Lesson 2:
Chapter 2

Caleb Moxley

BSC Mathematics

2 September 15

§2.1 Relationships

In this chapter, we will investigate relationships between two variables which describe the **same case**. For instance, height and weight can be two variables which are related if we are talking about the height and weight of the same case - like the height and weight of a horse. Variables can only be related if they both measure the same case!

§2.1 Relationships

In this chapter, we will investigate relationships between two variables which describe the **same case**. For instance, height and weight can be two variables which are related if we are talking about the height and weight of the same case - like the height and weight of a horse. Variables can only be related if they both measure the same case!

Definition (association)

There is an **association between two variables** if knowing the value of one variable can tell you something about the value of another variable.

§2.1 Relationships

In this chapter, we will investigate relationships between two variables which describe the **same case**. For instance, height and weight can be two variables which are related if we are talking about the height and weight of the same case - like the height and weight of a horse. Variables can only be related if they both measure the same case!

Definition (association)

There is an **association between two variables** if knowing the value of one variable can tell you something about the value of another variable.

Example (association)

Is there an association between the balance on a loan and whether or not the loan is in default?

§2.1 Relationships

In this chapter, we will investigate relationships between two variables which describe the **same case**. For instance, height and weight can be two variables which are related if we are talking about the height and weight of the same case - like the height and weight of a horse. Variables can only be related if they both measure the same case!

Definition (association)

There is an **association between two variables** if knowing the value of one variable can tell you something about the value of another variable.

Example (association)

Is there an association between the balance on a loan and whether or not the loan is in default? If there is an association, is it weak or strong?

§2.1 Relationships

In many relationships, one variable may help explain the other.

§2.1 Relationships

In many relationships, one variable may help explain the other.

Definition (explanatory and response variable)

In a study the variable which is measured is the **response** variable. The variable being changed within the experiment which explains or causes the change in the measured variable is the **explanatory** variable.

§2.1 Relationships

In many relationships, one variable may help explain the other.

Definition (explanatory and response variable)

In a study the variable which is measured is the **response** variable. The variable being changed within the experiment which explains or causes the change in the measured variable is the **explanatory** variable.

In a study measuring the effects of light on attention, what is the explanatory variable? What is the response variable? (Assume that attention is measured by the time in which a subject can complete a menial task.)

§2.2 Scatterplots

Scatterplots are the most basic graphical representation of relationships.

§2.2 Scatterplots

Scatterplots are the most basic graphical representation of relationships.

Definition (scatterplot)

A **scatterplot** graphs the points created by two measurements on individual cases. Explanatory variables usually are placed on the horizontal axis.

§2.2 Scatterplots

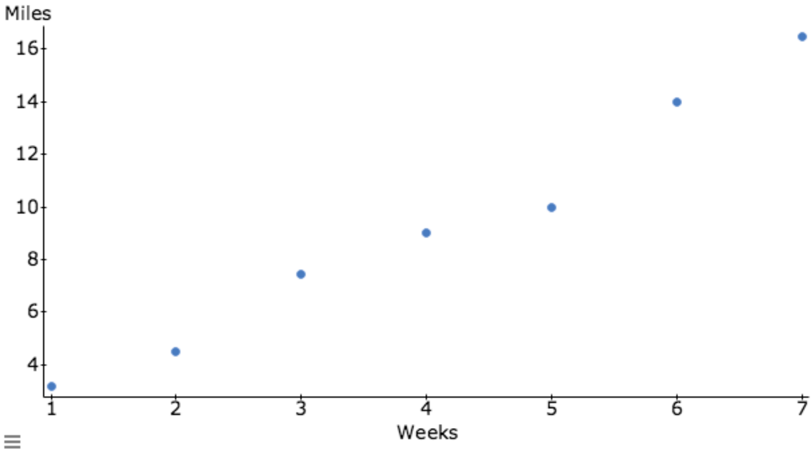
Scatterplots are the most basic graphical representation of relationships.

Definition (scatterplot)

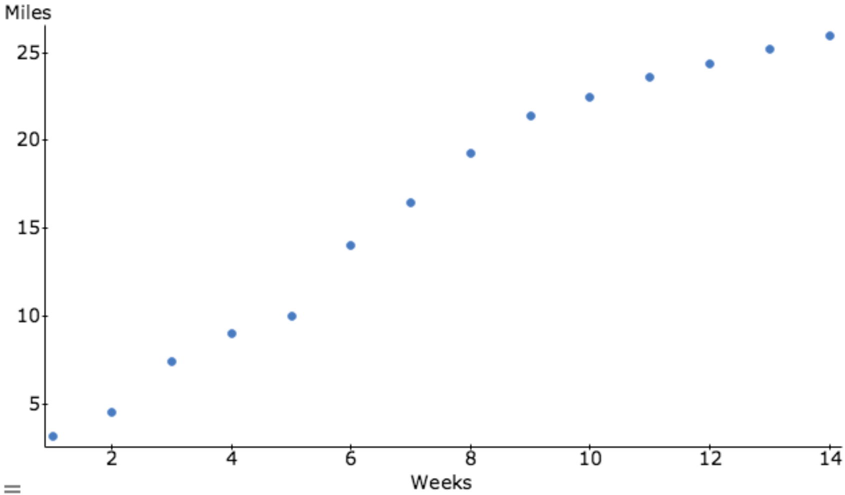
A **scatterplot** graphs the points created by two measurements on individual cases. Explanatory variables usually are placed on the horizontal axis.

You can see positive/negative associations in scatterplots. Positive associations are those where the variables increase together. Negative associations are those where one variable increases as the other decreases. You can also see the **strength** of an association from the scatterplot.

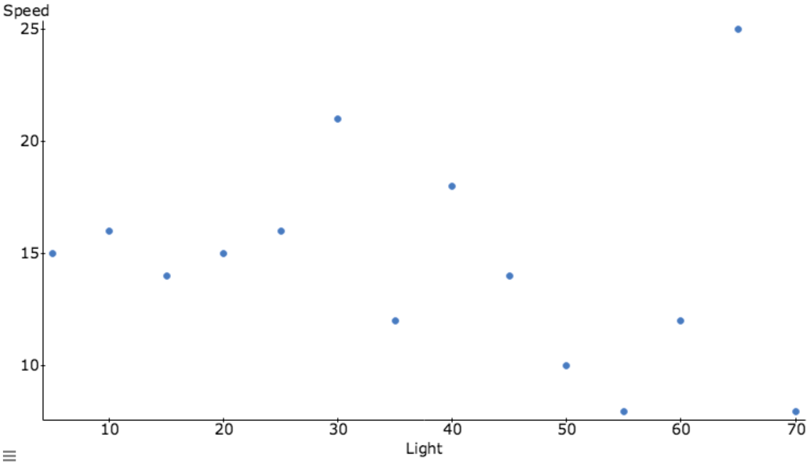
§2.2 Scatterplots



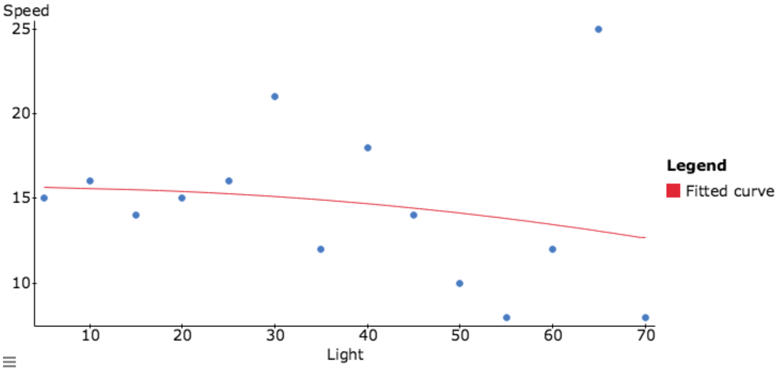
§2.2 Scatterplots



§2.2 Scatterplots



§2.2 Scatterplots



§2.3 Correlation

Definition (correlation)

The correlation between two variables is given by

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

§2.3 Correlation

Definition (correlation)

The correlation between two variables is given by

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Correlation can be calculated for any two variables! They need not be explanatory/response pairs or even associated in any way. It simply measures how well the data fit a straight line pattern.

§2.3 Correlation

Definition (correlation)

The correlation between two variables is given by

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Correlation can be calculated for any two variables! They need not be explanatory/response pairs or even associated in any way. It simply measures how well the data fit a straight line pattern. Correlation values close to 1 or -1 are high correlation values. Correlation values near 0 are low correlation values.

§2.3 Correlation

Here's the data for the miles ran after a certain number of weeks of training:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

§2.3 Correlation

Here's the data for the miles ran after a certain number of weeks of training:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

Example

It's easy to check that $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, and $s_y = 4.798$. So, we calculate r:

$$\frac{1}{6}\left(\left(\frac{1-4}{2.160}\right)\left(\frac{3.2-9.239}{4.798}\right)+\cdots+\left(\frac{7-4}{2.160}\right)\left(\frac{16.5-9.239}{4.798}\right)\right).$$

§2.3 Correlation

Here's the data for the miles ran after a certain number of weeks of training:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

Example

It's easy to check that $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, and $s_y = 4.798$. So, we calculate r:

$$\frac{1}{6}\left(\left(\frac{1-4}{2.160}\right)\left(\frac{3.2-9.239}{4.798}\right)+\cdots+\left(\frac{7-4}{2.160}\right)\left(\frac{16.5-9.239}{4.798}\right)\right).$$

This gives $r = 0.988$.

§2.3 Correlation

Some notes on r:

§2.3 Correlation

Some notes on r:

1. The sign of r indicates the nature of the relationship.

§2.3 Correlation

Some notes on r:

1. The sign of r indicates the nature of the relationship.
2. Correlation can only be calculated for **linear** relationships. Non-linear relationships can be identified by their scatterplot, and they should not have correlation calculated for them.

§2.3 Correlation

Some notes on r:

1. The sign of r indicates the nature of the relationship.
2. Correlation can only be calculated for **linear** relationships. Non-linear relationships can be identified by their scatterplot, and they should not have correlation calculated for them.
3. Correlation can only be calculated when we have matched pairs of data. Thus, there should always be the same number of x- and y-values.

§2.4 Least-Squares Regression

Definition (regression line)

A **regression line** is a straight line that describes how a response variable (y) changes as an explanatory variable (x) changes.

§2.4 Least-Squares Regression

Definition (regression line)

A **regression line** is a straight line that describes how a response variable (y) changes as an explanatory variable (x) changes. The **least-squares regression line** is the regression line that minimizes the vertical distances of data points (x_i, y_i) from the line. The least-squares regression line is unique for data sets that include at least two pairs.

§2.4 Least-Squares Regression

Definition (regression line)

A **regression line** is a straight line that describes how a response variable (y) changes as an explanatory variable (x) changes. The **least-squares regression line** is the regression line that minimizes the vertical distances of data points (x_i, y_i) from the line. The least-squares regression line is unique for data sets that include at least two pairs.

Theorem

The least-squares regression line is always given by the formula $\hat{y} = b_0 + b_1 x$ *where* $b_1 = r\frac{s_y}{s_x}$ *and* $b_0 = \bar{y} - b_1\bar{x}$.

§2.4 Least-Squares Regression

We already calculated several pieces of summary data for the pairs below:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

We have $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, $s_y = 4.798$, and $r = 0.988$.

We already calculated several pieces of summary data for the pairs below:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

We have $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, $s_y = 4.798$, and $r = 0.988$.

Thus, $b_1 = 0.988\frac{4.798}{2.160} = 2.195$

§2.4 Least-Squares Regression

We already calculated several pieces of summary data for the pairs below:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

We have $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, $s_y = 4.798$, and $r = 0.988$.

Thus, $b_1 = 0.988\frac{4.798}{2.160} = 2.195$ and $b_0 = 9.239 - 2.195(4) = 0.46$.

§2.4 Least-Squares Regression

We already calculated several pieces of summary data for the pairs below:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

We have $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, $s_y = 4.798$, and $r = 0.988$.

Thus, $b_1 = 0.988\frac{4.798}{2.160} = 2.195$ and $b_0 = 9.239 - 2.195(4) = 0.46$. So we have that the least-squares regression line is

$$\hat{y} = 0.46 + 2.195x.$$

§2.4 Least-Squares Regression

We already calculated several pieces of summary data for the pairs below:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

We have $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, $s_y = 4.798$, and $r = 0.988$.

Thus, $b_1 = 0.988\frac{4.798}{2.160} = 2.195$ and $b_0 = 9.239 - 2.195(4) = 0.46$. So we have that the least-squares regression line is

$$\hat{y} = 0.46 + 2.195x.$$

We can use this to create best predicted values!

We already calculated several pieces of summary data for the pairs below:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

We have $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, $s_y = 4.798$, and $r = 0.988$.

Thus, $b_1 = 0.988\frac{4.798}{2.160} = 2.195$ and $b_0 = 9.239 - 2.195(4) = 0.46$. So we have that the least-squares regression line is

$$\hat{y} = 0.46 + 2.195x.$$

We can use this to create best predicted values! The number of miles you might expect the trainee to run after 2.5 weeks would be about

§2.4 Least-Squares Regression

We already calculated several pieces of summary data for the pairs below:

W	1	2	3	4	5	6	7
M	3.2	4.5	7.45	9.02	10	14	16.5

We have $\bar{x} = 4$, $s_x = 2.160$, $\bar{y} = 9.239$, $s_y = 4.798$, and $r = 0.988$.

Thus, $b_1 = 0.988\frac{4.798}{2.160} = 2.195$ and $b_0 = 9.239 - 2.195(4) = 0.46$. So we have that the least-squares regression line is

$$\hat{y} = 0.46 + 2.195x.$$

We can use this to create best predicted values! The number of miles you might expect the trainee to run after 2.5 weeks would be about 0.46+(2.195)(2.5)=5.95.

§Caveats

You should be aware about the following issues:

§Caveats

You should be aware about the following issues:

- Causation and correlation are not the same thing!

§Caveats

You should be aware about the following issues:

- Causation and correlation are not the same thing!
- Two variables can be correlated linearly in a perfect fashion but not change by the same relative amounts.

§Caveats

You should be aware about the following issues:

- Causation and correlation are not the same thing!
- Two variables can be correlated linearly in a perfect fashion but not change by the same relative amounts.
- The r statistic is not resistant against outliers.