Probability and Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

May 7, 2019

Topics

Objectives

Objectives:

- Obtain the slope and y-intercept of a fitted regression line.
- Use a fitted regression line to predict a value of the response variable for a given value of the explanatory variable.
- Use a fitted regression line to quantify a typical change in the response variable for a given change in the explanatory variable.
- Obtain and interpret fitted values and residuals.
- Interpret the R-squared statistic as a measure of how well a fitted regression line fits the data.

Linear Regression (12.1, 12.2)

- A ***linear regression analysis*** consists of obtaining the **equation** of the **line** that best fits the bivariate data in a scatterplot.

Linear Regression (12.1, 12.2)

- A ***linear regression analysis*** consists of obtaining the **equation** of the **line** that best fits the bivariate data in a scatterplot.
- Performing a **regression analysis** is useful for:

Linear Regression (12.1, 12.2)

- A ***linear regression analysis*** consists of obtaining the **equation** of the **line** that best fits the bivariate data in a scatterplot.
- Performing a **regression analysis** is useful for:
  1. **Predicting** the value of Y from a given value X (using the **equation** of the line).

Linear Regression (12.1, 12.2)

- A ***linear regression analysis*** consists of obtaining the **equation** of the **line** that best fits the bivariate data in a scatterplot.
- Performing a **regression analysis** is useful for:
  1. **Predicting** the value of Y from a given value X (using the **equation** of the line).
  2. **Quantifying** a typical **change** in Y associated with a given **change** in X (using the **slope** of the line).

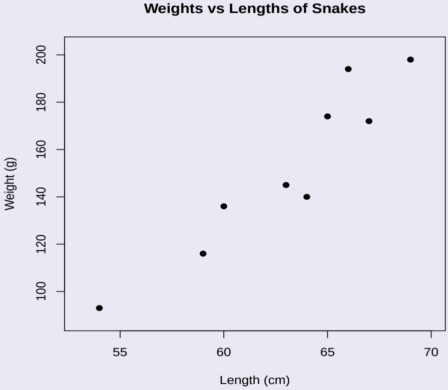
Linear Regression (12.1, 12.2)

- A ***linear regression analysis*** consists of obtaining the **equation** of the **line** that best fits the bivariate data in a scatterplot.
- Performing a **regression analysis** is useful for:
  1. **Predicting** the value of Y from a given value X (using the **equation** of the line).
  2. **Quantifying** a typical **change** in Y associated with a given **change** in X (using the **slope** of the line).
  3. Adding the line to the scatterplot to enhance its **appearance**.

Example

Here are the data on **lengths** and **weights** of snakes and the scatterplot, to which we add the **fitted regression line.**

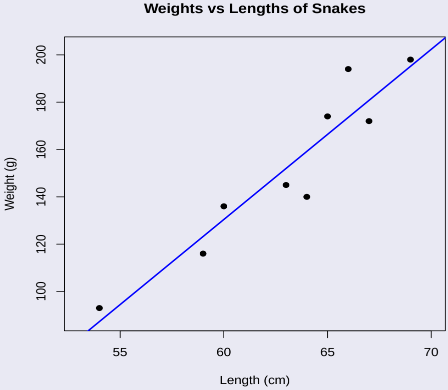
Snake	Length	Weight
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145



Example

Here are the data on **lengths** and **weights** of snakes and the scatterplot, to which we add the **fitted regression line.**

Snake	Length	Weight
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145



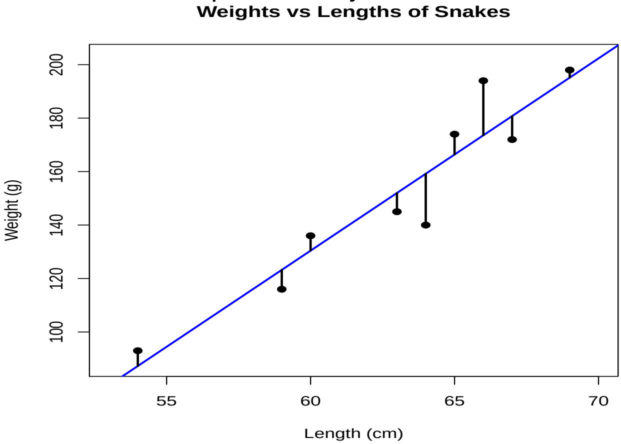
The **equation** of the **fitted regression line** is

$$\hat{y} = -301.09 + 7.19x \,,$$

(where y is **weight** and x is **length**).

- What **weight** would we **predict** for a snake whose **length** is **62** cm?
- What's a typical **change** in **weight** for each **1** cm **elongation**? What would we expect the **change** in **weight** to be for a **5** cm **elongation**?

- A line is considered to fit the data ”well” if the **vertical deviations** of the points away from it are **small.**



- **The Principle of Least Squares**: The "best fitting" line is the one that minimizes the **sum** of **squared vertical deviations**

$$\sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2$$

of the observed y_i values away from the line.

- The **slope** b_1 and and y**-intercept** b_0 of the line that **minimizes** the **sum of squared deviations** are computed from the data by:

- The **slope** b_1 and and y**-intercept** b_0 of the line that **minimizes** the **sum of squared deviations** are computed from the data by:

Fitted Regression Line Slope:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The **slope** b_1 and and y**-intercept** b_0 of the line that **minimizes** the **sum of squared deviations** are computed from the data by:

Fitted Regression Line Slope:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Fitted Regression Line Intercept:

$$b_0 = \bar{y} - b_1\bar{x}$$

Proof: Treating the x_i's and y_i's as constants, we define a **two-variable function** of b_0 and b_1 by

$$f(b_0,\, b_1) \;=\; \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2\,.$$

Proof: Treating the $\boldsymbol{x_i}$'s and $\boldsymbol{y_i}$'s as constants, we define a **two-variable function** of $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ by

$$f(b_0,\, b_1) \;=\; \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2\,.$$

Using **two-variable calculus**, we take **partial derivatives** with respect to $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$, set the derivatives equal to **zero**, and solve the resulting **system** of **two equations** in **two unknowns**:

Proof: Treating the $\boldsymbol{x_i}$'s and $\boldsymbol{y_i}$'s as constants, we define a **two-variable function** of $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ by

$$f(b_0,\, b_1) \;=\; \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \,.$$

Using **two-variable calculus**, we take **partial derivatives** with respect to $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$, set the derivatives equal to **zero**, and solve the resulting **system** of **two equations** in **two unknowns**:

$$\frac{d}{db_0} f(b_0,\, b_1) \;=\; 0 \qquad \text{and} \qquad \frac{d}{db_1} f(b_0,\, b_1) \;= 0$$

Proof: Treating the $\boldsymbol{x_i}$'s and $\boldsymbol{y_i}$'s as constants, we define a **two-variable function** of $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ by

$$f(b_0,\, b_1) \;=\; \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2 \,.$$

Using **two-variable calculus**, we take **partial derivatives** with respect to $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$, set the derivatives equal to **zero**, and solve the resulting **system** of **two equations** in **two unknowns**:

$$\frac{d}{db_0} f(b_0,\, b_1) \;=\; 0 \qquad \text{and} \qquad \frac{d}{db_1} f(b_0,\, b_1) \;= 0$$

It can be shown that the $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ given on the previous slide solve these equations.

- The resulting ”best fitting” line is called the ***fitted regression line***:

Fitted Regression Line:

$$\hat{y} \ = \ b_0 + b_1 x \, .$$

- The resulting ”best fitting” line is called the ***fitted regression line***:

Fitted Regression Line:

$$\hat{y} \;=\; b_0 + b_1 x\,.$$

The ”hat” over the y is to remind us that it’s the **fitted regression line.**

- **Some Cautionary Notes About Regression**:

- **Some Cautionary Notes About Regression**:
    1. Beware of ***extrapolation*** (using the line to predict y for values of x outside the range of the data at hand).

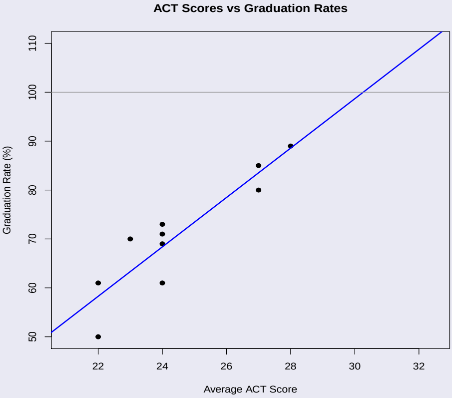
- **Some Cautionary Notes About Regression**:
  1. Beware of ***extrapolation*** (using the line to predict y for values of x outside the range of the data at hand).
  2. Beware of ***influential*** outliers. **Outliers** in the x **direction** can be particularly **influential**.

- **Some Cautionary Notes About Regression**:
  1. Beware of ***extrapolation*** (using the line to predict y for values of x outside the range of the data at hand).
  2. Beware of ***influential*** outliers. **Outliers** in the x **direction** can be particularly **influential**.
- The next example illustrates the danger of **extrapolation**.

Example

ACT exam scores are often used to predict **graduation rates** at universities. The average **ACT score** and **percentage** of freshmen who **graduate** are presented below for ten large universities.

University	ACT Average	Graduation Rate (%)
Illinois	27	80
Indiana	24	69
Iowa	24	61
Michigan	27	85
Michigan State	23	70
Minnesota	22	50
Northwestern	28	89
Ohio State	22	61
Purdue	24	71
Wisconsin	24	73



The equation of the **fitted regression line** is

$$\hat{y} = -52.8 + 5.05x.$$

(where y is **graduation rate** and x is average **ACT score**).

The equation of the **fitted regression line** is

$$\hat{y} = -52.8 + 5.05x.$$

(where y is **graduation rate** and x is average **ACT score**).

Another university's average **ACT score** is **32**.

The equation of the **fitted regression line** is

$$\hat{y} = -52.8 + 5.05x.$$

(where y is **graduation rate** and x is average **ACT score**).

Another university's average **ACT score** is **32.**

Would a **prediction** of its **graduation rate** based on the regression be an **extrapolation**?

The equation of the **fitted regression line** is

$$\hat{y} = -52.8 + 5.05x.$$

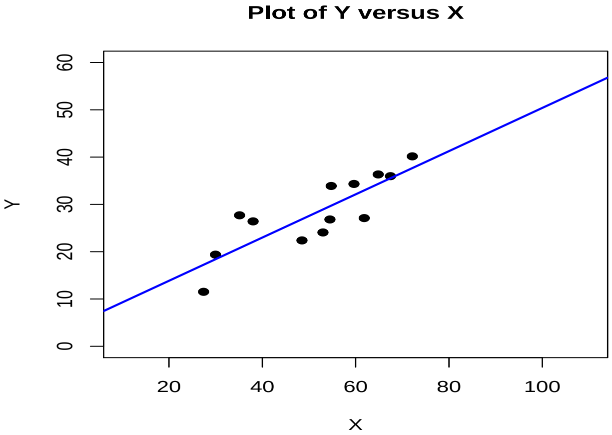
(where y is **graduation rate** and x is average **ACT score**).

Another university's average **ACT score** is **32.**

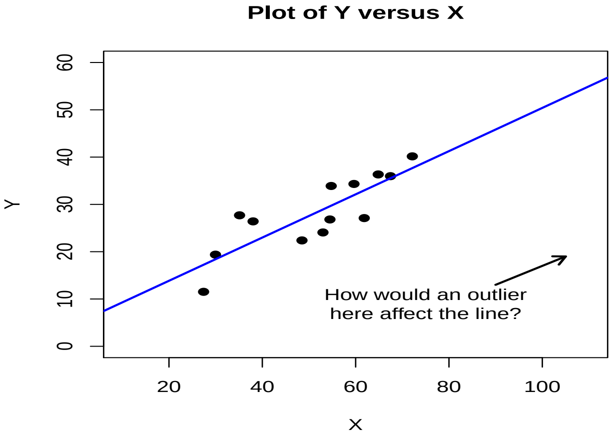
Would a **prediction** of its **graduation rate** based on the regression be an **extrapolation**?

Would the **prediction** be trustworthy?

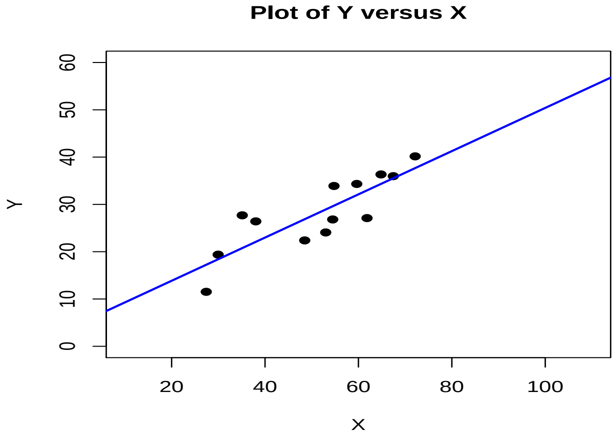
- Some outliers are **influential.**



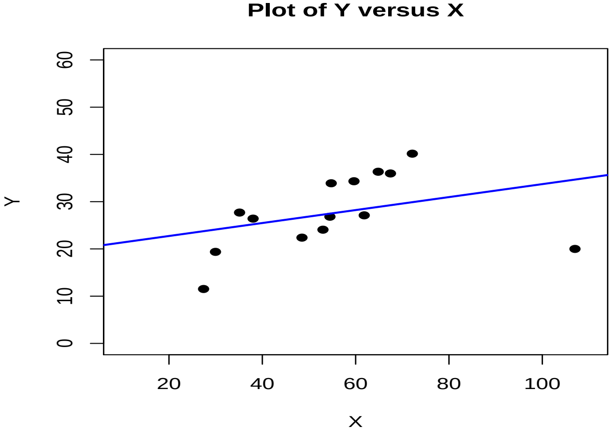
- Some outliers are **influential.**



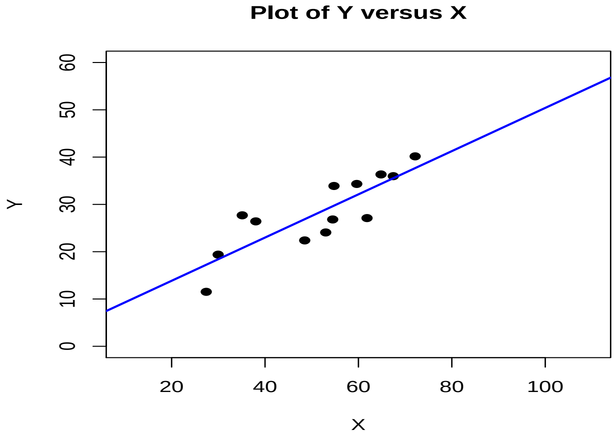
- Some outliers are **influential.**



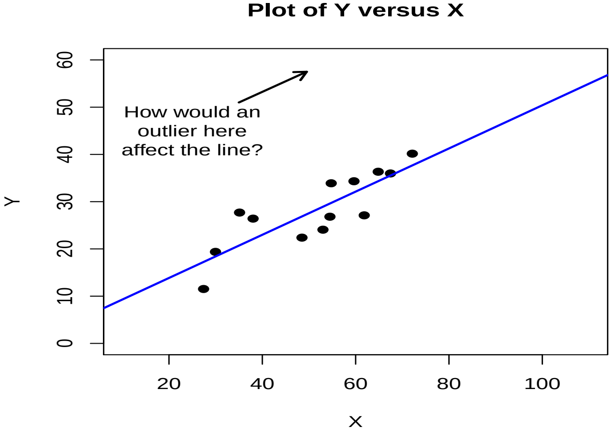
- Some outliers are **influential.**



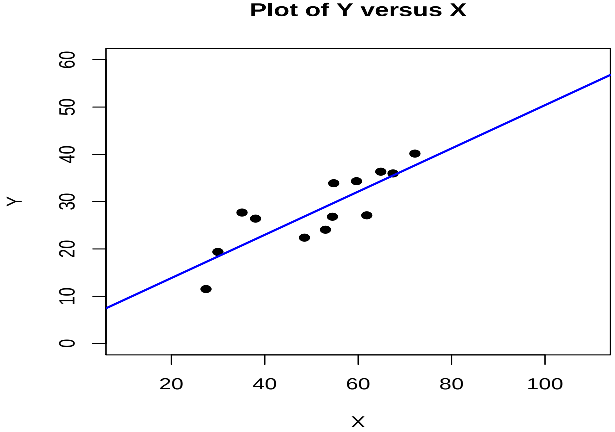
- Other outliers are **not** influential.



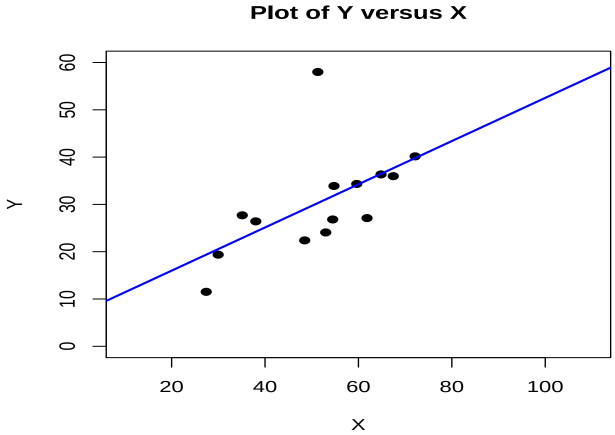
- Other outliers are **not** influential.



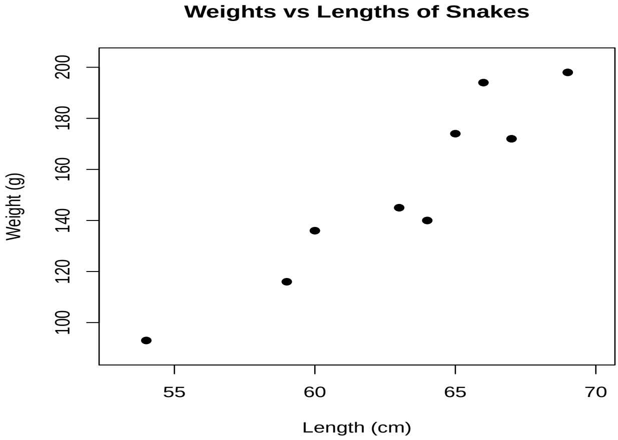
- Other outliers are **not** influential.



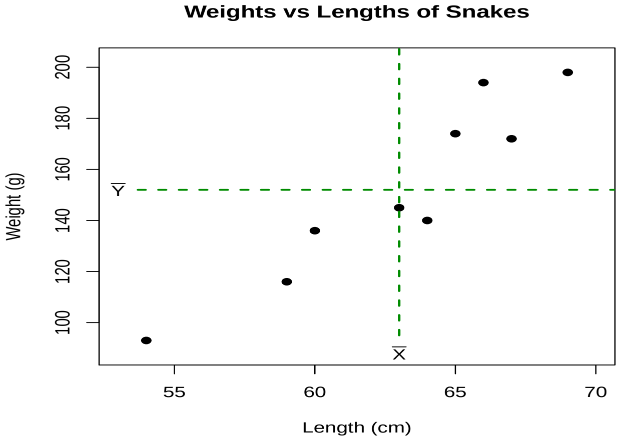
- Other outliers are **not** influential.



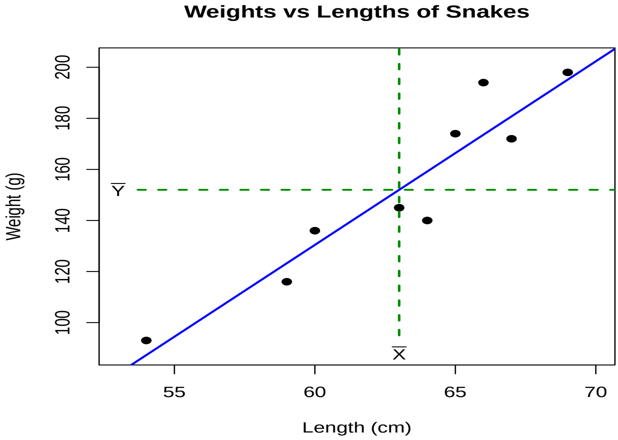
- The **fitted regression line** always goes through the **point of averages** $(\bar{x},\ \bar{y})$.



- The **fitted regression line** always goes through the **point of averages** $(\bar{x},\ \bar{y})$.



- The **fitted regression line** always goes through the **point of averages** $(\bar{x},\ \bar{y})$.



- It can be shown (via algebraic manipulation of the previous formula) that an **equivalent formula** for the **slope** $\boldsymbol{b_1}$ is:

Alternative Formula for Fitted Regression Line Slope:

$$b_1 \;=\; r \times \frac{s_y}{s_x}$$

where $\boldsymbol{r}$ is the **correlation** and $\boldsymbol{s_x}$ and $\boldsymbol{s_y}$ are the $\boldsymbol{x}$ and $\boldsymbol{y}$ sample **standard deviations.**

- In particular:

- In particular:
  - The **slope** $\boldsymbol{b_1}$ *always* has the **same sign** as the **correlation** $\boldsymbol{r}$ (because s_x and s_y are positive).

- In particular:
  - The **slope $\boldsymbol{b_1}$** *always* has the **same sign** as the **correlation $\boldsymbol{r}$** (because s_x and s_y are positive).
  - A one-standard deviation change in x leads to only an r-standard deviation change in y. This phenomenon is called ***regression to the mean***.

Fitted Values and Residuals (12.2)

- Statistics that measure how well a regression line fits the data are based on the **vertical deviations** of the points away from the fitted line.

Fitted Values and Residuals (12.2)

- Statistics that measure how well a regression line fits the data are based on the **vertical deviations** of the points away from the fitted line.
- To obtain the **vertical deviations**, we'll need the ***fitted values*** (also called ***predicted values***), denoted $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$:

Fitted Values and Residuals (12.2)

- Statistics that measure how well a regression line fits the data are based on the **vertical deviations** of the points away from the fitted line.

- To obtain the **vertical deviations**, we'll need the ***fitted values*** (also called ***predicted values***), denoted $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$:

Fitted (or **Predicted**) **Values** : For each $i = 1, 2, \ldots, n$,

$$\hat{y}_i = b_0 + b_1 x_i,$$

where x_i is the value of the explanatory variable for the ith individual in the data set.

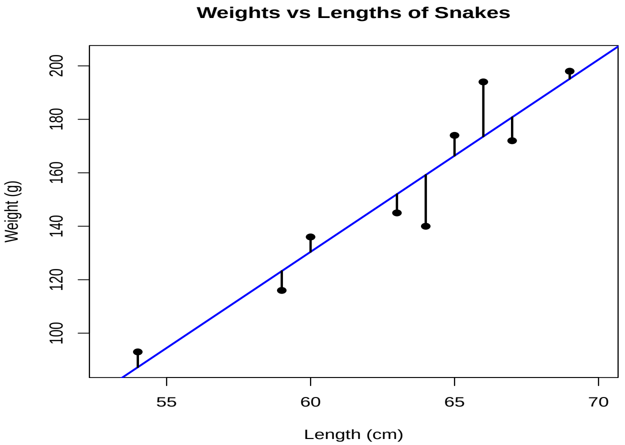
- The **vertical deviations** of the points away from the fitted line are called ***residuals***, denoted e_i:

- The **vertical deviations** of the points away from the fitted line are called ***residuals***, denoted e_i:

Residuals: For each $i = 1, 2, \ldots, n$,

$$e_i \;=\; y_i - \hat{y}_i,$$

where y_i is the value of the response variable for the ith individual in the data set.


Weights vs Lengths of Snakes
Weight (g)
Length (cm)
100
120
140
160
180
200
55
60
65
70

- **Residuals** are **positive** or **negative** depending on whether the point lies above or below the fitted line. They always **sum to zero**, i.e.

$$\sum_{i=1}^{n} e_i = 0.$$

- **Residuals** are **positive** or **negative** depending on whether the point lies above or below the fitted line. They always **sum to zero**, i.e.

$$\sum_{i=1}^{n} e_i \ = \ 0.$$

- **Residuals** are the net effect on the response y of **all *other* variables *besides*** the explanatory variable x.

- **Residuals** are **positive** or **negative** depending on whether the point lies above or below the fitted line. They always **sum to zero**, i.e.

$$\sum_{i=1}^{n} e_i = 0.$$

- **Residuals** are the net effect on the response y of **all *other* variables *besides*** the explanatory variable x.

  Example: ***Other* variables *besides* length** that affect a snake's **weight**, and contribute to the **residuals**, include the snake's bone density, circumference, diet/caloric intake, metabolic rate, etc.

The R^2 (12.2)

- One statistic that measures how well the line fits the data is the ***residual sum of squares***, also called the ***error sum of squares***, denoted **SSE**:

Residual (or **Error**) **Sum of Squares**:

$$\text{SSE} = \sum_{i=1}^{n} e_i^2$$

- **SSE** depends on the sample size n, so a better way to measure the fit is the "average" squared residual (using $n-2$), denoted $\boldsymbol{s^2}$:

Mean Squared Residual (or **Error**):

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

- **SSE** depends on the sample size n, so a better way to measure the fit is the ”average” squared residual (using $n-2$), denoted $\boldsymbol{s^2}$:

Mean Squared Residual (or **Error**):

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

(We divide by $n-2$ instead of n because it results in a statistic s^2 that more accurately estimates the population variance σ^2 away from the population regression line.)

- The ***square root*** of the **mean squared residual**, s, represents the size of a ***typical*** **vertical deviation** away from the fitted line.

- s^2 and s measure how well the line fits the data, but their values depend on the **measurement scale** of y.

- s^2 and s measure how well the line fits the data, but their values depend on the **measurement scale** of y.
- The ***coefficient of determination*** (or "R-squared"), denoted r^2, also measures how well the line fits, but its value **doesn't** depend on the **scale** of y:

Coefficient of Determination (or **R-Squared**):

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where **SSE** is *error sum of squares* and **SST** is the ***total sum of squares*** defined as

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

- The value of r^2 tells us **how well** the line **fits** the data:

- The value of r^2 tells us **how well** the line **fits** the data:
  - r^2 values **near zero** imply a very **poor** fit.

- The value of r^2 tells us **how well** the line **fits** the data:
  - r^2 values **near zero** imply a very **poor** fit.
  - r^2 values **close to 1.0** imply a very **good fit**.

- The value of r^2 tells us **how well** the line **fits** the data:
  - r^2 values **near zero** imply a very **poor** fit.
  - r^2 values **close to 1.0** imply a very **good fit**.
- **Interpretation** of r^2:
  r^2 represents the **proportion of variation** *in the responses $y_1, y_2, \ldots, y_n$ that can be explained by differences among $x_1, x_2, \ldots, x_n$ and the linear relationship of y to x.*

- To understand the **interpretation** from the last slide, note that:

- To understand the **interpretation** from the last slide, note that:
    - **SST** measures the **total variation** in the y_i's.

- To understand the **interpretation** from the last slide, note that:
    - **SST** measures the **total variation** in the y_i's.
    - **SSE** measures **residual variation** in the y_i's due to **all other** variables ***besides*** the explanatory variable x.

- Therefore:

$$\frac{\text{SSE}}{\text{SST}} = $$

Proportion of variation in $y_1, y_2, \ldots, y_n$ that's due to differences among values of **other variables besides x.**

- Therefore:

$$\frac{\text{SSE}}{\text{SST}} = \text{Proportion of variation in } y_1, y_2, \ldots, y_n \text{ that's due to differences among values of } \textbf{other variables besides } \boldsymbol{x}\textbf{.}$$

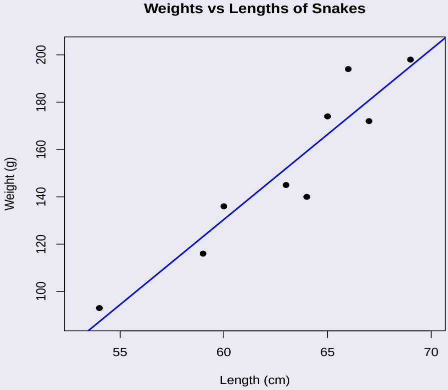
and so $\boldsymbol{r^2}$ is

$$1 - \frac{\text{SSE}}{\text{SST}} = \text{Proportion of variation in } y_1, y_2, \ldots, y_n \text{ that's due to differences among the values of } \boldsymbol{x_1, x_2, \ldots, x_n}.$$

Example

Here again are the data on **lengths** and **weights** of snakes and the scatterplot with the **fitted regression line**.

Snake	Length	Weight
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145



The **equation** of the **fitted regression line** is

$$\hat{y} = -301.09 + 7.19x \,,$$

(where y is **weight** and x is **length**).

The **equation** of the **fitted regression line** is

$$\hat{y} = -301.09 + 7.19x\,,$$

(where y is **weight** and x is **length**).

Statistical software gives:

$$\text{SSE} \;=\; \mathbf{1093.7} \qquad \text{and} \qquad \text{SST} \;=\; \mathbf{9990.0},$$

The **equation** of the **fitted regression line** is

$$\hat{y} = -301.09 + 7.19x\,,$$

(where $\boldsymbol{y}$ is **weight** and $\boldsymbol{x}$ is **length**).

Statistical software gives:

$$\text{SSE} \;=\; \mathbf{1093.7} \qquad \text{and} \qquad \text{SST} \;=\; \mathbf{9990.0},$$

so the **R-squared** value is

$$\boldsymbol{r^2} \;=\; 1 - \frac{1093.7}{9990.0} \;=\; \mathbf{0.89}.$$

Thus **89%** of the variation in snakes' **weights** is attributable to differences among their **lengths.**

Thus **89%** of the variation in snakes' **weights** is attributable to differences among their **lengths.**

The other **11%** of the variation in **weights** is attributable to differences among the values of **all the *other* variables *besides* length** that affect **weight** (such as bone density, circumference, diet/caloric intake, metabolic rate, etc.)

- It can be shown that the **coefficient of determination** (**R-squared**) r^2 is the **square** of the **correlation** r, i.e.

$$r^2 \;=\; (r)^2\,.$$

Example

The **correlation** between **lengths** and **weights** of snakes is

$$\boldsymbol{r \;=\; 0.944}\,.$$

- It can be shown that the **coefficient of determination** (**R-squared**) $\boldsymbol{r^2}$ is the **square** of the **correlation** $\boldsymbol{r}$, i.e.

$$r^2 \;=\; (r)^2\,.$$

Example

The **correlation** between **lengths** and **weights** of snakes is

$$\boldsymbol{r \;=\; 0.944}\,.$$

By **squaring** the **correlation** we get the same **R-squared** value

$$\boldsymbol{r^2} \;=\; 0.944^2 \;=\; \boldsymbol{0.89}$$

that was obtained in the last example.