Linear Regression and Correlation

MAT 152 – Statistical Methods I

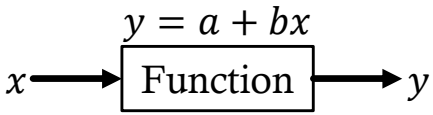
Lecture 1

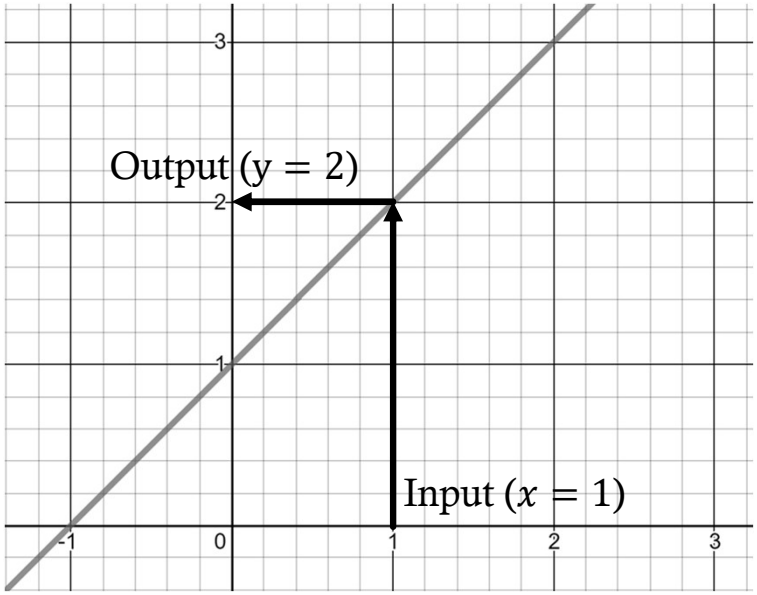
Instructor: Dustin Roten

Fall 2020

Reviewing Linear Equations

- Linear equations are **functions** that represent straight lines. There are inputs (independent variables, x) and outputs (dependent variables, y).
- Linear equations contain two pieces of information:
  - The "slope" or "rate of change"
  - The "y-intercept"
- Linear equations can be written as $y = a + bx$
  - a represents the y-intercept
  - b represents the slope





Example

Suzy wishes to open a small bicycle shop where she plans to build and sell custom bikes. She wants to officially open once she has 15 bikes in her inventory. It takes her two days to build one bicycle. She already has three constructed. If Suzy works at a constant rate, how many days will pass before she can open her shop?

What we know:

The independent variable is time (in days) [x]

The dependent variable is the number of bikes [y]

$a = 3$ (The number of bikes already constructed)

$b = \frac{1 \text{ bike}}{2 \text{ days}} = 0.5 \frac{\text{bikes}}{\text{day}}$ (The rate of change)

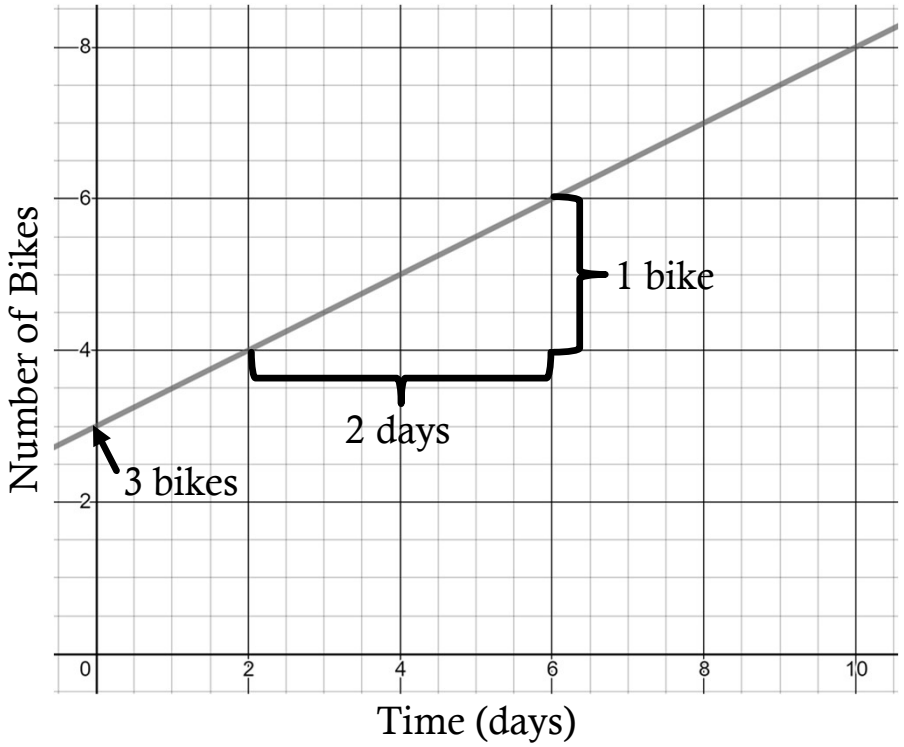
$y = 3 + 0.5x$

Time (in days) x	**Number of Bikes** y
0	3
1	3.5
2	4
3	4.5
4	5
5	5.5
6	6

Example (Cont.)

$y = 3 + 0.5x$

Time (in days) x	Number of Bikes y
0	3
1	3.5
2	4
3	4.5
4	5
5	5.5
6	6



Example (Cont.)

Suzy wishes to open a small bicycle shop where she plans to build and sell custom bikes. She wants to officially open once she has 15 bikes in her inventory. It takes her two days to build one bicycle. She already has three constructed. If Suzy works at a constant rate, how many days will pass before she can open her shop?

What time (in days) makes the following statement true?

$$15 = 3 + 0.5x$$

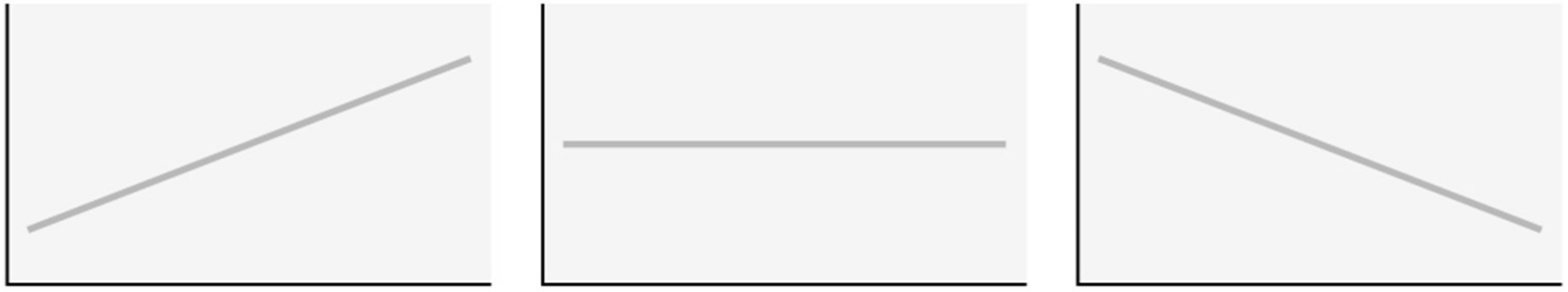
solve for x.

$$x = \frac{15 - 3}{0.5} = 24$$

Suzy can open after 24 days.

Slopes

- Slopes can be positive, zero, or negative.



(a) Positive Slope

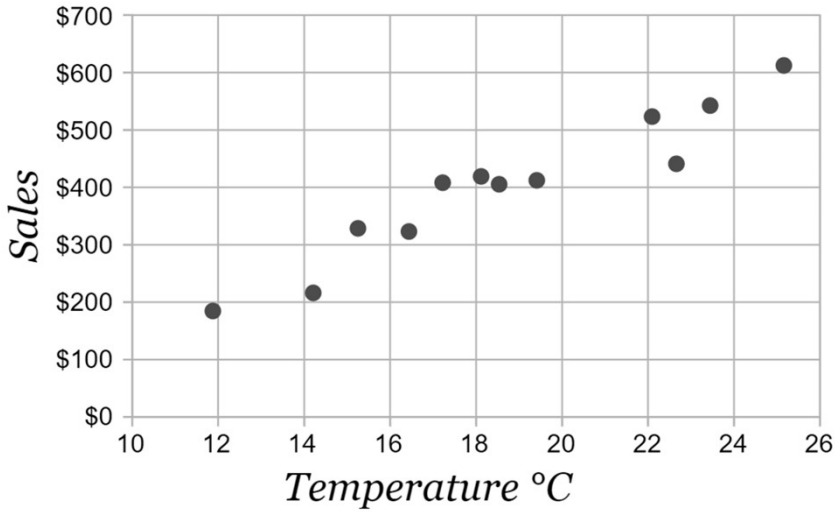
(b) Zero Slope

(c) Negative Slope

Scatter Plots

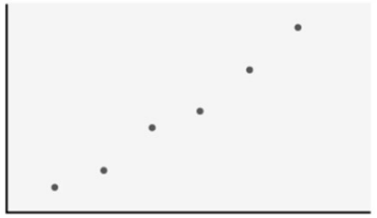
- Scatter plots show the direction of a relationship in **bivariate** data. (Bivariate means two variables.)
- Consider the relationship between outdoor temperature and ice cream sales (plotted to the right).
- It is reasonable to assume that ice cream sales increases when it is hotter outside.

Scatter Plot of Ice Cream Sales vs Temperature

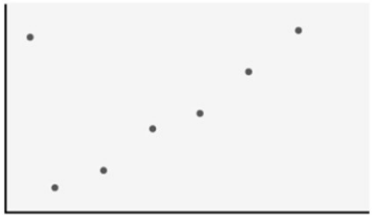


Scatter Plots

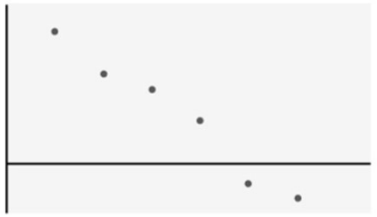
- A scatter plot shows the **direction** of a relationship between the variables. A clear direction happens when there is one of the following:
  - High values of one variable occur with high values of the other variable.
  - Low values of one variable occur with low values of the other variable.
  - High values of one variable occur with low values of the other variable.
- The more visible the direction in a scatter plot, the stronger the relationship between the variables.
- NOTE: Not all relationships are linear!



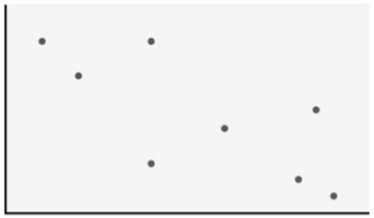
(a) Positive linear pattern (strong)



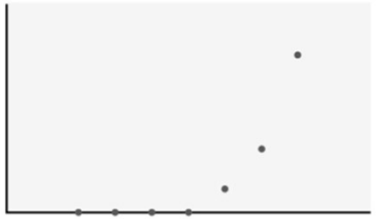
(b) Linear pattern w/ one deviation



(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)

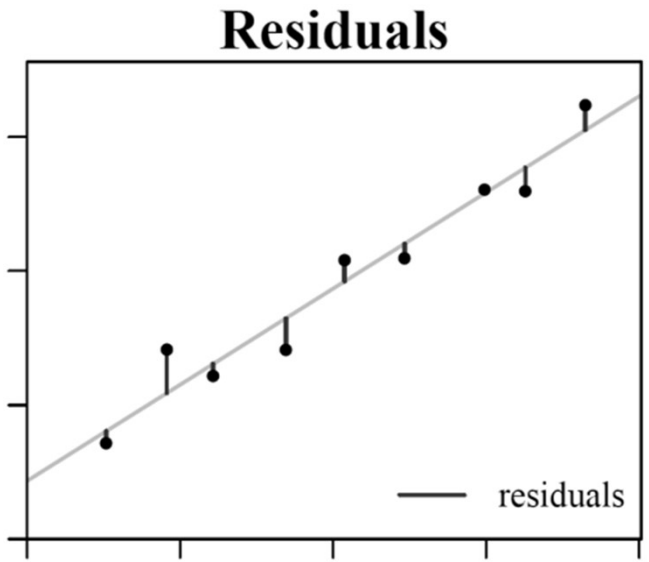


(a) Exponential growth pattern

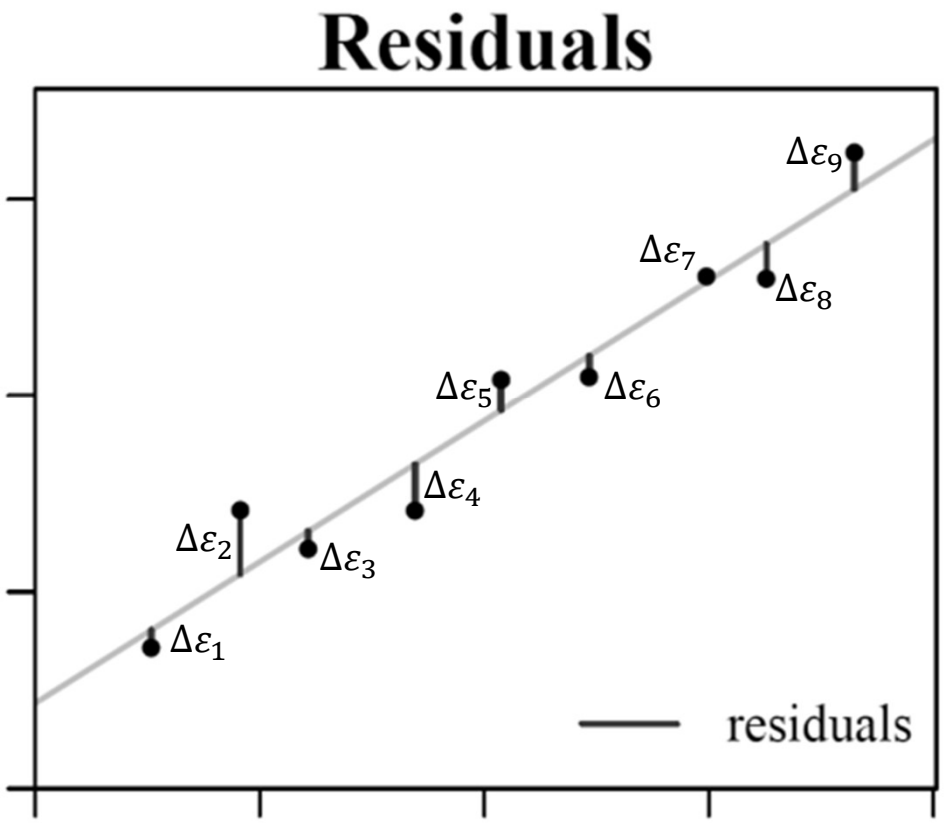
(b) No pattern

Linear Relationships

- If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line is calculated by the **linear regression** process.
- This process constructs the "line of best fit".
- The linear regression process minimizes the distances from each point to the line of best fit.



Linear Relationships



$\Delta\varepsilon_2, \Delta\varepsilon_6, \Delta\varepsilon_9$ are positive residuals

$\Delta\varepsilon_1, \Delta\varepsilon_3, \Delta\varepsilon_4, \Delta\varepsilon_6, \Delta\varepsilon_8$ are negative residuals

$\Delta\varepsilon_7$ is zero

Since some residuals are negative and some are positive, adding them up may result in zero. So, these values are first squared before adding:

$$(\Delta\varepsilon_1)^2 + \cdots + (\Delta\varepsilon_9)^2 = \sum_{i=1}^{9} (\Delta\varepsilon_i)^2$$

This is the **sum of squared errors** which must be minimized by the linear regression process.

Least-Squares Regression Line

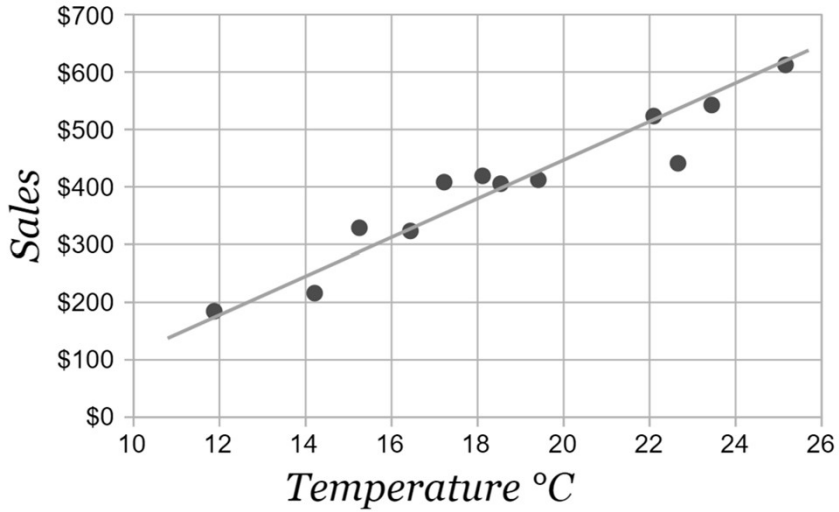
- Thus, the least-squares regression line has the form:

$$\hat{y} = a + bx$$

- $a = \bar{y} - b\bar{x}$
- $b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = r\left(\frac{s_y}{s_x}\right)$

- a is still the y-intercept
- b is still the slope (rate of change)
- The line of best fit goes through $(\bar{x}, \bar{y})$.
- r is the correlation coefficient.

Scatter Plot of Ice Cream Sales vs Temperature



Example

A small clothing store wanted to investigate the influence that rainfall had on their umbrella sales. To do this, daily rainfall was recorded along with daily umbrella sales. The data are presented in the plot below.

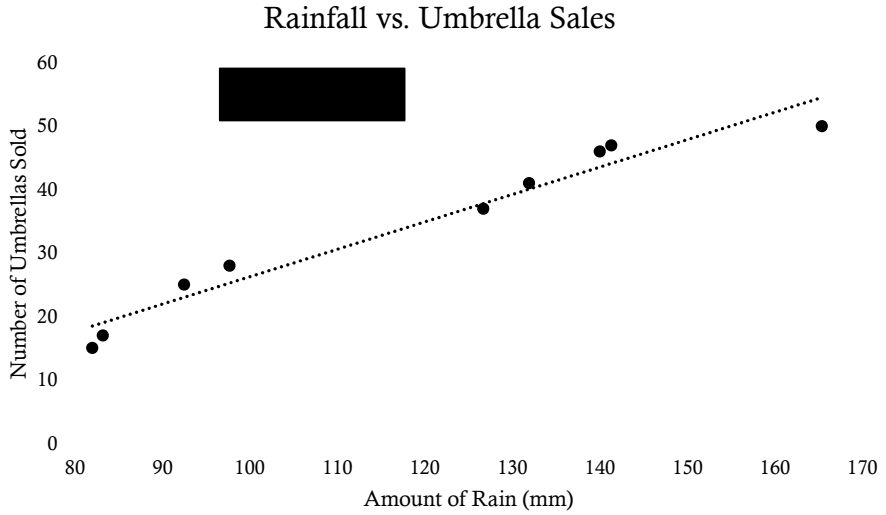
Independent variable: Rainfall

Dependent variable: Umbrella sales

$\hat{y} = a + bx$

$\hat{y} = -16.969 + 0.4325x$

0.4325 umbrellas are sold per 1mm of rain.



Example (Cont.)

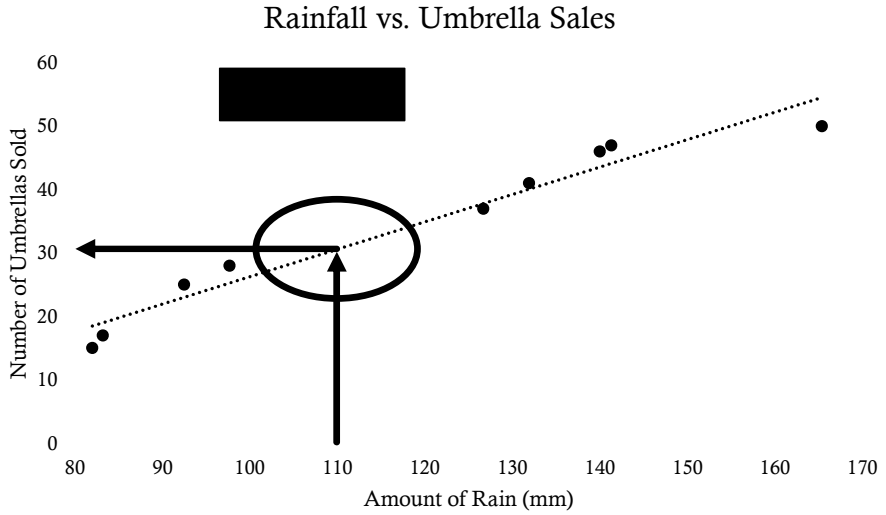
How many umbrellas are expected to sell if it rains 110mm?

$\hat{y} = -16.969 + 0.4325x$

$\hat{y} = -16.969 + 0.4325 \cdot 110 = 30.6$

Roughly 31 umbrellas will be sold if it rains 110mm

Making inferences BETWEEN data points is called **interpolation**.



Example (Cont.)

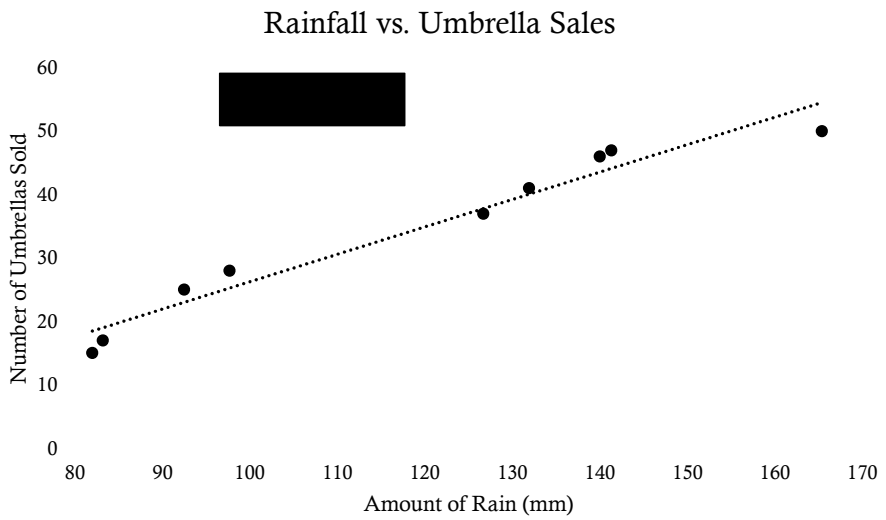
How many umbrellas are expected to sell if it rains 180mm?

$\hat{y} = -16.969 + 0.4325x$

$\hat{y} = -16.969 + 0.4325 \cdot 180 = 60.881$

Roughly 61 umbrellas will be sold if it rains 180mm

Making inferences BEYOND the data points is called **extrapolation**. DO NOT EXTRAPOLTE TOO FAR AWAY FROM THE DATA!



A Quick Review

- Scatter plots show trends in bivariate data.
- The linear regression equation has the form $\hat{y} = a + bx$.
  - a is the y-intercept
  - b is the slope or rate of change
- The linear regression process places the least-squares regression line as close as possible to all data points.
- This line of best fit passes through the point $(\bar{x}, \bar{y})$.
- Non-linear relationships exist!