# Simple linear regression

Chrysafis Vogiatzis

Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Lecture 30

ILLINOIS
ISE | Industrial & Enterprise Systems Engineering
GRAINGER COLLEGE OF ENGINEERING

# Welcome to Part 4

The three classifications of modern statistical methods:

1. **Descriptive statistics**: techniques to describe, visualize, and present information and data.
2. **Inferential statistics**: techniques to draw conclusions for a large, unknown population based on observations from a smaller group (sample).
3. **Model building**: techniques to identify relationships between data points (when those exist) and build models that can make predictions about the future.

In this last part of the class, we will focus on model building.

# Welcome to Part 4

The three classifications of modern statistical methods:

1. **Descriptive statistics**: techniques to describe, visualize, and present information and data.
2. **Inferential statistics**: techniques to draw conclusions for a large, unknown population based on observations from a smaller group (sample).
3. **Model building**: techniques to identify relationships between data points (when those exist) and build models that can make predictions about the future.

In this last part of the class, we will focus on model building.

ILLINOIS

# Welcome to Part 4

The three classifications of modern statistical methods:

1. **Descriptive statistics**: techniques to describe, visualize, and present information and data.
2. **Inferential statistics**: techniques to draw conclusions for a large, unknown population based on observations from a smaller group (sample).
3. **Model building**: techniques to identify relationships between data points (when those exist) and build models that can make predictions about the future.

In this last part of the class, we will focus on model building.

ILLINOIS

# Welcome to Part 4

The three classifications of modern statistical methods:

1. **Descriptive statistics**: techniques to describe, visualize, and present information and data.
2. **Inferential statistics**: techniques to draw conclusions for a large, unknown population based on observations from a smaller group (sample).
3. **Model building**: techniques to identify relationships between data points (when those exist) and build models that can make predictions about the future.

In this last part of the class, we will focus on model building.

ILLINOIS

# Descriptive statistics

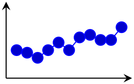
Sample

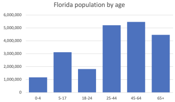
$X_1, X_2, \ldots, X_n$

$\Longrightarrow$

- Summary statistics:
  - mean
  - variance
  - median
  - mode
  - percentiles

$\Longrightarrow$





# Inferential statistics

From sample:

$X_1, X_2, \ldots, X_n$

Infer

$\Longrightarrow$

To a population:

$\mu, \sigma, p$

Point estimation:

$\hat{\Theta}$

- bias.
- variance.
- MSE.

Interval estimation:

*Confidence interval*

- unknown mean, var., proportion.
- confidence level $1 - \alpha$.

Hypothesis test:

$H_0$ or $H_1$

- one or two populations.
- mean, variance, proportion.
- $\alpha, \beta$, $P$-values.

ILLINOIS

# Model building

- Goal #1: investigate whether a **relationship** exists between the variables of our model.
- Goal #2: measure how **strong** that relationship is.
- Goal #3: **predict** future responses given information.

1. **Regression.**
   - For continuous outcomes $y$.
   - Given a variable $x$, predict the value of variable $y$.




2. **Classification.**
   - For discrete outcomes $y$.
   - Given a variable $x$, predict where $y$ belongs to.

# Model building

- Goal #1: investigate whether a **relationship** exists between the variables of our model.
- Goal #2: measure how **strong** that relationship is.
- Goal #3: **predict** future responses given information.

1. **Regression.**
   - For continuous outcomes $y$.
   - Given a variable $x$, predict the value of variable $y$.

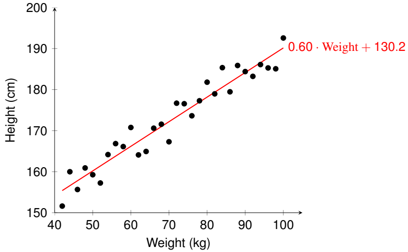
2. **Classification.**
   - For discrete outcomes $y$.
   - Given a variable $x$, predict where $y$ belongs to.

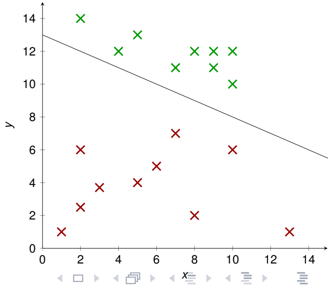
# Model building

- Goal #1: investigate whether a **relationship** exists between the variables of our model.
- Goal #2: measure how **strong** that relationship is.
- Goal #3: **predict** future responses given information.

1. **Regression.**
   - For continuous outcomes $y$.
   - Given a variable $x$, predict the value of variable $y$.

2. **Classification.**
   - For discrete outcomes $y$.
   - Given a variable $x$, predict where $y$ belongs to.



# Regression

**Notation**

- *independent variables* $x_j$, $j = 1, \ldots, k$. predictors
- *dependent variable* $y$. response
- *data* $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$, $i = 1, \ldots, n$.
- *goal*: $\hat{y} = f(x_1, \ldots, x_k)$.

**Linear regression**:

- $f$ is linear:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

**Simple linear regression**:

- $k = 1$ – only one independent variable $x$:

$$\hat{y} = \beta_0 + \beta_1 x.$$

# Regression

**Notation**

- *independent variables* $x_j$, $j = 1, \ldots, k$. predictors
- *dependent variable* $y$. response
- *data* $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$, $i = 1, \ldots, n$.
- *goal*: $\hat{y} = f(x_1, \ldots, x_k)$.

**Linear regression**:

- $f$ is linear:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

**Simple linear regression**:

- $k = 1$ – only one independent variable $x$:

$$\hat{y} = \beta_0 + \beta_1 x.$$

ILLINOIS

# Regression

**Notation**

- *independent variables* $x_j$, $j = 1, \ldots, k$. predictors
- *dependent variable* $y$. response
- *data* $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$, $i = 1, \ldots, n$.
- *goal*: $\hat{y} = f\,(x_1, \ldots, x_k)$.

**Linear regression**:

- $f$ is linear:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

**Simple linear regression**:

- $k = 1$ – only one independent variable $x$:

$$\hat{y} = \beta_0 + \beta_1 x.$$

ILLINOIS

# Motivating example

## Example

*A webstore has collected the following data on the weekly visitors of the website and the profits from the past 20 weeks. They want to investigate that relationship and see whether they can direct more clicks towards their store. The data they have collected is as follows:*

| n | Visitors | Profit | n | Visitors | Profit |
|---|---|---|---|---|---|
| 1 | 907 | 11.2 | 2 | 926 | 11.05 |
| 3 | 506 | 6.84 | 4 | 741 | 9.21 |
| 5 | 789 | 9.42 | 6 | 889 | 10.08 |
| 7 | 874 | 9.45 | 8 | 510 | 6.73 |
| 9 | 529 | 7.24 | 10 | 420 | 6.12 |
| 11 | 679 | 7.63 | 12 | 872 | 9.43 |
| 13 | 924 | 9.46 | 14 | 607 | 7.64 |
| 15 | 452 | 6.92 | 16 | 729 | 8.95 |
| 17 | 794 | 9.33 | 18 | 844 | 10.23 |
| 19 | 1010 | 11.77 | 20 | 621 | 7.41 |

# Motivating example

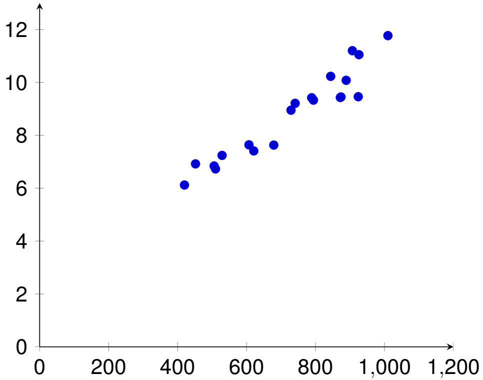
The table was hard to read. So here is the same data in a slightly more visual format..




1. Do you see a relationship between profits and visits?
2. Is the relationship linear?
3. Is the relationship strong?
4. Can I predict profits based on # visitors?

# Motivating example

The table was hard to read. So here is the same data in a slightly more visual format..

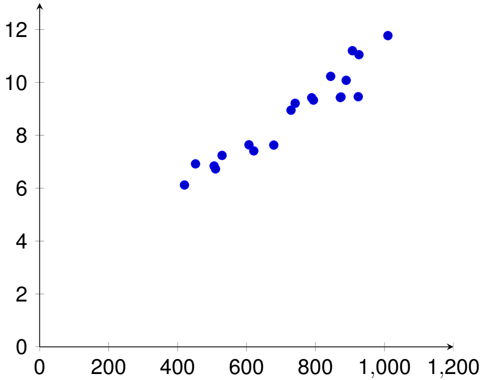



1. Do you see a relationship between profits and visits?
2. Is the relationship linear?
3. Is the relationship strong?
4. Can I predict profits based on # visitors?

ILLINOIS

# Motivating example

The table was hard to read. So here is the same data in a slightly more visual format..

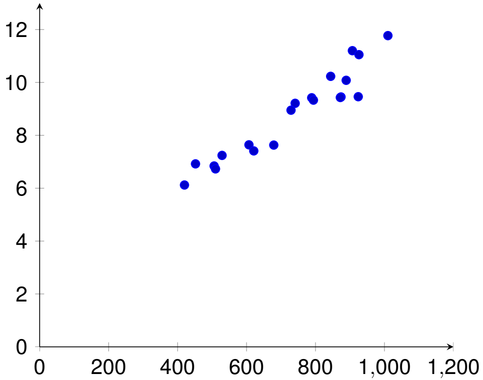



1. Do you see a relationship between profits and visits?
2. Is the relationship linear?
3. Is the relationship strong?
4. Can I predict profits based on # visitors?

ILLINOIS

# Motivating example

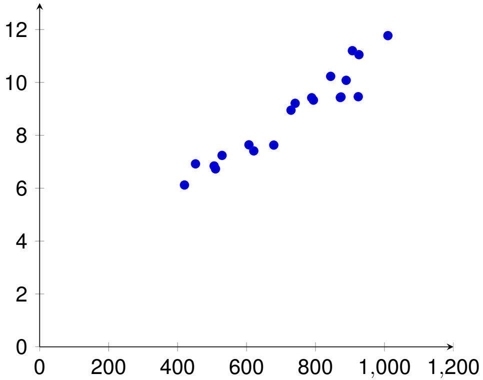
The table was hard to read. So here is the same data in a slightly more visual format..




1. Do you see a relationship between profits and visits?
2. Is the relationship linear?
3. Is the relationship strong?
4. Can I predict profits based on # visitors?

# Motivating example

The table was hard to read. So here is the same data in a slightly more visual format..

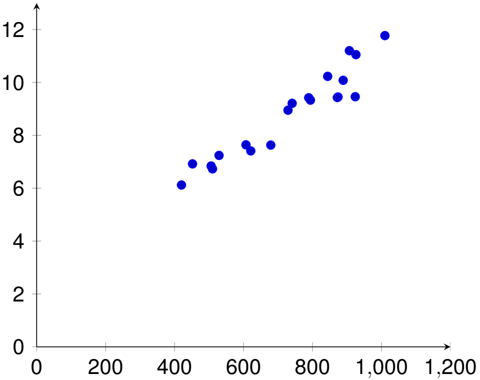



1. Do you see a relationship between profits and visits?
2. Is the relationship linear?
3. Is the relationship strong?
4. Can I predict profits based on # visitors?

ILLINOIS

# Motivating example

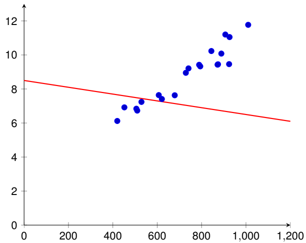
The table was hard to read. So here is the same data in a slightly more visual format..



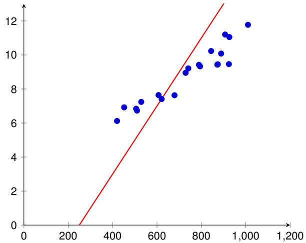

1. Do you see a relationship between profits and visits?
2. Is the relationship linear?
3. Is the relationship strong?
4. Can I predict profits based on # visitors?

ILLINOIS

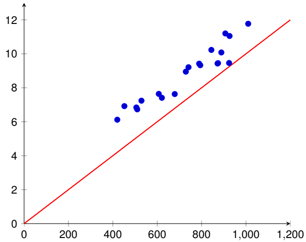
# Best line?



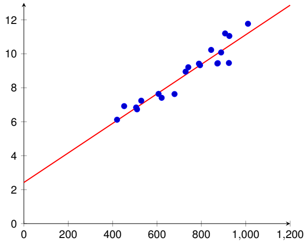

Bad line.




Incorrect “trend”.



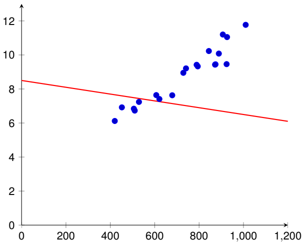

Underestimates.



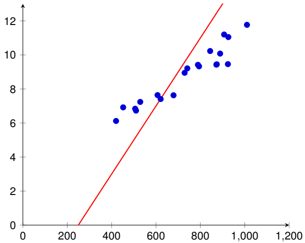

Good line!

ILLINOIS

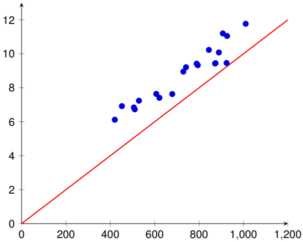
# Best line?



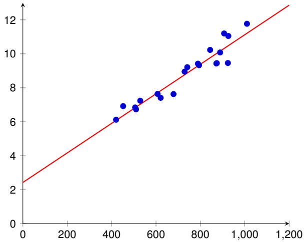

Bad line.




Incorrect “trend”.

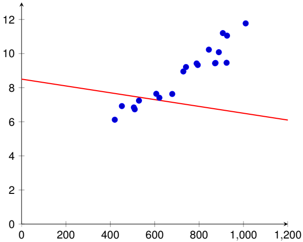



Underestimates.

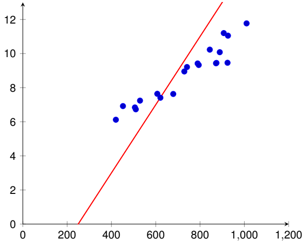



Good line!

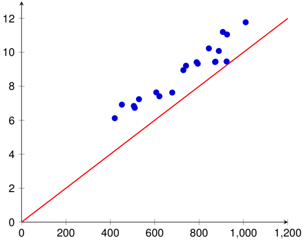
# Best line?



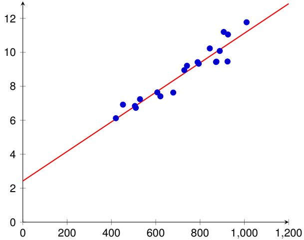

Bad line.




Incorrect “trend”.

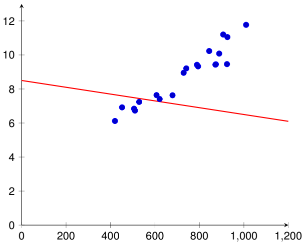



Underestimates.

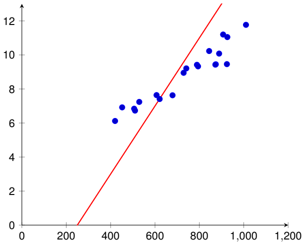



Good line!

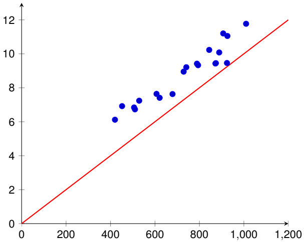
# Best line?



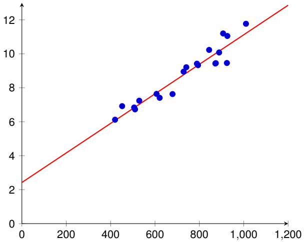

Bad line.




Incorrect “trend”.

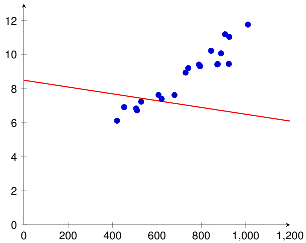



Underestimates.

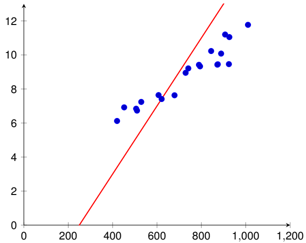



Good line!

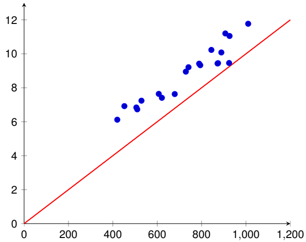
# Best line?



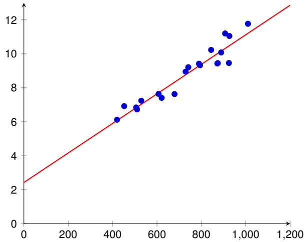

Bad line.




Incorrect “trend”.




Underestimates.




Good line!

# Least squares line

- The "best" line is the one that **minimizes the total deviations.**
- How to define deviations?

Main idea: every data point $(x_i, y_i)$ should satisfy:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i.}$$

- $\beta_0$: intercept.
- $\beta_1$: slope.
- $\epsilon_i$: noise related to data point $(x_i, y_i)$.

Noises $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

Finally, let the "total deviations" be measured by the squares of all noises:

$$L = \sum_{i=1}^{n} \epsilon_i^2.$$

# Least squares line

- The "best" line is the one that **minimizes the total deviations.**
- How to define deviations?

Main idea: every data point $(x_i, y_i)$ should satisfy:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i.}$$

- $\beta_0$: intercept.
- $\beta_1$: slope.
- $\epsilon_i$: noise related to data point $(x_i, y_i)$.

Noises $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Finally, let the "total deviations" be measured by the squares of all noises:

$$L = \sum_{i=1}^{n} \epsilon_i^2.$$

ILLINOIS

# Least squares line

- The “best” line is the one that **minimizes the total deviations.**
- How to define deviations?

Main idea: every data point $(x_i, y_i)$ should satisfy:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i.}$$

- $\beta_0$: intercept.
- $\beta_1$: slope.
- $\epsilon_i$: noise related to data point $(x_i, y_i)$.

Noises $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Finally, let the “total deviations” be measured by the squares of all noises:

$$L = \sum_{i=1}^{n} \epsilon_i^2.$$

ILLINOIS

# Least squares line

- The "best" line is the one that **minimizes the total deviations.**
- How to define deviations?

Main idea: every data point $(x_i, y_i)$ should satisfy:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i.}$$

- $\beta_0$: intercept.
- $\beta_1$: slope.
- $\epsilon_i$: noise related to data point $(x_i, y_i)$.

Noises $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

Finally, let the "total deviations" be measured by the squares of all noises:

$$L = \sum_{i=1}^{n} \epsilon_i^2.$$

ILLINOIS

# Least squares line

- The "best" line is the one that **minimizes the total deviations.**
- How to define deviations?

Main idea: every data point $(x_i, y_i)$ should satisfy:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \epsilon_i.}$$

- $\beta_0$: intercept.
- $\beta_1$: slope.
- $\epsilon_i$: noise related to data point $(x_i, y_i)$.

Noises $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

Finally, let the "total deviations" be measured by the squares of all noises:

$$L = \sum_{i=1}^{n} \epsilon_i^2.$$

ILLINOIS

# Least squares

We want to *minimize* $L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$.

- Take derivative, set to zero!
- What are our variables? $\beta_0, \beta_1$

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \implies$$

$$\implies \boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \implies$$

$$\implies \boxed{\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2}}$$

# Least squares

We want to *minimize* $L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$.

- Take derivative, set to zero!
- What are our variables? $\beta_0, \beta_1$.

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \implies$$

$$\implies \boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \implies$$

$$\implies \boxed{\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2}}$$

# Least squares

We want to *minimize* $L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$.

- Take derivative, set to zero!
- What are our variables?

$\beta_0, \beta_1$.

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \implies$$

$$\implies \boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \implies$$

$$\implies \boxed{\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2}}$$

ILLINOIS

## Example

*Recall the data from before. Let $x$ = visitors and $y$ = profit.*

| $n$ | Visitors | Profit | $n$ | Visitors | Profit |
|---|---|---|---|---|---|
| 1 | 907 | 11.2 | 2 | 926 | 11.05 |
| 3 | 506 | 6.84 | 4 | 741 | 9.21 |
| 5 | 789 | 9.42 | 6 | 889 | 10.08 |
| 7 | 874 | 9.45 | 8 | 510 | 6.73 |
| 9 | 529 | 7.24 | 10 | 420 | 6.12 |
| 11 | 679 | 7.63 | 12 | 872 | 9.43 |
| 13 | 924 | 9.46 | 14 | 607 | 7.64 |
| 15 | 452 | 6.92 | 16 | 729 | 8.95 |
| 17 | 794 | 9.33 | 18 | 844 | 10.23 |
| 19 | 1010 | 11.77 | 20 | 621 | 7.41 |

**Answer**: First, calculate $\sum x_i, \sum y_i, \sum x_i y_i, \sum x_i^2$.

- $\hat{\beta}_1 = \dfrac{n\sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{i=1}^{n} y_i\right)}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = 0.0087.$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 8.8055 - 0.0087 \cdot 731.15 = 2.423.$

ILLINOIS

## Example

*Recall the data from before. Let $x$ = visitors and $y$ = profit.*

| n | Visitors | Profit | n | Visitors | Profit |
|---|---|---|---|---|---|
| 1 | 907 | 11.2 | 2 | 926 | 11.05 |
| 3 | 506 | 6.84 | 4 | 741 | 9.21 |
| 5 | 789 | 9.42 | 6 | 889 | 10.08 |
| 7 | 874 | 9.45 | 8 | 510 | 6.73 |
| 9 | 529 | 7.24 | 10 | 420 | 6.12 |
| 11 | 679 | 7.63 | 12 | 872 | 9.43 |
| 13 | 924 | 9.46 | 14 | 607 | 7.64 |
| 15 | 452 | 6.92 | 16 | 729 | 8.95 |
| 17 | 794 | 9.33 | 18 | 844 | 10.23 |
| 19 | 1010 | 11.77 | 20 | 621 | 7.41 |

**Answer**: First, calculate $\sum x_i, \sum y_i, \sum x_i y_i, \sum x_i^2$.

- $\hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 0.0087.$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 8.8055 - 0.0087 \cdot 731.15 = 2.423.$

## Example

*Recall the data from before. Let $x$ = visitors and $y$ = profit.*

| n | Visitors | Profit | n | Visitors | Profit |
|---|---|---|---|---|---|
| 1 | 907 | 11.2 | 2 | 926 | 11.05 |
| 3 | 506 | 6.84 | 4 | 741 | 9.21 |
| 5 | 789 | 9.42 | 6 | 889 | 10.08 |
| 7 | 874 | 9.45 | 8 | 510 | 6.73 |
| 9 | 529 | 7.24 | 10 | 420 | 6.12 |
| 11 | 679 | 7.63 | 12 | 872 | 9.43 |
| 13 | 924 | 9.46 | 14 | 607 | 7.64 |
| 15 | 452 | 6.92 | 16 | 729 | 8.95 |
| 17 | 794 | 9.33 | 18 | 844 | 10.23 |
| 19 | 1010 | 11.77 | 20 | 621 | 7.41 |

**Answer**: First, calculate $\sum x_i, \sum y_i, \sum x_i y_i, \sum x_i^2$.

- $\hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 0.0087.$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 8.8055 - 0.0087 \cdot 731.15 = 2.423.$

ILLINOIS

## Example

*Recall the data from before. Let $x$ = visitors and $y$ = profit.*

| n | Visitors | Profit | n | Visitors | Profit |
|---|---|---|---|---|---|
| 1 | 907 | 11.2 | 2 | 926 | 11.05 |
| 3 | 506 | 6.84 | 4 | 741 | 9.21 |
| 5 | 789 | 9.42 | 6 | 889 | 10.08 |
| 7 | 874 | 9.45 | 8 | 510 | 6.73 |
| 9 | 529 | 7.24 | 10 | 420 | 6.12 |
| 11 | 679 | 7.63 | 12 | 872 | 9.43 |
| 13 | 924 | 9.46 | 14 | 607 | 7.64 |
| 15 | 452 | 6.92 | 16 | 729 | 8.95 |
| 17 | 794 | 9.33 | 18 | 844 | 10.23 |
| 19 | 1010 | 11.77 | 20 | 621 | 7.41 |

**Answer**: First, calculate $\sum x_i, \sum y_i, \sum x_i y_i, \sum x_i^2$.

- $\hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = 0.0087.$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 8.8055 - 0.0087 \cdot 731.15 = 2.423.$