Introduction to Linear Models

Prof. Wells

STA 209, 2/15/23

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Investigate the linear model
- Discuss predictions and residuals
- Explore a formula for finding the line of best fit

Section 1

Introduction to Linear Regression

Overview

"All models are wrong, but some are useful."

— George Box, statistician

Overview

"All models are wrong, but some are useful."

— George Box, statistician

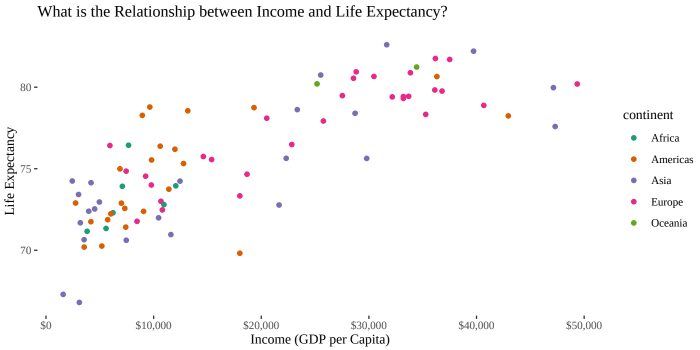
- Linear regression is both an accessible and potent tool in statistical analysis.

Overview

"All models are wrong, but some are useful."

— George Box, statistician

- Linear regression is both an accessible and potent tool in statistical analysis.

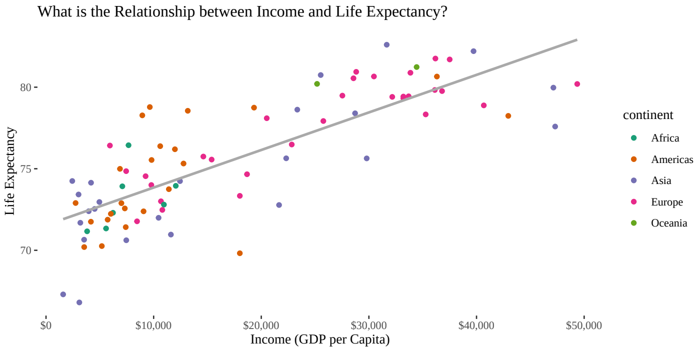


Overview

"All models are wrong, but some are useful."

— George Box, statistician

- Linear regression is both an accessible and potent tool in statistical analysis.



Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

$$Y = \beta_0 + \beta_1 X \qquad \text{with } \beta_0, \beta_1 \text{ fixed constants}$$

Relationships for Quantitative Variables

- Quantitative variables, by nature, are amenable to algebraic manipulation.
- Given two quantitative variables X and Y, we construct a mathematical model that expresses the values of Y as a function of the values of X:

$$Y = f(X)$$

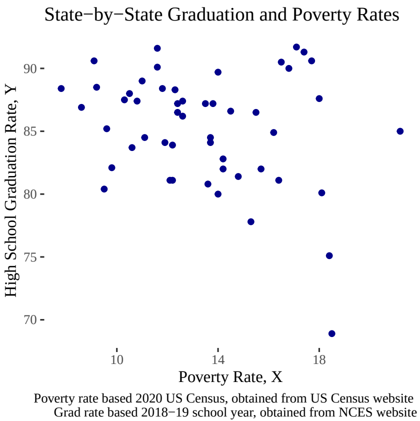
- Linear functions are the simplest of all mathematical functions, and so are the starting place for modeling

$$Y = \beta_0 + \beta_1 X \qquad \text{with } \beta_0, \beta_1 \text{ fixed constants}$$

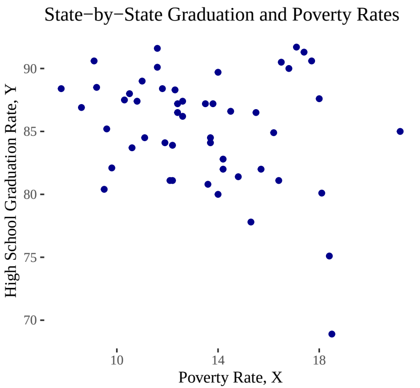
- Of course, in the wild, the observed values of Y will **not** be perfectly predicted by the values of X.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \text{where } \epsilon \text{ is the error}$$

Scatterplots and Linear Relationships I



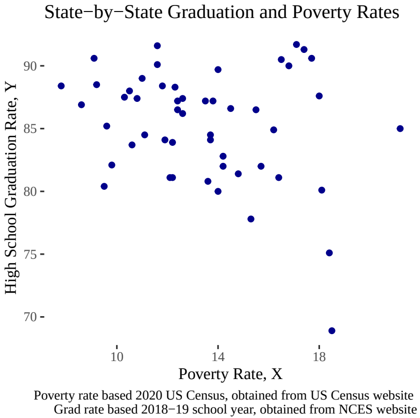
Scatterplots and Linear Relationships I



Poverty rate based 2020 US Census, obtained from US Census website
Grad rate based 2018−19 school year, obtained from NCES website

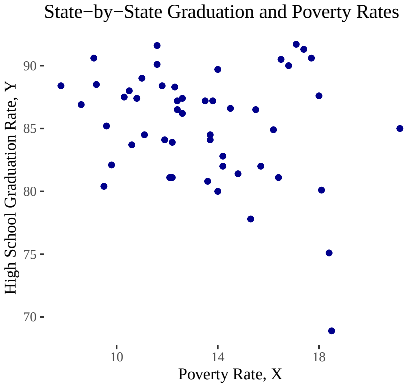
- Explanatory Variable:

Scatterplots and Linear Relationships I



- Explanatory Variable:
  - Poverty Rate (X)

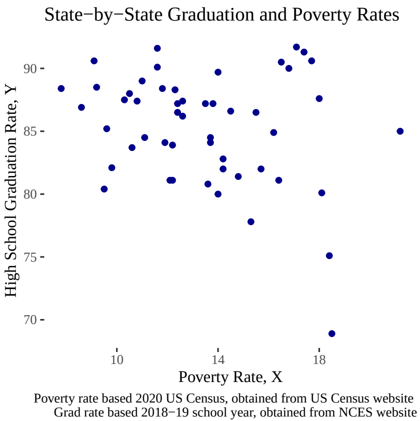
Scatterplots and Linear Relationships I



Poverty rate based 2020 US Census, obtained from US Census website
Grad rate based 2018−19 school year, obtained from NCES website

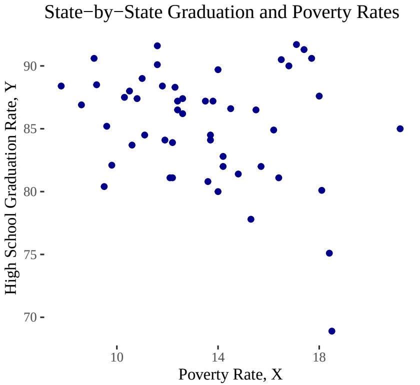
- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:

Scatterplots and Linear Relationships I



- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)

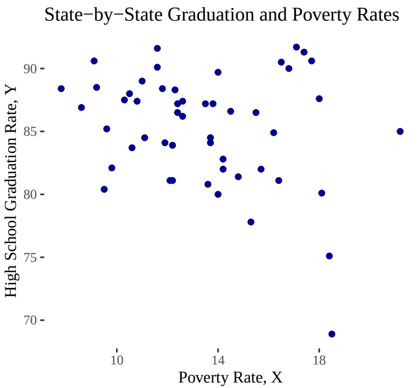
Scatterplots and Linear Relationships I



Poverty rate based 2020 US Census, obtained from US Census website
Grad rate based 2018−19 school year, obtained from NCES website

- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)
- Relationship:

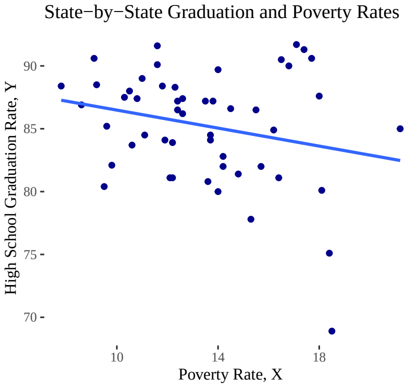
Scatterplots and Linear Relationships I



Poverty rate based 2020 US Census, obtained from US Census website
Grad rate based 2018−19 school year, obtained from NCES website

- Explanatory Variable:
  - Poverty Rate (X)
- Response Variable:
  - High School Graduation Rate (Y)
- Relationship:
  - Linear, negative, moderate

Scatterplots and Linear Relationships II

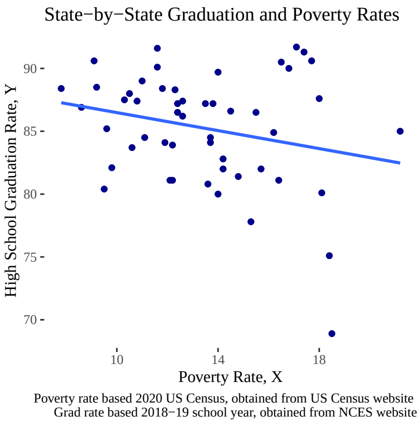


Poverty rate based 2020 US Census, obtained from US Census website
Grad rate based 2018−19 school year, obtained from NCES website

- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 90 - 0.4X$$

Scatterplots and Linear Relationships II

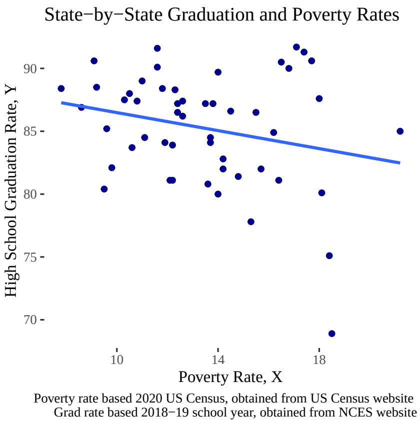


- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 90 - 0.4X$$

- $\hat{Y}$ indicates this is an estimate of Y

Scatterplots and Linear Relationships II

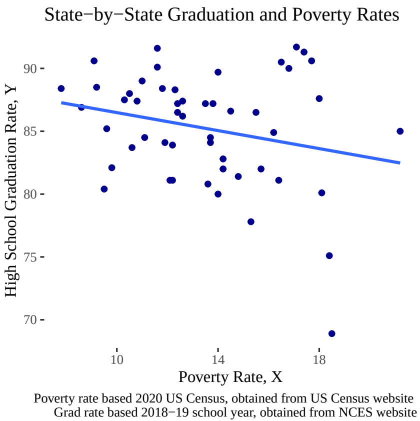


- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 90 - 0.4X$$

- $\hat{Y}$ indicates this is an estimate of Y
- **Slope** of $\beta_1 = -0.4$ means every 1 unit increase in Poverty corresponds to a 0.4 unit decrease on average in Graduation.

Scatterplots and Linear Relationships II

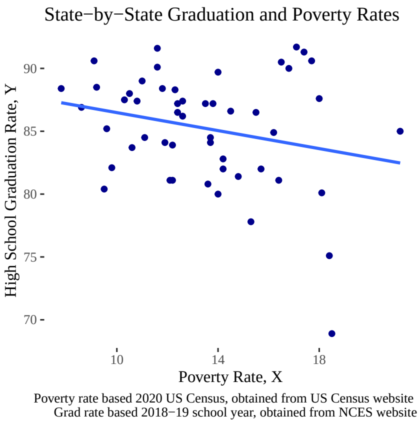


- Model (hand-fitted):

$$\hat{Y} = \beta_0 + \beta_1 X = 90 - 0.4X$$

- $\hat{Y}$ indicates this is an estimate of Y
- **Slope** of $\beta_1 = -0.4$ means every 1 unit increase in Poverty corresponds to a 0.4 unit decrease on average in Graduation.
- **Intercept** of $\beta_0 = 90$ means model predicts graduation rate of 90% when poverty rate is 0%.

Scatterplots and Linear Relationships III

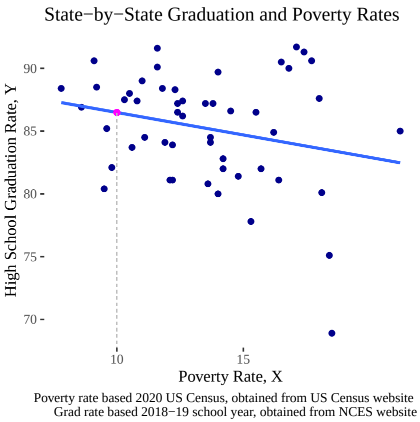


- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 10%?

Scatterplots and Linear Relationships III

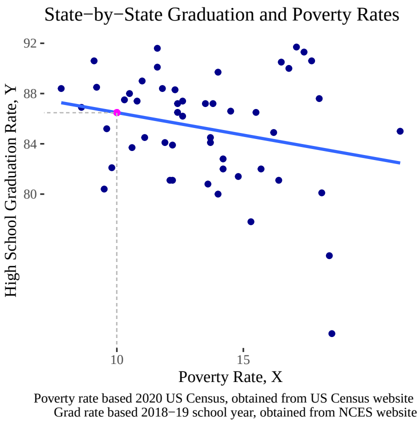


- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 10%?

Scatterplots and Linear Relationships III



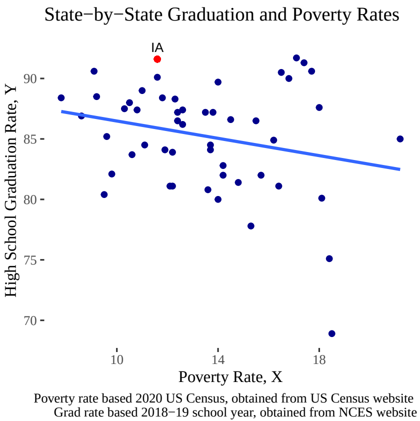
- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- What does the model predict to be the graduation rate for a state with theoretical poverty rate 7%?

$$\hat{Y} = 90 - 0.4 \cdot 10 = 86$$

Scatterplots and Linear Relationships IV



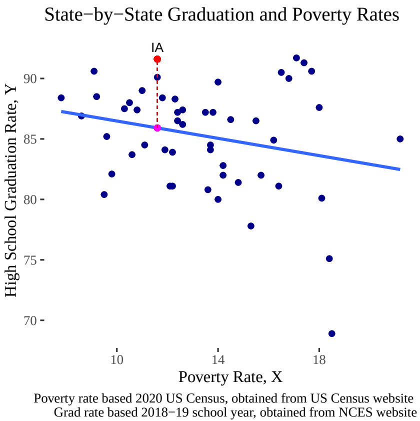
- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- Iowa has a poverty rate of 11.6. What does the model predict is Iowa's graduation rate?

$$\hat{Y} = 90 - 0.4 \cdot 11.6 = 85.36$$

Scatterplots and Linear Relationships IV



- Model:

$$\hat{Y} = 90 - 0.4 \cdot X$$

- Iowa has a poverty rate of 11.6. What does the model predict is Iowa's graduation rate?

$$\hat{Y} = 90 - 0.4 \cdot 11.6 = 85.36$$

But Iowa's actual graduation rate is 91.6

Residuals

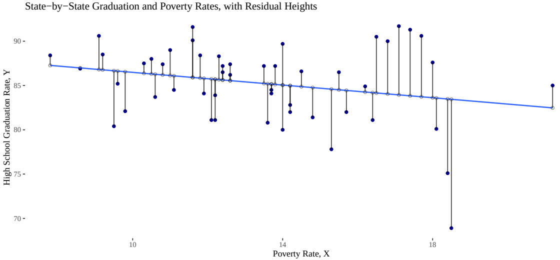
- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

$$e_i = Y_i - \hat{Y}_i$$

Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

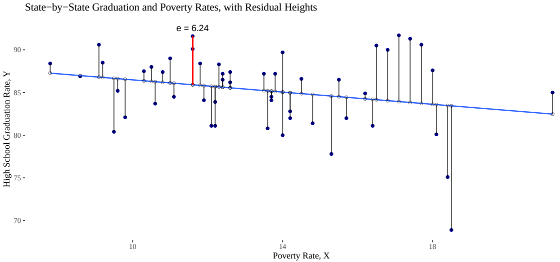
$$e_i = Y_i - \hat{Y}_i$$



Residuals

- **Residuals** are the leftover variation in the data after accounting for model fit.
- Each observation (X_i, Y_i) has its own residual e_i, which is the difference between the observed (Y_i) and predicted $(\hat{Y}_i)$ value:

$$e_i = Y_i - \hat{Y}_i$$

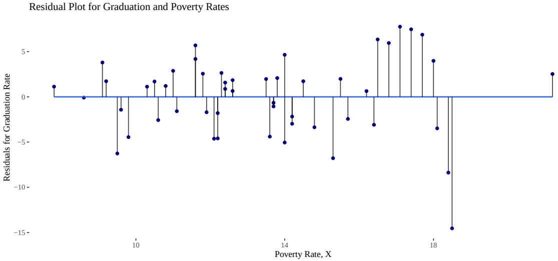


- Iowa's residual is

$$e = Y - \hat{Y} = 91.6 - 85.36 = 6.24$$

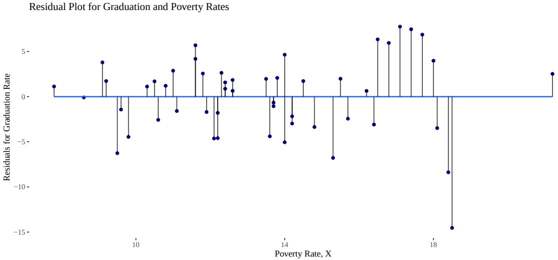
Residual Plot

- To visualize the degree of accuracy of a linear model, we use residual plots:



Residual Plot

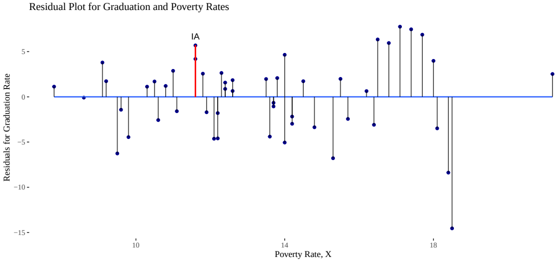
- To visualize the degree of accuracy of a linear model, we use residual plots:



- Points preserve original x-position, but with y-position equal to residual.

Residual Plot

- To visualize the degree of accuracy of a linear model, we use residual plots:



- Points preserve original x-position, but with y-position equal to residual.

Section 2

Quantifying Goodness-of-Fit

Correlation Coefficient

- The **Correlation Coefficient** R describes the strength of a *linear* relationship between two quantitative variables, and is always a number between -1 and 1.

Correlation Coefficient

- The **Correlation Coefficient** R describes the strength of a *linear* relationship between two quantitative variables, and is always a number between -1 and 1.
- The *sign* of R indicates the direction of relationship, while the *magnitude* of R indicates the strength.

Correlation Coefficient

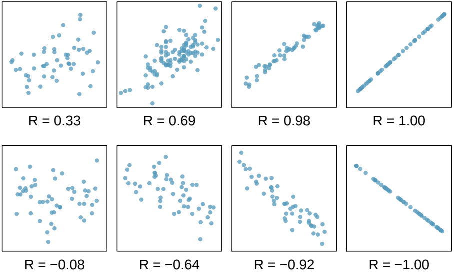
- The **Correlation Coefficient** R describes the strength of a *linear* relationship between two quantitative variables, and is always a number between -1 and 1.
- The *sign* of R indicates the direction of relationship, while the *magnitude* of R indicates the strength.

If $\|R\|$ is between...	Then the linear relationship is...
0.7 and 1	strong
0.3 and .07	moderate
0 and 0.3	weak

Correlation Coefficient

- The **Correlation Coefficient** R describes the strength of a *linear* relationship between two quantitative variables, and is always a number between -1 and 1.
- The *sign* of R indicates the direction of relationship, while the *magnitude* of R indicates the strength.

If $\lvert R \rvert$ is between...	Then the linear relationship is...
0.7 and 1	strong
0.3 and .07	moderate
0 and 0.3	weak



Mathematical Definition of Correlation

- The Correlation Coefficient is defined via formula using the means ($\bar{x}$, $\bar{y}$) and standard deviations (s_x, s_y) of the variables X and Y:

Mathematical Definition of Correlation

- The Correlation Coefficient is defined via formula using the means ($\bar{x}$, $\bar{y}$) and standard deviations (s_x, s_y) of the variables X and Y:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Mathematical Definition of Correlation

- The Correlation Coefficient is defined via formula using the means ($\bar{x}$, $\bar{y}$) and standard deviations (s_x, s_y) of the variables X and Y:

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Suppose that when X is above its mean, then Y also tends to be above its mean. And similarly, when X is below its mean, then Y is also.

Mathematical Definition of Correlation

- The Correlation Coefficient is defined via formula using the means ($\bar{x}$, $\bar{y}$) and standard deviations (s_x, s_y) of the variables X and Y:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

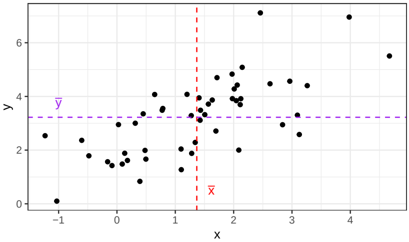
- Suppose that when X is above its mean, then Y also tends to be above its mean. And similarly, when X is below its mean, then Y is also.
  - What will the scatterplot of X and Y look like?

Mathematical Definition of Correlation

- The Correlation Coefficient is defined via formula using the means ($\bar{x}$, $\bar{y}$) and standard deviations (s_x, s_y) of the variables X and Y:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Suppose that when X is above its mean, then Y also tends to be above its mean. And similarly, when X is below its mean, then Y is also.
  - What will the scatterplot of X and Y look like?

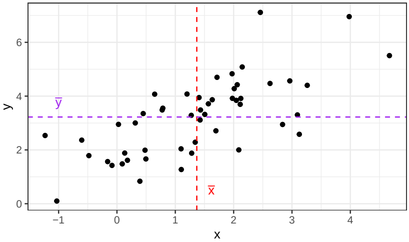


Mathematical Definition of Correlation

- The Correlation Coefficient is defined via formula using the means $(\bar{x}, \bar{y})$ and standard deviations (s_x, s_y) of the variables X and Y:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Suppose that when X is above its mean, then Y also tends to be above its mean. And similarly, when X is below its mean, then Y is also.
  - What will the scatterplot of X and Y look like?



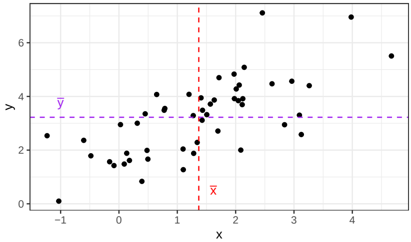
- Based on the formula, is correlation positive or negative?

Mathematical Definition of Correlation

- The Correlation Coefficient is defined via formula using the means $(\bar{x}, \bar{y})$ and standard deviations (s_x, s_y) of the variables X and Y:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Suppose that when X is above its mean, then Y also tends to be above its mean. And similarly, when X is below its mean, then Y is also.
  - What will the scatterplot of X and Y look like?



- Based on the formula, is correlation positive or negative?

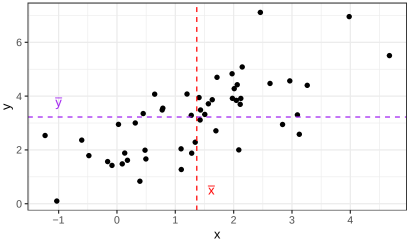
$$R = \sum(\text{Pos.})(\text{Pos.}) + \sum(\text{Neg.})(\text{Neg.})$$

Mathematical Definition of Correlation

- The Correlation Coefficient is defined via formula using the means $(\bar{x}, \bar{y})$ and standard deviations (s_x, s_y) of the variables X and Y:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Suppose that when X is above its mean, then Y also tends to be above its mean. And similarly, when X is below its mean, then Y is also.
    - What will the scatterplot of X and Y look like?



- Based on the formula, is correlation positive or negative?

$$R = \sum(\text{Pos.})(\text{Pos.}) + \sum(\text{Neg.})(\text{Neg.})$$

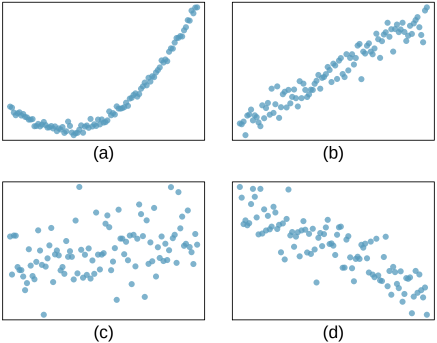
$$R = 0.6995848$$

Correlation is not Association

- Correlation measures strength of *LINEAR* relationship:

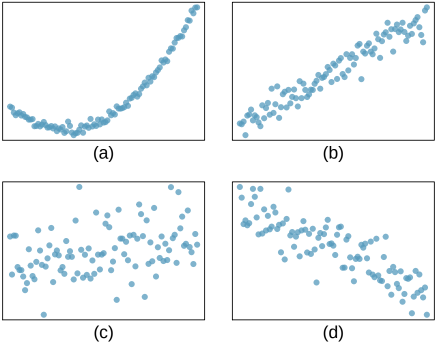
Correlation is not Association

- Correlation measures strength of *LINEAR* relationship:
- Which of the following has the strongest correlation (largest value of $|R|$)?



Correlation is not Association

- Correlation measures strength of *LINEAR* relationship:
- Which of the following has the strongest correlation (largest value of $|R|$)?



- Answer: (b), not (a)

Correlation isn't the Whole Story

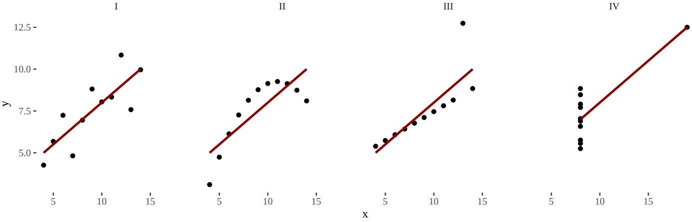
- Computing a correlation coefficient is no substitute for data visualization.

Correlation isn't the Whole Story

- Computing a correlation coefficient is no substitute for data visualization.
- All of the following have identical, strong positive correlation ($R = 0.82$):

Correlation isn't the Whole Story

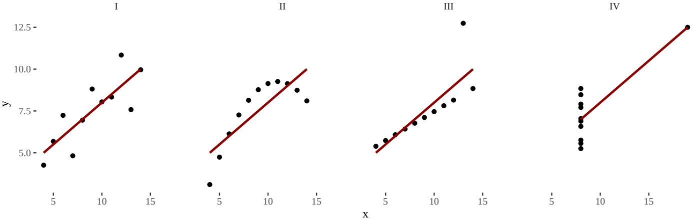
- Computing a correlation coefficient is no substitute for data visualization.
- All of the following have identical, strong positive correlation ($R = 0.82$):




```
##                I   II  III   IV
## Correlation 0.82 0.82 0.82 0.82
```

Correlation isn't the Whole Story

- Computing a correlation coefficient is no substitute for data visualization.
- All of the following have identical, strong positive correlation ($R = 0.82$):




```
##                I   II  III   IV
## Correlation 0.82 0.82 0.82 0.82
```

- However, each graphic tells a radically different story about the relationship between the variables.

Section 3

Fitting a Line by Least-Squares Regression

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values
  $$|e_1| + |e_2| + \cdots + |e_n|$$
  - Option 2: Minimize the sum of squares
  $$e_1^2 + e_2^2 + \cdots + e_n^2$$
- Option 2 is usually preferred.
  1. Most commonly used.

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.
  1. Most commonly used.
  2. More computationally efficient.

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.
  1. Most commonly used.
  2. More computationally efficient.
  3. Has theoretical advantages (by analogy with distance and pythagorean thm.)

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.
  1. Most commonly used.
  2. More computationally efficient.
  3. Has theoretical advantages (by analogy with distance and pythagorean thm.)
  4. Appropriately weights one large residuals as "worse" than many small ones.

Measure for **BEST** Line

- The line of best fit to a scatterplot should minimize residuals, meaning:
  - Option 1: Minimizing the sum of absolute values

$$|e_1| + |e_2| + \cdots + |e_n|$$

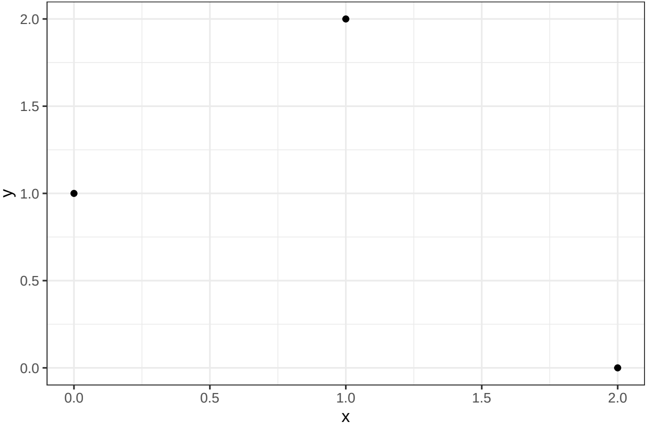
  - Option 2: Minimize the sum of squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Option 2 is usually preferred.
  1. Most commonly used.
  2. More computationally efficient.
  3. Has theoretical advantages (by analogy with distance and pythagorean thm.)
  4. Appropriately weights one large residuals as "worse" than many small ones.
  5. Has well-understood properties for inference

Line of Best Fit

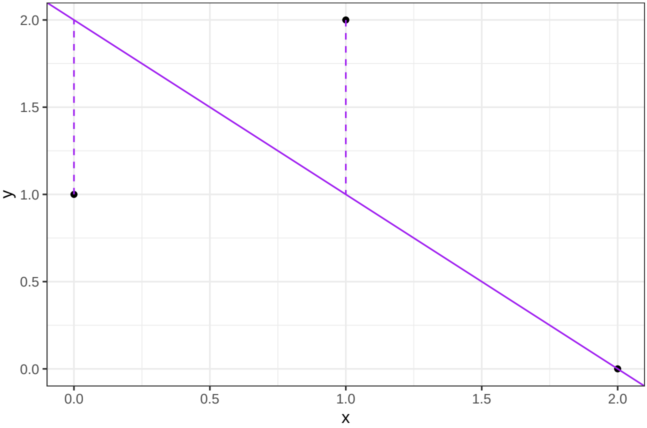
- For the three data points below, consider candidates for the line of best fit:



$$\text{Goal: minimize } e_1^2 + e_2^2 + e_3^2$$

Line of Best Fit

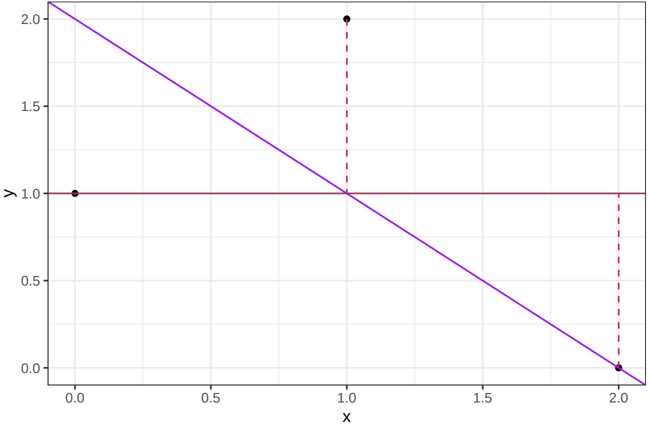
- For the three data points below, consider candidates for the line of best fit:



$$\text{Purple line}: \quad e_1^2 + e_2^2 + e_3^2 = 1^2 + 0^2 + 1^2 = 2$$

Line of Best Fit

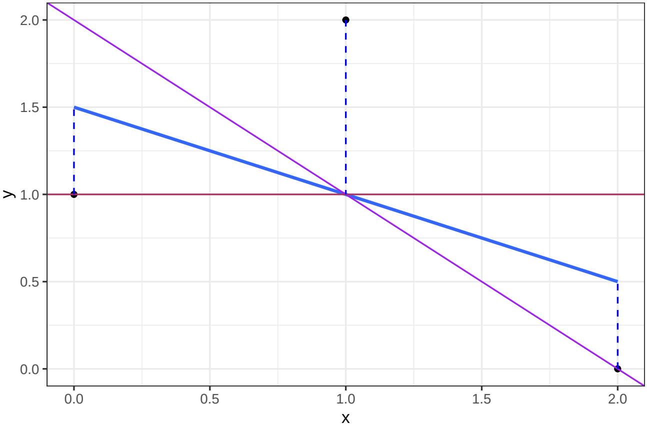
- For the three data points below, consider candidates for the line of best fit:



$$\text{Maroon line}: \quad e_1^2 + e_2^2 + e_3^2 = 0^2 + 1^2 + 1^2 = 2$$

Line of Best Fit

- For the three data points below, consider candidates for the line of best fit:



$$\text{Blue line}: \quad e_1^2 + e_2^2 + e_3^2 = 0.5^2 + 1^2 + 0.5^2 = 1.5$$

A Formula for the Least Squares Regression Line

- Suppose n observations for variables X and Y are collected:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

with means $\bar{x}, \bar{y}$, standard deviations s_x, s_y, and correlation R.

A Formula for the Least Squares Regression Line

- Suppose n observations for variables X and Y are collected:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

with means $\bar{x}, \bar{y}$, standard deviations s_x, s_y, and correlation R.

- The **Least Squares Regression Line** modeling Y as a function of X is

$$\hat{Y} = \beta_0 + \beta_1 X$$

A Formula for the Least Squares Regression Line

- Suppose n observations for variables X and Y are collected:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

with means $\bar{x}, \bar{y}$, standard deviations s_x, s_y, and correlation R.

- The **Least Squares Regression Line** modeling Y as a function of X is

$$\hat{Y} = \beta_0 + \beta_1 X$$

where the slope β_1 is given by

$$\beta_1 = \frac{s_y}{s_x} R$$

A Formula for the Least Squares Regression Line

- Suppose n observations for variables X and Y are collected:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

  with means $\bar{x}, \bar{y}$, standard deviations s_x, s_y, and correlation R.

- The **Least Squares Regression Line** modeling Y as a function of X is

$$\hat{Y} = \beta_0 + \beta_1 X$$

  where the slope β_1 is given by

$$\beta_1 = \frac{s_y}{s_x} R$$

  and where the intercept is given by

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

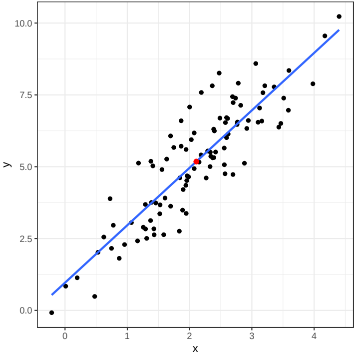
- The line *always* goes through the point $(\bar{x}, \bar{y})$

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

- The line *always* goes through the point $(\bar{x}, \bar{y})$




```
##     mean_x   mean_y
## 1 2.108887 5.179967
```

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

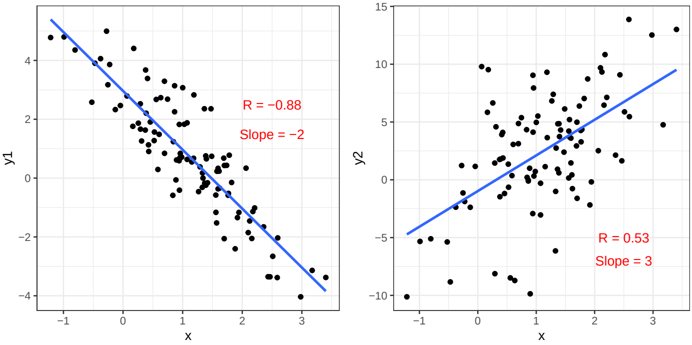
- The slope has the same sign as the correlation coefficient

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

- The slope has the same sign as the correlation coefficient



Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

- The slope is close to 0 when either $R \approx 0$ *or* when s_x is much bigger than s_y

Properties of the Least-Squares Regression line

- The least squares line is

$$\hat{Y} = \beta_0 + \beta_1 X \qquad \beta_1 = \frac{s_y}{s_x} R \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

- The slope is close to 0 when either $R \approx 0$ *or* when s_x is much bigger than s_y