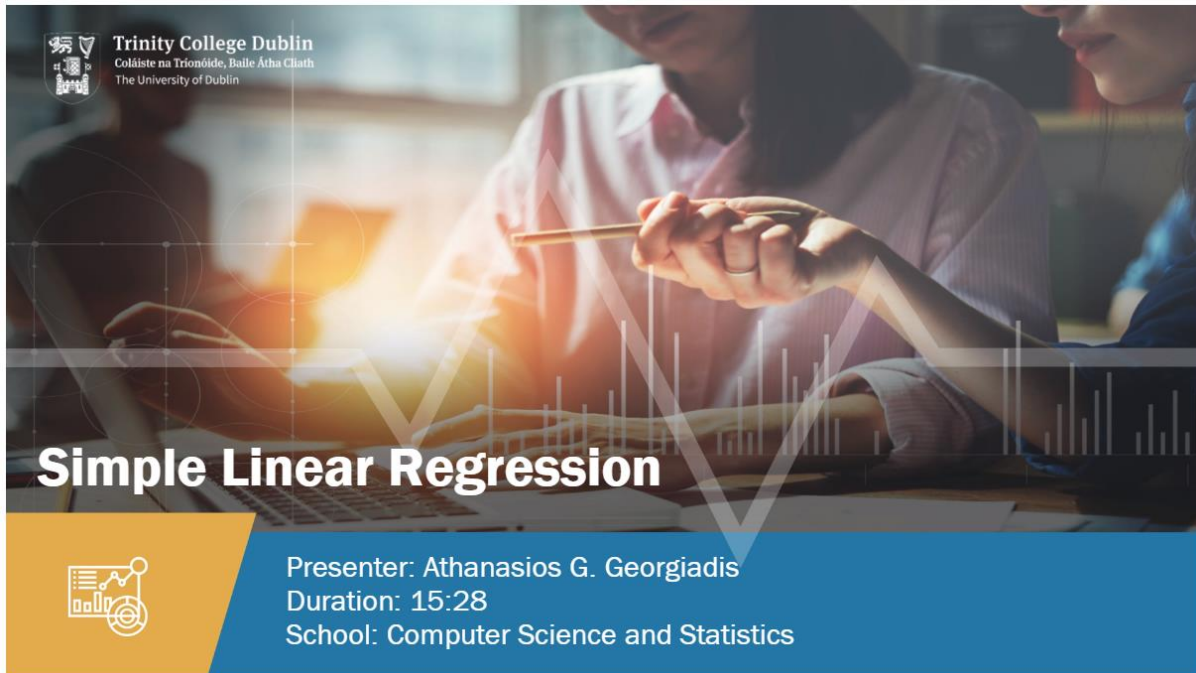




Simple Linear Regression

Slide 1: Introduction	2
Slide 2: What is Regression?	2
Slide 3: Simple Linear Regression with One Predictor Variable X.....	3
Slide 4: The Simple Linear Regression Model	4
Slide 5: Least Squares Line of Best Fit	5
Tab 1: Determining the Line of Best Fit	6
Slide 6: Height to Weight Regression Example.....	7
Slide 7: Entering the Data in R and Drawing the Scatterplot	8
Slide 8: The Regression Equation.....	9
Slide 9: Visualising the Regression	9
Slide 10: Predictions	10
Slide 11: Testing the Quality of the Fit of the Model	10
Slide 12: Poor R^2	11
Slide 13: Diagnostics	12
Tab 1: Outliers.....	12
Tab 2: High Leverage Points.....	16
Slide 14: Summary.....	19

Slide 1: Introduction



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Simple Linear Regression

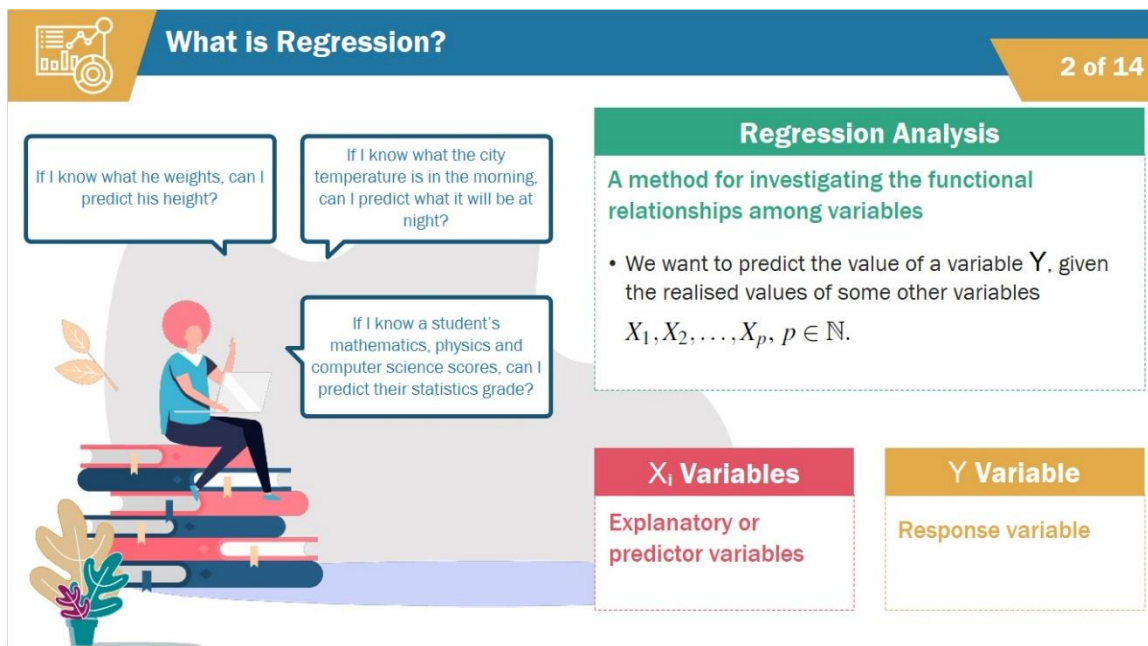
Presenter: Athanasios G. Georgiadis
Duration: 15:28
School: Computer Science and Statistics

Hello, and welcome. My name is Athanasios Georgiadis and I'm the instructor for this presentation on simple linear regression.

During this presentation we will explain what is meant by regression, focus on the simple linear regression, and elaborate on the principle of least squares.

We will present a representative example of linear regression and we will conclude with diagnostics.

Slide 2: What is Regression?



What is Regression?

2 of 14

If I know what he weights, can I predict his height?

If I know what the city temperature is in the morning, can I predict what it will be at night?

If I know a student's mathematics, physics and computer science scores, can I predict their statistics grade?

Regression Analysis

A method for investigating the functional relationships among variables

- We want to predict the value of a variable Y , given the realised values of some other variables $X_1, X_2, \dots, X_p, p \in \mathbb{N}$.

X_i Variables	Y Variable
Explanatory or predictor variables	Response variable

Let's start with some motivational questions:

If I knew the weight of a person, could I predict his height?

If I knew the temperature in a city at a certain time in the morning could I predict the temperature at night-time?

If I knew the grade of a Student in Mathematics, Physics and Computer Science, could I predict hers or his grade in Statistics?

Regression analysis is a method for investigating the functional relationship among variables.

In Regression analysis, we try to predict the value of a variable Y , given the realized values of some other variables X_1, X_2, \dots, X_p .

The X_i -variables are called the explanatory or predictor variables.

The Y -variable is called the response variable.

Slide 3: Simple Linear Regression with One Predictor Variable X



Simple Linear Regression With One Predictor Variable X

1 of 1

- The **regression of the random variable Y on the variable X** is the expected value of Y , when X takes a specific value x :
$$\mathbb{E}(Y|X = x)$$
- The regression of Y on X is called **linear** when it can be modelled as the equation of a line:
$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x.$$
- This is a simple regression because there is exactly one predictor variable X .


In this presentation we focus on the case when we have one predictor variable X .

The *regression of the random variable Y on the variable X* is the expected value of Y , when X takes a specific value x , and it is referred as a *simple regression*, because we have exactly one predictor variable X .

The regression of Y on X is called *linear* when it can be modelled as the equation of a line

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x.$$

Slide 4: The Simple Linear Regression Model



The Simple Linear Regression Model

4 of 14

Unknown Parameters $\beta_0 + \beta_1 X$

- For a simple linear regression model $E(Y|X = x) = \beta_0 + \beta_1 x$, the unknown parameters are:
 - The intercept, β_0
 - The slope, β_1
 - We need to estimate their values, based on our data.
- We collect data in pairs of observations of the variables Y and X :
 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), \quad n \in \mathbb{N}.$

Random Error e_i

- A simple linear regression model has a random error e_i
- Let Y_1, Y_2, \dots, Y_n be independent realisations of the random variable Y, observed at the values x_1, x_2, \dots, x_n of the predictor random variable X .
 - Then for every $i = 1, \dots, n$:
 $Y_i = E(Y|X = x_i) + e_i = \beta_0 + \beta_1 x_i + e_i.$
 - The e_i is called the random error in Y_i .
- Random error:
 - Is the variation that exists in Y_i due to random phenomena
 - Does not depend on any X and Y variables
 - Has an expectation of zero
 $E(e_i|X) = 0.$

For a simple linear regression model, the *unknown parameters* β_0 and β_1 are referred as the *intercept* and the *slope* of the model respectively.

These are unknown population parameters. We need to *estimate* their values based on our data. For this purpose, we collect data in pairs of observations of the variables X and Y:

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

for some data size n .


Let Y_1, Y_2, \dots, Y_n be independent realizations of the random variable Y, observed at the values x_1, x_2, \dots, x_n of the predictor random variable X. Then for every $i = 1, \dots, n$:

$$Y_i = E(Y | X = x_i) + e_i,$$

this e_i is called the *random error* in Y_i .

The random error represents the variation that exists in Y_i due to random phenomena that cannot be predicted or explained. The random error does not depend on any of X and Y variables and its expectation is zero.


Slide 5: Least Squares Line of Best Fit




Least Squares Line of Best Fit

5 of 14

- Let $y_i = \beta_0 + \beta_1 x_i + e_i$,
be the linear relationship between x_i and y_i , involving the errors e_i
- We want to find **estimators** $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 respectively,
such that the so-called model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, be "as close as possible" to y_i
- Above \hat{y}_i will be referred to as the **fitted value** of y_i and the entire line as **the line of best fit**.



Determining the Line of Best Fit

 Click the tab to learn more. Then, click Next to continue.

Let:

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

be the linear relationship between x_i and y_i , involving the errors e_i .

We aim to find *estimators* $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 respectively, such that the so-called fitted model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

is "as close as possible" to y_i . Above \hat{y}_i will be referred as the *fitted* value of y_i and the entire line as the *line of best fit*.

Click the tab to find out more about the process of determining the line of best fit. When you are ready, click next to continue.

Tab 1: Determining the Line of Best Fit

Determining the Line of Best Fit

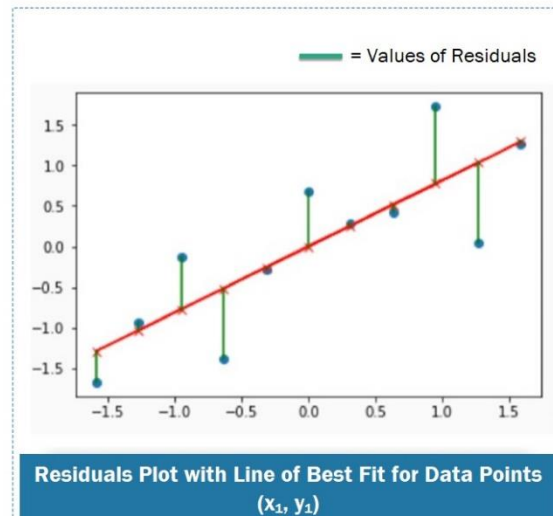
Defining the Residuals



- To determine the line of best fit, we define the residuals.

$$\hat{e}_i := y_i - \hat{y}_i$$

is defined as the difference between the actual and predicted values of Y .



For determining the line of best fit, we need to define the *residuals*. The residual \hat{e}_i is defined as the difference between the actual and the predicted values of Y .

In a hypothetical example, we draw the line of the best fit (in red) for some data-points (x_i, y_i) and mark the values of the residuals (in green).

Tab 1.1: Principle of Least Squares

Determining the Line of Best Fit

Principle of Least Squares



- Let $\hat{y}_i = b_0 + b_1 x_i$ equal an arbitrary fitted line.
- The **principle of least squares**:
 - Determines the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ as the values of b_0 and b_1 such that the **Residual Sum of Squares** can attain its minimum value

$$\text{RSS} := \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

Let $\hat{y}_i = b_0 + b_1 x_i$ an arbitrary fitted line.

The *Principle of Least Squares* determines the *least squares estimators*

$\hat{\beta}_0$ and $\hat{\beta}_1$ as the values of the arbitrary b_0 and b_1 such that the *Residual Sum of Squares* RSS, presented explicitly this equation can attain its minimum value.

Tab 1.2: Least Squares Line of Best Fit: Mathematical Equations

Determining the Line of Best Fit		
Least Squares Line of Best Fit: Mathematical Equations		
Precise Value of $\hat{\beta}_1$	Precise Value of $\hat{\beta}_0$	Least Squares Line of Best Fit
$\hat{\beta}_1 := \frac{SXY}{SXX} := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ <ul style="list-style-type: none"> Always derived directly by R 	$\hat{\beta}_0 := \bar{y} - \hat{\beta}_1 \bar{x}$ <ul style="list-style-type: none"> Always derived directly by R 	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

The precise values $\hat{\beta}_1$ and $\hat{\beta}_0$, that minimise RSS, can be found using mathematical techniques and are presented in these equations. We will always derive them directly by R language.

The *least squares line of best fit* takes the form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Slide 6: Height to Weight Regression Example

Height - Weight Regression Example			
6 of 14			
<ul style="list-style-type: none"> Let's look at an example demonstrating simple linear regression with one predictor variable X. 			
What was the Prediction Being Explored?	What was Collected?	What was Counted?	What were the Collected Pairs (x_i, y_i)?
<ul style="list-style-type: none"> The regression of the weight on the height of some population. 	<ul style="list-style-type: none"> A sample of $n = 15$ people. 	<ul style="list-style-type: none"> Height (in cm) Weight (in kg) 	<ul style="list-style-type: none"> (150,62), (155,70), (160,71), (165,77), (175,81), (170,80), (152,65), (156,71), (172,80), (174,79), (177,83), (180,85), (181,88), (184,90), (189,92).
<p>? Is it possible to predict the weight of any individual of some known height?</p>			


We will now look at an example that demonstrates simple linear regression with one predictor variable X .

Scientists are exploring the regression of the Weight on the Height of some population. In other words, they want to predict the Weight of any individual of some known Height.

They collected a sample of $n = 15$ people and they counted their Height (in cm) and Weight (in kilos). The collected pairs (x_i, y_i) are presented here:

If we want to establish whether we can make this weight – height prediction, we need to undertake the regression analysis!

Slide 7: Entering the Data in R and Drawing the Scatterplot



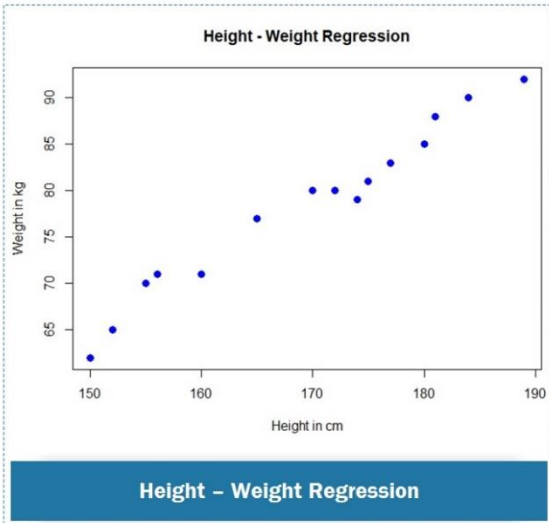
Entering Data in R and Drawing the Scatterplot

7 of 14

- We enter the data in R to create a scatterplot.

```
1 # Enter the data
2 x <- c(150, 155, 160, 165, 175, 170, 152, 156, 172, 174, 177,
3       180, 181, 184, 189)
4 y <- c(62, 70, 71, 77, 81, 80, 65, 71, 80, 79, 83, 85, 88, 90, 92)
5 # Visualize the Scatterplot
6 plot(x,y,col = "blue",main = "Height - Weight Regression",
       cex = 1.3, pch = 16, xlab = "Height in cm", ylab = "Weight in kg")
```

- There is a clear linear pattern between variables X and Y .



Height - Weight Regression

Creating the scatterplot is “Step 1” in regression analysis.

To obtain the scatterplot, we enter the data in R using the following command.

This visual representation of the data may indicate the existence of a linear (or other) relation between the variables.

Here we observe a clear linear pattern between the variables X which is the Height and Y which is the Weight.

Slide 8: The Regression Equation



The Regression Equation

8 of 14

- We can use “lm” (linear-model) from R to find the values of the $\hat{\beta}_i$ coefficients.

```
1 # Apply lm()  
2 relation <- lm(y ~ x)  
3 print(relation)
```

```
Call:  
lm(formula = y ~ x)  
  
Coefficients:  
(Intercept)          x  
-42.4025         0.7126
```

- The least squares line of best fit is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -42.4025 + 0.7126x$$

We can use “lm” (linear-model) command from R, as in the frame below, to find the values of the $\hat{\beta}_i$ coefficients.

Here the equation of the least squares line of best fit is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -42.4025 + 0.7126x.$$

Slide 9: Visualising the Regression

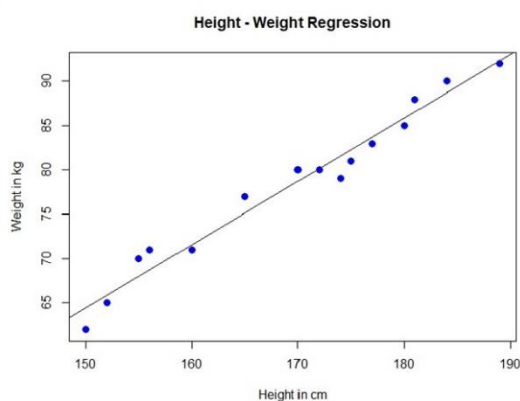


Visualising the Regression

9 of 14

- We can visualise the regression fit using the R code commands:


```
1 # Visualize the Regression Graphically  
2 plot(x, y, col = "blue", main = "Height - Weight Regression",  
3      abline(lm(y ~ x)), cex = 1.3, pch = 16, xlab = "Height in  
   cm", ylab = "Weight in kg")
```



Height - Weight Regression

We can visualise the regression fit using the commands given below.

Slide 10: Predictions



Predictions

10 of 14

- R allows us to predict the weight Y of a person of a given height.
 - For example, for a person of height $X = 178$ cm:


```
1 a <- data.frame(x = 178)
2 result <- predict(relation,a)
3 print(result)
```

1
84.44265
- The expected weight is around $\hat{y} \simeq 84.4\text{kg}$

R language allows us to further to predict the Weight Y of a person of a given Height.

For example, for a person of Height $X = 178\text{cm}$, the expected Weight \hat{y} is around 84.4kilos, as we can extract using this command in R.

Slide 11: Testing the Quality of the Fit of the Model



Testing the Quality of the Fit of the Model

11 of 14

- The **coefficient of determination R^2** is the most common measure for testing how good your model is.
 - This is defined by the proportion of the total variability explained by the regression model:
$$R^2 := \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} := \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$$


Note:

- $0 \leq R^2 \leq 1$
- R^2 is close to 1 in models that fit the data well.
- R^2 is close to 0 in models that poorly fit the data.

? In the height-weight example, is there a good fit?

- We can also use R to determine what is a good fit.

```
1 # Print R^2 from the summary of the relation
2 summary(relation)$r.squared
```
- Using R , $R^2 \simeq 97\%$.
 - 97% of the total variability is explained by the regression model, making it very satisfactory.

 Take time to view the information on this slide.

Our next aim is to test the quality of the fit of the model, as this will have an impact on how much we can trust our model.

The most common measure for test how good is your model is the *coefficient of determination R^2* . As you can see from this equation, this is defined by the proportion of the total variability explained by the regression model. Note that

R^2 is a number between 0 and 1.


For models that fit the data well, R^2 is close to 1.

For models that poorly fit the data, R^2 is close to 0.

On the Height-Weight example, using R , we find that R^2 is around 97%. This means that around 97% of the total variability is explained by the regression model. Which is very satisfactory!

Take time to view the information on this slide. When you are ready, click next to continue.

Slide 12: Poor R^2

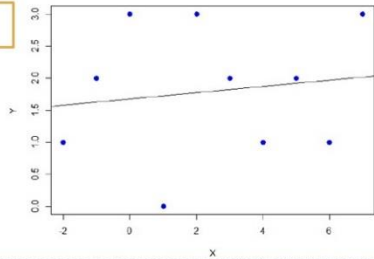


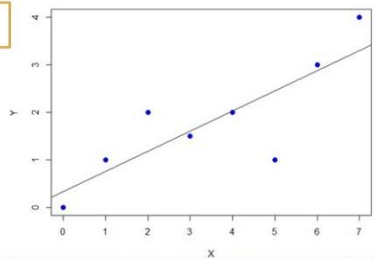
Poor R^2

12 of 14

- We can determine a poor fit using R^2 .
- R^2 evaluates the model, giving a percentage as the final answer.

! • A high R^2 is not necessarily good every single time.
• A low R^2 is not necessarily always bad.

Dataset A
A "poor fit"
• $R^2 \approx 2\%$


Dataset B
A "good, but imperfect fit"
• $R^2 \approx 67\%$


Let's take a look at the following dataset A in the graph, which presents $R^2 \approx 2\%$. This agrees completely with the feeling of a "poor fit" that we all have when we see the scatterplot and the line of best fit.


Finally, dataset B with $R^2 \approx 67\%$ is presented in the graph on the right.

We observe a fit that still looks good, but not perfect. It is clearly better than the example with $R^2 \approx 2\%$ and worse than the Height-Weight example, where $R^2 \approx 97\%$.

R^2 evaluates the model, giving a percentage as final answer.

As an important final remark here, we mention that sometimes a high R^2 is not necessarily good every single time and a low R^2 is not necessarily always bad.

Slide 13: Diagnostics



Diagnostics


13 of 14

Outliers

High Leverage Points

Introduction

- There are several possible reasons for breaking the fit of a model.
- The data may contain influential points that strongly affect the model.
- These may present as measurement errors because of unusual observations that are unlikely to happen.

 Click each tab to learn more. Then, click Next to continue.

There are several possible reasons for breaking the fit of a model. We proceed to explore them now.

Our data may contain *influential points* that strongly affect the model. Such points may pop up as measurement errors because of very unusual observations that are unlikely to happen.

Let us focus in two of them: the *outliers* and the points of *high leverage*.

Click each tab to learn more. When you are ready, click next to continue.

Tab 1: Outliers

Outliers

What is an Outlier?

An outlier:

- Is a point which differs significantly from other observations
- Usually presents an extreme value on the corresponding residual
- This corresponds to a break of the zero-expectation on the values of random errors, which was an assumption for the model to be valid.

If carrying out descriptive statistical analysis of residuals' values:

- Any point with a residual out of the interval $[Q_1 - 3D, Q_3 + 3D]$, where Q_1 is the first quartile, Q_3 the third quartile and $D := Q_3 - Q_1$, the interquartile range is an outlier

We start with the Outliers.

A point is referred as an *outlier* when it differs significantly from other observations. Such a point usually presents an extreme value on the corresponding residual.

This corresponds to a break of the zero-expectation on the values of random errors, which was an assumption for the model to be valid.

An empirical rule asserts that when we do the descriptive statistical analysis of the residuals' values, then any point with a residual out of the interval $[Q_1 - 3D, Q_3 + 3D]$, where Q_1 is the first quartile, Q_3 is the third quartile and $D := Q_3 - Q_1$, the interquartile range, is an outlier. In different books, slightly different intervals may appear.

Tab 1.1: Outlier Example: Higher-Lower Temperature Regression

Outliers

Outlier Example: Higher-Lower Temperature Regression

- Scientists studied the lower and higher temperature within $n = 10$ days in a specific city.
- The values are listed in couples (x_i, y_i) :
(0,10), (0,11), (1,13), (2,13), (2,12),
(3,20), (3,14), (4,14), (5,15), (5,16)
- The information is entered in R.
 - We expand the regression of the higher temperature (Y) on the lower temperature (X).

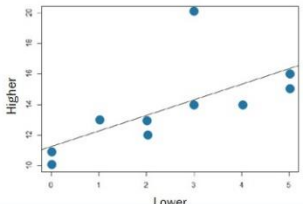
1 `l <- c(0, 0, 1, 2, 2, 3, 3, 4, 5, 5)`

2 `h <- c(10, 11, 13, 13, 12, 20, 14, 14, 15, 16)`

3 `plot(l, h, col = "blue", main = "Lower - Higher temperature Regression",`

4 `abline(lm(h~l)), cex = 1.3, pch = 16, xlab = "Lower", ylab = "Higher")`

• We use R to see the scatterplot together with the fitted line.



Lower-Higher Temperature Regression

Let us see an example.

Scientists studied the Lower and Higher temperature within $n = 10$ days in a specific City. The values they found are listed as the couples below.

We assist them by entering everything in R and expanding the regression of the Higher temperature (Y) on the Lower temperature (X).

We use R to see the scatterplot together with the fitted line.

Tab 1.2: Residuals

Outliers

Residuals

- We can detect the outlier with residuals' values or their plots.
- We obtain the descriptive statistics metric for the residuals using R.

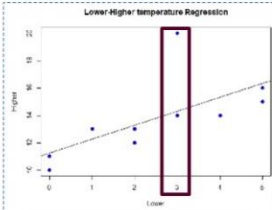
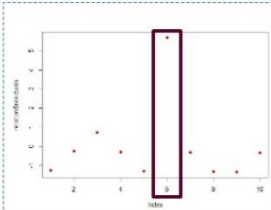
```
1 relation <- lm(h~l)
2 print(summary(relation))
```

Call:
lm(formula = h ~ l)

Residuals:					
Min	1Q	Median	3Q	Max	
-1.3410	-1.2836	-0.3246	-0.2672	5.6918	

- Any residual with a value larger than $Q_3 + 3D = 2.782$, corresponds to an outlier.

? What is the point that corresponds to this extreme residual value?

- We ask R to list the residuals.

```
1 round(relation$residuals,2)

> round(relation$residuals,2)
      1      2      3      4      5      6      7      8      9     10
-1.26 -0.26  0.72 -0.29 -1.29  5.69 -0.31 -1.32 -1.34 -0.34
```

- We conclude that the sixth data point, (3, 20) is the outlier.

The outlier appears clearly in the graph (for $x = 3$), but it could also be detected with the values of the residuals or their plots.

We obtain the descriptive statistics metric for the residuals using R.

R gives the first and third quartile. Under the empirical rule, any residual with a value larger than $Q_3 + 3D$, which here equals to 2.782, corresponds to an outlier.

Here there exists such a value, with a residual around 5! What is the point that corresponds to this extreme residual value? We ask R to list the residuals and we conclude that the 6th data point; (3, 20) is the outlier.

The outlier corresponding point appears clearly in the plot of the residuals too.

Tab 1.3: Is There Any Treatment for Outliers?

Outliers

Is There Any Treatment for Outliers?

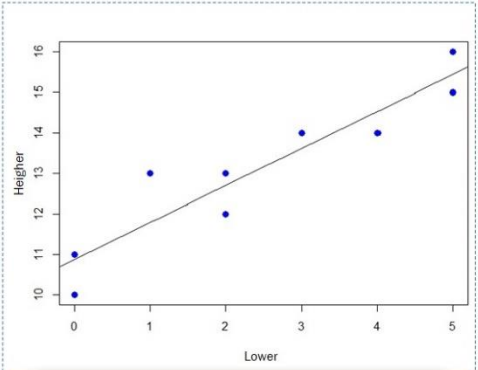
■ ■ ■ ■ ■

? How do we deal with outliers?

- We can carefully exclude it from the dataset and re-analyse the restricted data.

```
1 l <- c(0, 0, 1, 2, 2, 3, 4, 5, 5)
2 h <- c(10, 11, 13, 13, 12, 14, 14, 15, 16)
3 plot(l, h, col = "blue", main = "Lower - Higher temperature Regression ",
4      abline(lm(h~l)), cex = 1.3, pch = 16, xlab = "Lower", ylab = "Heigher")
```

- We remove the point and re-apply the usual commands in R.
 - We now have a nice fit.



Lower-Higher Temperature Regression

Is there any recommended treatment for dealing with outliers?

When an outlier is identified, we may (carefully) exclude it from the dataset and re-analyse, as seen here. We remove the point and reapply the usual commands in R.

Now we have a nice fit.

Tab 1.4: The New Fit: Regression Equation and R^2

Outliers

The New Fit: Regression Equation and R^2

■ ■ ■ ■ ■

```
1 relation <- lm(h~l)
2 print ( relation )
```

Coefficients:
(Intercept) 1
10.8824 0.9118

- The new data, excluding the outlier, is used to obtain the regression equation:
 $\text{Higher} = 10.88824 + 0.9118 \text{ Lower}$

- R^2 is around 87% which is very satisfactory.

```
1 summary ( relation )$r.squared
```

[1] 0.8696833

We refit and the regression equation is obtained by the new data that excludes the outlier as in this first equation. We can see that R^2 is around 87% which is very satisfactory.

Tab 1.5: Plot of the Residuals

Outliers

Plot of the Residuals

■ ■ ■ ■ ■

- Residuals are also very satisfactory.

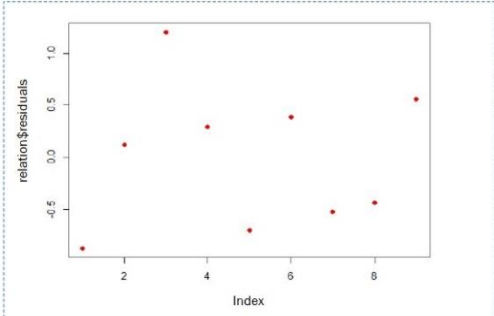
```
1 round(relation$residuals,2)
```

	1	2	3	4	5	6	7	8	9
	-0.88	0.12	1.21	0.29	-0.71	0.38	-0.53	-0.44	0.56

- This can be verified by their plot:

```
1 plot(relation$residuals, pch = 16, col = "red")
```

Plot of Residuals



Note:

- Zero expectation and x, y dependence will be discussed in other sessions.

! Residuals must be very sparse and their values must be distributed randomly around zero without following any pattern.

The same holds for the residuals, as it can be observed by their values or their plot.

Note that the residuals must be very sparse, and their values must be distributed randomly around zero without following any pattern (as it happens above). We will discuss the zero expectation and x, y dependence in more detail in the next sessions.

Tab 2: High Leverage Points

High Leverage Points

What Constitutes a High Leverage Point?


■ ■ ■ ■ ■

High leverage points:

- Are data points which exercise considerable influence on the fitted model
- They appear for unusually large or small values of X.

Cook's distance:

- Measures how "negative influence" exercises a high leverage point
- For large values of Cook's distance, we suggest excluding the point from the study.
 - The cut-off is defined to be 1.
 - If Cook's distance of a point extends beyond 1, we remove it from the study.

 There is a link to information on Cook's distance in the Extend section of your session homepage.

Another reason for breaking the fit of a model may be because of high leverage points.

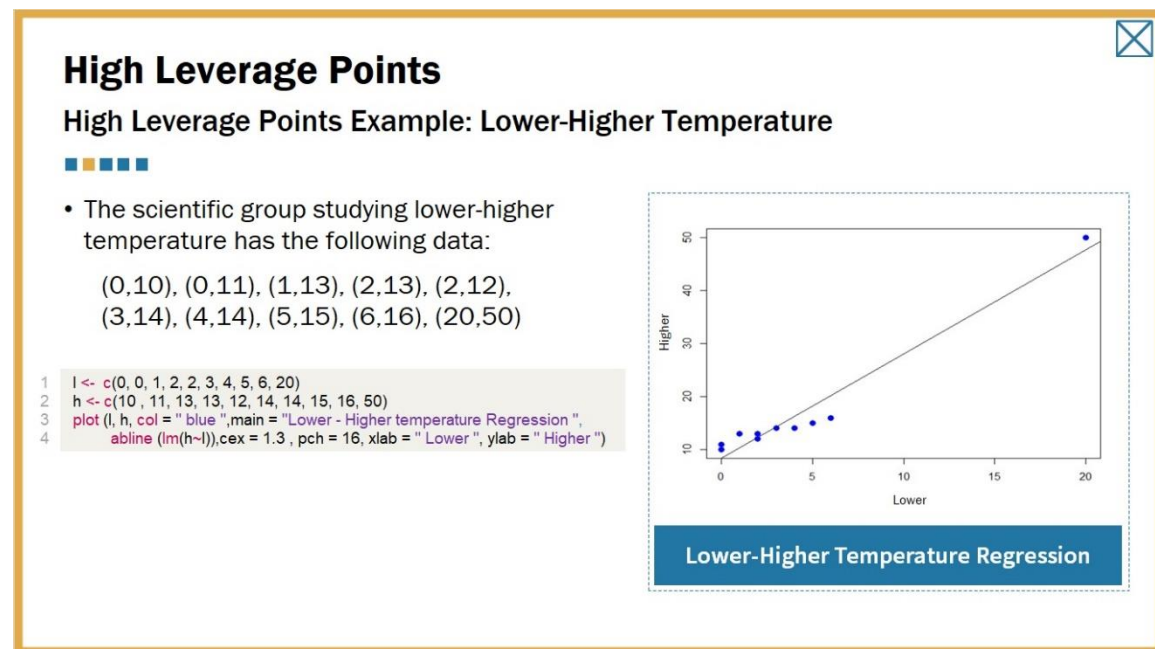
Data points which exercise considerable influence on the fitted model are called *high leverage points*.

Such points appear for unusually large or small values of x . Cook's distance measures how "negative influence" exercises a high leverage point.

There is a link to more information on Cook's distance in the Extend section of your session homepage.

For large values of Cook's distance, it is suggested to exclude the point from the study. The cut-off is defined to be 1. If Cook's distance of a point extends 1, we remove it from the study.

Tab 2.1: Example: Lower-Higher Temperature



Let's take a look at an example.

Assume that the scientific group which studies the Lower-Higher temperature problem, has the following data.

We draw the least squares line (with the usual commands in R) and we observe that it is "dictated" by the point with $x = 20$.

Tab 2.2: Cook's Distance

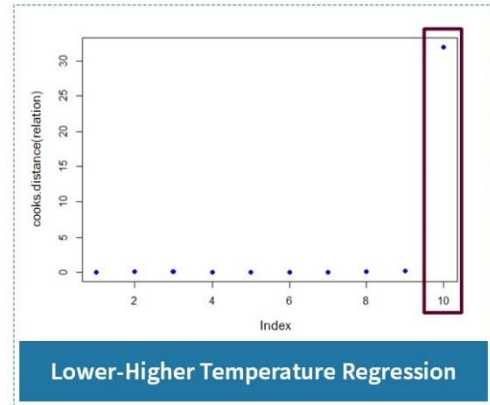
High Leverage Points

Cook's Distance



- Cook's distance justifies the "bad influence" of the specific point.

```
1 plot(cooks.distance(relation), pch = 16, col = "blue")
```



Cook's distance justifies the "bad influence" of the specific point, as it can be observed by the graph obtained by R, since the value of Cook's distance is huge in this point.

Tab 2.3: Is There Any Treatment?

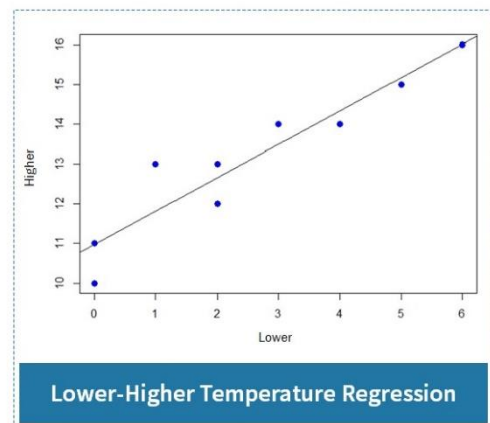
High Leverage Points

Is There Any Treatment?



- We can delete the influential point and re-fit.
 - This improves the fit.

```
1 l <- c(0, 0, 1, 2, 2, 3, 4, 5, 6)
2 h <- c(10, 11, 13, 13, 12, 14, 14, 15, 16)
3 plot(l, h, col = "blue", main = "Lower - Higher temperature Regression",
4      abline(lm(h~l)), cex = 1.3, pch = 16, xlab = "Lower", ylab = "Higher")
```

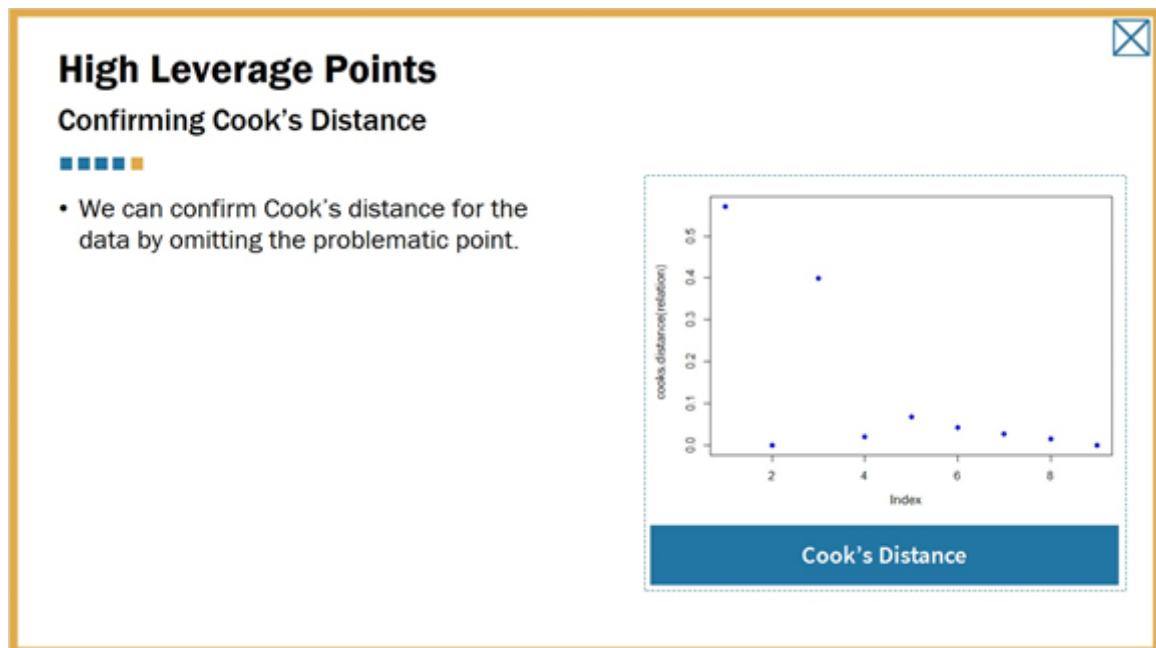


Is there any treatment for this high leverage point?

We delete the influential point from the dataset and re-fit:


It is far better now!

Tab 2.4: We Confirm Cook's Distance



In addition, we can also confirm Cook's distance for the data by omitting the problematic point.

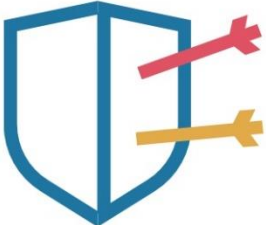
Slide 14: Summary



Summary

14 of 14

- Having completed this presentation, you should now be able to:
 - Fit a simple linear regression model
 - Evaluate the fit
 - Detect influential points and propose a cure



Developed by Trinity Online Services CLG with the School of Computer Science and Statistics, Trinity College Dublin, The University of Dublin

Having completed this presentation, you should now be able to:

- Fit a simple linear regression model,
- Evaluate the fit, and
- Detect influential points and propose a cure.