Lecture 6

CORRELATION AND LINEAR REGRESSION

Introduction

In this lecture we look at relationships between **pairs** of **continuous variables**. These might be

Introduction

In this lecture we look at relationships between **pairs** of **continuous variables**. These might be

- **height** and **weight**

In this lecture we look at relationships between **pairs** of **continuous variables**. These might be

- **height** and **weight**
- **market value** and **number of transactions**

Introduction

In this lecture we look at relationships between **pairs** of **continuous variables**. These might be

- **height** and **weight**
- **market value** and **number of transactions**
- **temperatures** and **sales of ice cream**

Introduction

In this lecture we look at relationships between **pairs** of **continuous variables**. These might be

- **height** and **weight**
- **market value** and **number of transactions**
- **temperatures** and **sales of ice cream**

We can examine the relationship between our two variables through a **scatter diagram** (plot one against the other).

Introduction

In this lecture we look at relationships between **pairs** of **continuous variables**. These might be

- **height** and **weight**
- **market value** and **number of transactions**
- **temperatures** and **sales of ice cream**

We can examine the relationship between our two variables through a **scatter diagram** (plot one against the other).

Today we will look at how to examine such relationships more formally.

Example: ice cream sales

The following data are monthly ice cream sales at Luigi Minchella's ice cream parlour.

Month	Average Temp (ºC)	Sales (£000's)
January	4	73
February	4	57
March	7	81
April	8	94
May	12	110
June	15	124
July	16	134
August	17	139
September	14	124
October	11	103
November	7	81
December	5	80

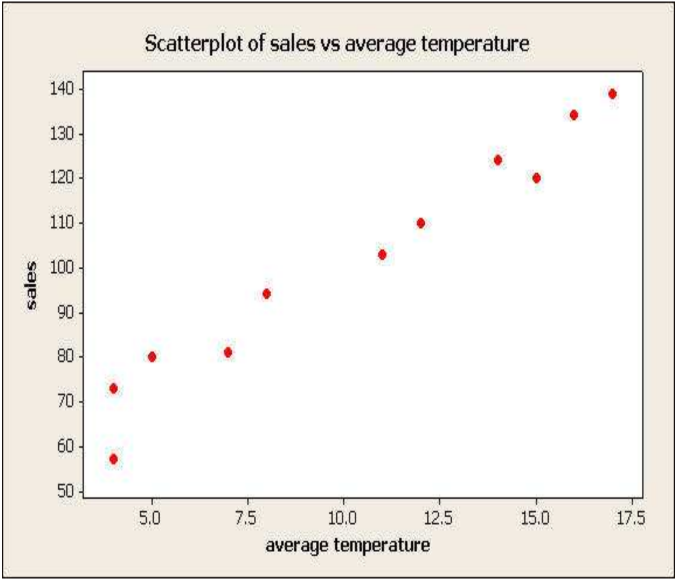
Month	Average Temp (°C)	Sales (£000's)
January	4	73
February	4	57
March	7	81
April	8	94
May	12	110
June	15	124
July	16	134
August	17	139
September	14	124
October	11	103
November	7	81
December	5	80

Is there any relationship between temperature and sales?

Month	Average Temp (°C)	Sales (£000's)
January	4	73
February	4	57
March	7	81
April	8	94
May	12	110
June	15	124
July	16	134
August	17	139
September	14	124
October	11	103
November	7	81
December	5	80

Is there any relationship between temperature and sales?

Can you DESCRIBE this relationship?


Scatterplot of sales vs average temperature
sales
140
130
120
110
100
90
80
70
60
50
5.0
7.5
10.0
12.5
15.0
17.5
average temperature

- The scatter plot goes "uphill", so we say there is a **positive** relationship between temperature and ice cream sales.

- The scatter plot goes “uphill”, so we say there is a **positive** relationship between temperature and ice cream sales.
- As temperatures rise, so do ice cream sales!

- The scatter plot goes "uphill", so we say there is a **positive** relationship between temperature and ice cream sales.
- As temperatures rise, so do ice cream sales!
- If the scatter plot had shown a "downhill" slope, we would have a **negative** relationship.

- The scatter plot goes "uphill", so we say there is a **positive** relationship between temperature and ice cream sales.
- As temperatures rise, so do ice cream sales!
- If the scatter plot had shown a "downhill" slope, we would have a **negative** relationship.
- We could draw a straight line through the middle of most of the points. Thus, there is also a **linear** association present.

- The scatter plot goes “uphill”, so we say there is a **positive** relationship between temperature and ice cream sales.
- As temperatures rise, so do ice cream sales!
- If the scatter plot had shown a “downhill” slope, we would have a **negative** relationship.
- We could draw a straight line through the middle of most of the points. Thus, there is also a **linear** association present.
- If we were to draw a line through the points, most points would lie close to this line. The closer the points lie to a line, the **stronger** the relationship.

We now look at how to **quantify** the relationship between two variables by calculating the **sample correlation coefficient**. This is

Correlation

We now look at how to **quantify** the relationship between two variables by calculating the **sample correlation coefficient**. This is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}},$$

where

Correlation

We now look at how to **quantify** the relationship between two variables by calculating the **sample correlation coefficient**. This is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}},$$

where

$$S_{XY} = \left(\sum xy\right) - n\bar{x}\bar{y},$$

Correlation

We now look at how to **quantify** the relationship between two variables by calculating the **sample correlation coefficient**. This is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}},$$

where

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y}, \\ S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2, \end{aligned}$$

Correlation

We now look at how to **quantify** the relationship between two variables by calculating the **sample correlation coefficient**. This is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}},$$

where

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y}, \\ S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2, \\ S_{YY} &= \left(\sum y^2\right) - n\bar{y}^2. \end{aligned}$$

- r always lies between -1 and $+1$.

- r always lies between -1 and $+1$.
- If r is close to $+1$, we have evidence of a strong **positive** (linear) association.

- r always lies between -1 and $+1$.
- If r is close to $+1$, we have evidence of a strong **positive** (linear) association.
- If r is close to -1, we have evidence of a strong **negative** (linear) association.

- r always lies between -1 and $+1$.
- If r is close to $+1$, we have evidence of a strong **positive** (linear) association.
- If r is close to -1, we have evidence of a strong **negative** (linear) association.
- If r is close to zero, there is probably **no** association!

- r always lies between -1 and $+1$.
- If r is close to $+1$, we have evidence of a strong **positive** (linear) association.
- If r is close to -1, we have evidence of a strong **negative** (linear) association.
- If r is close to zero, there is probably **no** association!
- A correlation coefficient close to zero does not imply no relationship at all, just no **linear** relationship.

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

x	**y**	$\mathbf{x}^2$	$\mathbf{y}^2$	**xy**
4	73			

The easiest way to calculate r is to draw up a table!

x	**y**	$\mathbf{x^2}$	$\mathbf{y^2}$	**xy**
4	73	16		

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

x	**y**	$\mathbf{x^2}$	$\mathbf{y^2}$	**xy**
4	73	16	5329	

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

x	**y**	$\mathbf{x^2}$	$\mathbf{y^2}$	**xy**
4	73	16	5329	292
4	57			

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

x	y	x^2	y^2	xy
4	73	16	5329	292
4	57	16		

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

x	**y**	$\mathbf{x^2}$	$\mathbf{y^2}$	**xy**
4	73	16	5329	292
4	57	16	3249	

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

x	y	x^2	y^2	xy
4	73	16	5329	292
4	57	16	3249	228

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

x	y	x^2	y^2	xy
4	73	16	5329	292
4	57	16	3249	228
7	81	49	6561	567

The easiest way to calculate r is to draw up a table!

x	y	x^2	y^2	xy
4	73	16	5329	292
4	57	16	3249	228
7	81	49	6561	567
8	94	64	8836	752

Example: ice cream sales

The easiest way to calculate r is to draw up a table!

x	y	x^2	y^2	xy
4	73	16	5329	292
4	57	16	3249	228
7	81	49	6561	567
8	94	64	8836	752
12	110	144	12100	1320
15	124	225	15376	1860
16	134	256	17956	2144
17	139	289	19321	2363
14	124	196	15376	1736
11	103	121	10609	1133
7	81	49	6561	567
5	80	25	6400	400
120	1200	1450	127674	13362

Notice in the formulae for S_{XY}, S_{XX} and S_{YY} we need the sample means $\bar{x}$ and $\bar{y}$. Thus,

Notice in the formulae for S_{XY}, S_{XX} and S_{YY} we need the sample means $\bar{x}$ and $\bar{y}$. Thus,

$$\begin{aligned} \bar{x} &= \frac{120}{12} \\ &= 10 \qquad \text{and} \end{aligned}$$

Notice in the formulae for S_{XY}, S_{XX} and S_{YY} we need the sample means $\bar{x}$ and $\bar{y}$. Thus,

$$\begin{aligned} \bar{x} &= \frac{120}{12} \\ &= 10 \qquad \text{and} \\ \bar{y} &= \frac{1200}{12} \\ &= 100. \end{aligned}$$

Similarly,

Similarly,

$$
\begin{aligned}
S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y} \\
&= 13362 - 12 \times 10 \times 100 \\
&= 1362,
\end{aligned}
$$

Similarly,

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y} \\ &= 13362 - 12 \times 10 \times 100 \\ &= 1362, \end{aligned}$$

$$\begin{aligned} S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2 \\ &= 1450 - 12 \times 10 \times 10 \\ &= 250 \qquad \text{and} \end{aligned}$$

Similarly,

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y} \\ &= 13362 - 12 \times 10 \times 100 \\ &= 1362, \end{aligned}$$

$$\begin{aligned} S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2 \\ &= 1450 - 12 \times 10 \times 10 \\ &= 250 \qquad \text{and} \end{aligned}$$

$$\begin{aligned} S_{YY} &= \left(\sum y^2\right) - n\bar{y}^2 \\ &= 127674 - 12 \times 100 \times 100 \\ &= 7674. \end{aligned}$$

Thus,

Thus,

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}}$$

Thus,

$$\begin{aligned} r &= \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}} \\ &= \frac{1362}{\sqrt{250 \times 7674}} \end{aligned}$$

Thus,

$$
\begin{aligned}
r &= \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}} \\
&= \frac{1362}{\sqrt{250 \times 7674}} \\
&= 0.983 \text{ (to 3 decimal places).}
\end{aligned}
$$

A few points to remember...

A few points to remember...

- If your calculated correlation coefficient does not lie between -1 and $+1$, you've done something wrong!

A few points to remember...

- If your calculated correlation coefficient does not lie between -1 and $+1$, you've done something wrong!
- Check that your correlation coefficient **agrees with your plot**

A few points to remember...

- If your calculated correlation coefficient does not lie between -1 and $+1$, you've done something wrong!
- Check that your correlation coefficient **agrees with your plot**
  - an "**uphill**" slope ties in with a **positive** correlation coefficient;

A few points to remember...

- If your calculated correlation coefficient does not lie between -1 and $+1$, you've done something wrong!
- Check that your correlation coefficient **agrees with your plot**
  - an "**uphill**" slope ties in with a **positive** correlation coefficient;
  - a "**downhill**" slope ties in with a **negative** correlation coefficient;

A few points to remember...

- If your calculated correlation coefficient does not lie between -1 and $+1$, you've done something wrong!
- Check that your correlation coefficient **agrees with your plot**
  - an "**uphill**" slope ties in with a **positive** correlation coefficient;
  - a "**downhill**" slope ties in with a **negative** correlation coefficient;
  - a **random scattering** of points indicates a correlation coefficient close to **zero**;

A few points to remember...

- If your calculated correlation coefficient does not lie between -1 and $+1$, you've done something wrong!
- Check that your correlation coefficient **agrees with your plot**
  - an "**uphill**" slope ties in with a **positive** correlation coefficient;
  - a "**downhill**" slope ties in with a **negative** correlation coefficient;
  - a **random scattering** of points indicates a correlation coefficient close to **zero**;
  - the closer the points lie to a straight line (either "uphill" *or* "downhill"), the closer to either $+1$ or -1 the correlation coefficient will be!

Simple linear regression

A correlation analysis helps to establish whether or not there is a linear relationship between two variables. However, it doesn't allow us to *use* this linear relationship.

Simple linear regression

A correlation analysis helps to establish whether or not there is a linear relationship between two variables. However, it doesn't allow us to *use* this linear relationship.

Regression analysis allows us to *use* the linear relationship between two variables. For example, with a regression analysis, we can predict the value of one variable given the value of another.

Simple linear regression

A correlation analysis helps to establish whether or not there is a linear relationship between two variables. However, it doesn't allow us to *use* this linear relationship.

Regression analysis allows us to *use* the linear relationship between two variables. For example, with a regression analysis, we can predict the value of one variable given the value of another.

To perform a regression analysis, we must assume that

A correlation analysis helps to establish whether or not there is a linear relationship between two variables. However, it doesn't allow us to *use* this linear relationship.

Regression analysis allows us to *use* the linear relationship between two variables. For example, with a regression analysis, we can predict the value of one variable given the value of another.

To perform a regression analysis, we must assume that

- the scatter plot of the two variables (roughly) shows a **straight line**, and

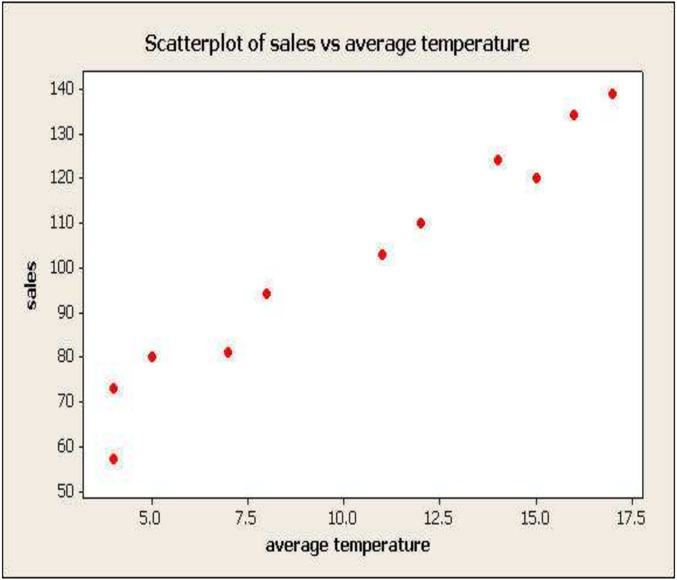
Simple linear regression

A correlation analysis helps to establish whether or not there is a linear relationship between two variables. However, it doesn't allow us to *use* this linear relationship.

Regression analysis allows us to *use* the linear relationship between two variables. For example, with a regression analysis, we can predict the value of one variable given the value of another.

To perform a regression analysis, we must assume that

- the scatter plot of the two variables (roughly) shows a **straight line**, and
- the **spread** in the Y–direction is roughly constant with X.


Scatterplot of sales vs average temperature
sales
140
130
120
110
100
90
80
70
60
50
5.0
7.5
10.0
12.5
15.0
17.5
average temperature

Look at the scatter plot of ice cream sales against temperature.

Look at the scatter plot of ice cream sales against temperature.

- A **"line of best fit"** can be drawn through the data;

Look at the scatter plot of ice cream sales against temperature.

- A **"line of best fit"** can be drawn through the data;
- This line could then be used to **make predictions** of ice cream sales based on temperature.

Look at the scatter plot of ice cream sales against temperature.

- A **"line of best fit"** can be drawn through the data;
- This line could then be used to **make predictions** of ice cream sales based on temperature.
- However, everyone's line is bound to be slightly different!

Look at the scatter plot of ice cream sales against temperature.

- A **"line of best fit"** can be drawn through the data;
- This line could then be used to **make predictions** of ice cream sales based on temperature.
- However, everyone's line is bound to be slightly different!
- And so everyone's predictions will be slightly different!

Look at the scatter plot of ice cream sales against temperature.

- A **"line of best fit"** can be drawn through the data;
- This line could then be used to **make predictions** of ice cream sales based on temperature.
- However, everyone's line is bound to be slightly different!
- And so everyone's predictions will be slightly different!
- The aim of regression analysis is to find the "best" line which goes through the data.

The regression equation

The regression equation

The **simple linear regression model** is given by

$$Y = \alpha + \beta X + \epsilon,$$

where

The regression equation

The **simple linear regression model** is given by

$$Y = \alpha + \beta X + \epsilon,$$

where

- Y is the **response variable** and

The regression equation

The **simple linear regression model** is given by

$$Y = \alpha + \beta X + \epsilon,$$

where

- Y is the **response variable** and
- X is the **explanatory variable**.

The regression equation

The **simple linear regression model** is given by

$$Y = \alpha + \beta X + \epsilon,$$

where

- Y is the **response variable** and
- X is the **explanatory variable**.
- α represents the **intercept** of the regression line (the point where the line "cuts" the Y–axis),

The regression equation

The **simple linear regression model** is given by

$$Y = \alpha + \beta X + \epsilon,$$

where

- Y is the **response variable** and
- X is the **explanatory variable**.
- α represents the **intercept** of the regression line (the point where the line "cuts" the Y–axis),
- β represents the **slope** of the regression line (i.e. how steep the line is), and

The regression equation

The **simple linear regression model** is given by

$$Y = \alpha + \beta X + \epsilon,$$

where

- Y is the **response variable** and
- X is the **explanatory variable**.
- α represents the **intercept** of the regression line (the point where the line "cuts" the Y–axis),
- β represents the **slope** of the regression line (i.e. how steep the line is), and
- ϵ is known as "**random error**"

The regression equation

The **simple linear regression model** is given by

$$Y = \alpha + \beta X + \epsilon,$$

where

- Y is the **response variable** and
- X is the **explanatory variable**.
- α represents the **intercept** of the regression line (the point where the line "cuts" the Y–axis),
- β represents the **slope** of the regression line (i.e. how steep the line is), and
- ϵ is known as "**random error**"

In practice, we assume ϵ is zero, and so the only things we need to find are α and β. But how?

- Remember, the aim of regression analysis is to find a line which goes through the **middle** of the data in the scatter plot, **closer to the points than any other line**;

- Remember, the aim of regression analysis is to find a line which goes through the **middle** of the data in the scatter plot, **closer to the points than any other line**;
- So the "best" line will minimise any gaps between the line and the data!

- Remember, the aim of regression analysis is to find a line which goes through the **middle** of the data in the scatter plot, **closer to the points than any other line**;
- So the "best" line will minimise any gaps between the line and the data!
- It turns out that the values of α and β which give this best line are

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} \qquad \text{and}$$

- Remember, the aim of regression analysis is to find a line which goes through the **middle** of the data in the scatter plot, **closer to the points than any other line**;
- So the "best" line will minimise any gaps between the line and the data!
- It turns out that the values of α and β which give this best line are

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} \quad \text{and}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Example: ice cream sales

For the ice cream sales data, we can find the "best" line using the formulae on the previous slide! Thus

Example: ice cream sales

For the ice cream sales data, we can find the "best" line using the formulae on the previous slide! Thus

$$\begin{aligned}\hat{\beta} &= \frac{S_{XY}}{S_{XX}} \\ &= \frac{1362}{250} \\ &= \mathbf{5.448},\end{aligned}$$

and

Example: ice cream sales

For the ice cream sales data, we can find the "best" line using the formulae on the previous slide! Thus

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{1362}{250} = \mathbf{5.448},$$

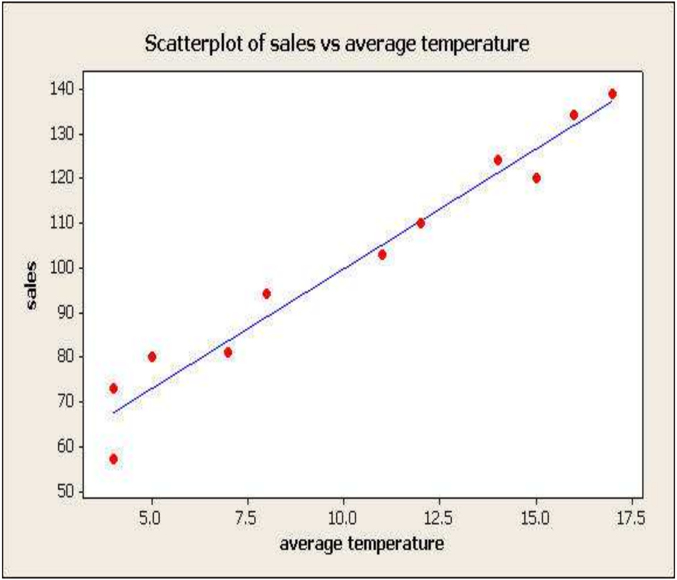
and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 100 - 5.448 \times 10 = \mathbf{45.52}.$$

Thus, the regression equation is

Thus, the regression equation is

$$\mathbf{Y} = \mathbf{45.52} + \mathbf{5.448X} + \epsilon$$



Scatterplot of sales vs average temperature
sales
140
130
120
110
100
90
80
70
60
50
5.0
7.5
10.0
12.5
15.0
17.5
average temperature

Making predictions

We can use our estimated regression equation to make **predictions** of ice cream sales based on average temperature.

For example: Predict monthly sales if the average temperature is 10°C.

For example: Predict monthly sales if the average temperature is 10°C.

We can use take a reading from our graph, or, more accurately, use our regression equation!

$$
\begin{aligned}
Y &= 45.52 + 5.448X \qquad \text{i.e.} \\
&= 45.52 + 5.448 \times 10 \\
&= 45.52 + 54.48 \\
&= 100,
\end{aligned}
$$

For example: Predict monthly sales if the average temperature is 10°C.

We can use take a reading from our graph, or, more accurately, use our regression equation!

$$\begin{aligned} Y &= 45.52 + 5.448X \qquad \text{i.e.} \\ &= 45.52 + 5.448 \times 10 \\ &= 45.52 + 54.48 \\ &= 100, \end{aligned}$$

i.e. if the average temperature is 10°C, we can expect sales of £100,000.