~” to a separate Boolean variable, the behavior was unexpected

The variable ***msk*** is a numpy array. Numpy overrides the inherent Python behavior of the “bitwise not” operator “**~**” to allow it’s use as shown.

```python
import numpy as np
#create a boolean mask of specified length
msk = np.random.rand(10) < 0.8
print([msk])
print([~msk])
```

```
[array([ True, False,  True, False,  True,  True,  True,  True,  True,
        True])]
[array([False,  True, False,  True, False, False, False, False, False,
       False])]
```

When we apply “~” to a base level Python type, it takes the underlying binary representation of the variable and converts 0s to 1s and 1s to 0s.

```python
bool_var1 = True
bool_var2 = ~bool_var1
print (bool_var1)
print (bool_var2)
```

```
True
-2
```

To perform a “logical not” operation on Python Boolean variables, we have to use the “not” keyword operator.

```python
bool_var1 = True
bool_var2 = not bool_var1
print (bool_var1)
print (bool_var2)
```

```
True
False
```

Multiple Linear Regression

Sample regression line formula we saw in simple linear regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{or} \quad \hat{y} = b_0 + b_1 x$$

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ \dots$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ \dots + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \end{bmatrix}$$

Predictor variables

Target variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244