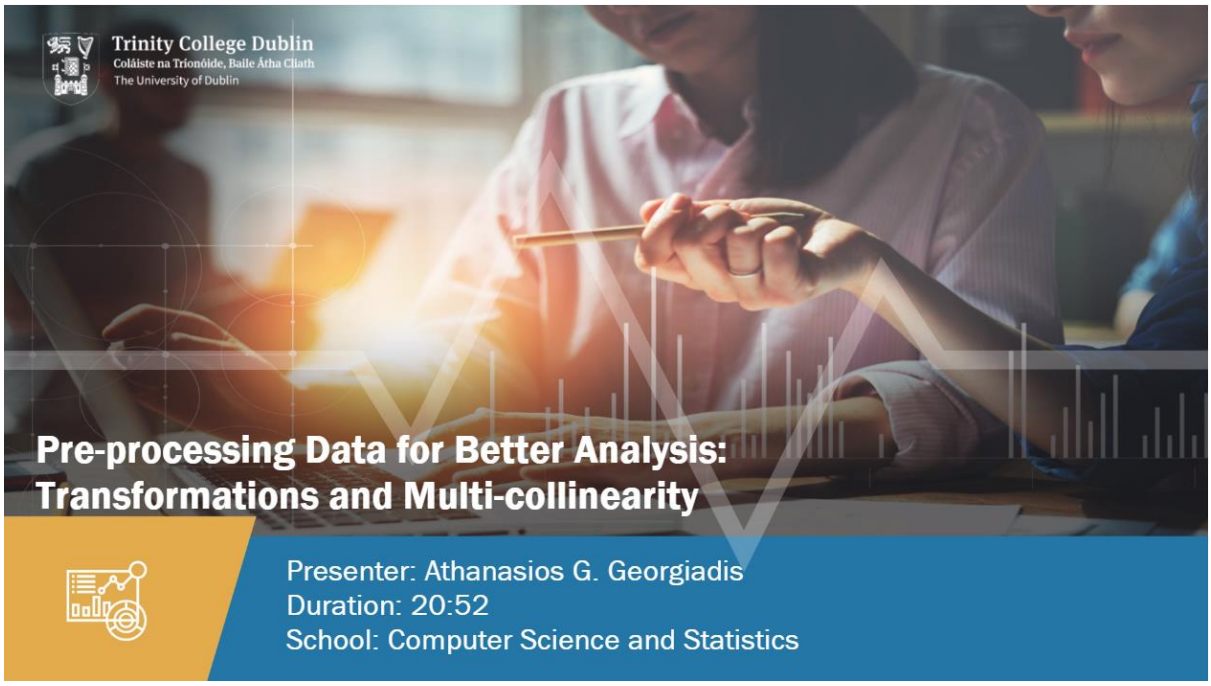
Pre-processing Data for Better Analysis: Transformations and Multi-collinearity

Slide 1: Introduction

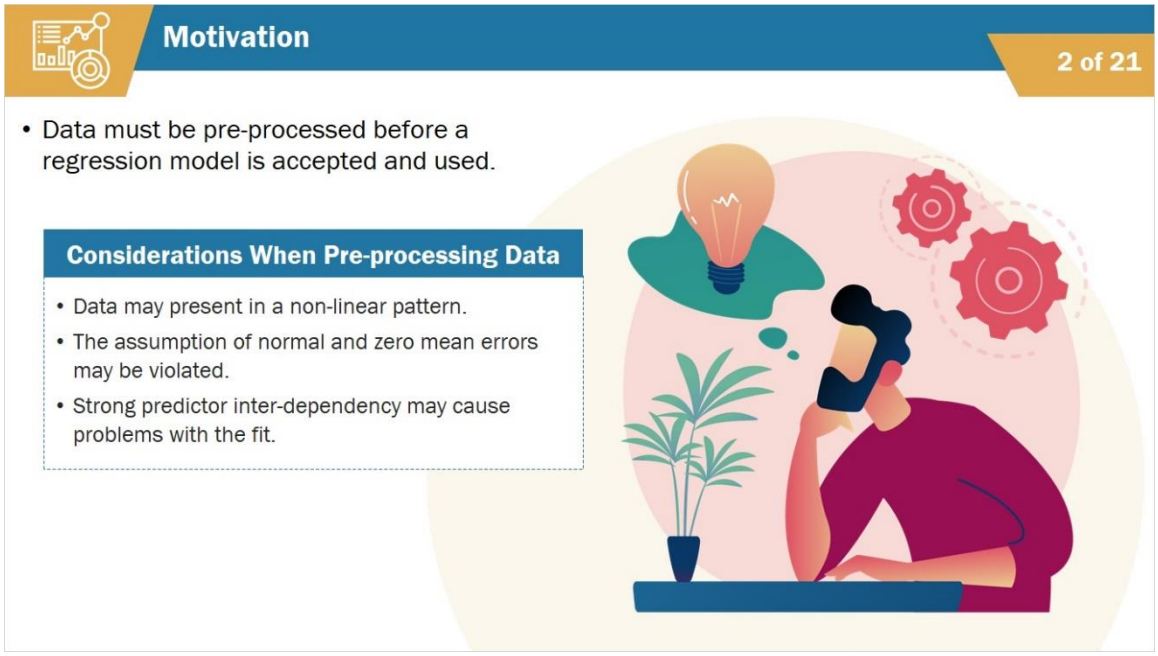


Hello and welcome to this presentation entitled **Pre-processing data for better analysis: transformations and multi-collinearity**. My name is Athanasios Georgiadis and I will lead this presentation.

We will look at:

- The motivation for this presentation
- Transforming predictors
- Transforming the response, and
- Multi-collinearity

Slide 2: Motivation



Before we accept a regression model and use it, we have to pre-process our data.

The data may present a non-linear pattern (as happened for the example in polynomial regression that we studied previously).

The assumption of normal and zero means errors may be violated.

The predictors may be strongly dependent on each other and this could cause problems with the fit.

In this presentation, we will see how we must act in these scenarios.

Slide 3: Section 1: Transforming Predictors

Slide 4: Scenario: Demonstration Transforming Predictors

Scenario: Demonstrating Transforming Predictors

4 of 21

Transforming Predictors

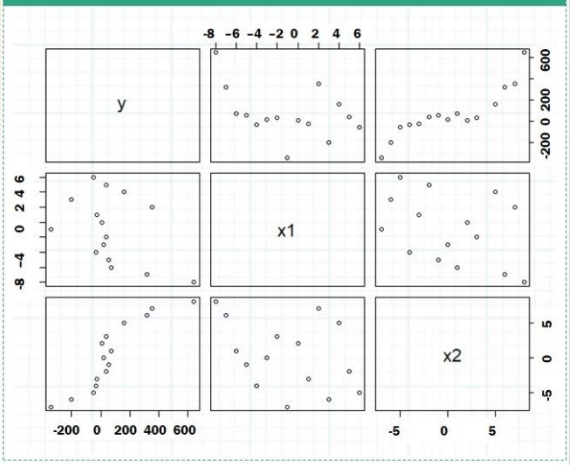
- Transforming predictors involves using a function of the predictor, instead of the predictor itself.
  - Instead of X_i consider using X_i^k, $1/X_i$, $log(X_i)$ etc.

- Enter the following dataset in R, with two predictors (X_1, X_2) and a response variable Y:

```
x1=c( -4 , -3 , -2 , -1 ,0 ,1 ,2 ,3 ,4 ,5 ,6 , -5 , -6 , -7 , -8)
x2=c( -4 ,0 ,3 , -7 ,2 , -3 ,7 , -6 ,5 , -2 , -5 , -1 ,1 ,6 ,8)
y=c( -30 ,20 ,35 , -340 ,10 , -25 ,350 , -200 ,160 ,40 , -52 ,55 ,
75 ,315 ,640)
ex1data = data . frame (y,x1 ,x2)
plot ( ex1data [, c("y", "x1", "x2")])
```

Take time to view the information on this slide.

Scatterplot



Transforming predictors involves using a function of the predictor instead of the predictor itself. For example, instead of using X_i, you could use a either a power of it X_i^k, a reciprocal, $1/X_i$, a logarithm $log(X_i)$, or any other function that could be helpful. Let's work through an example so you can see when and how transforming predictors is necessary.

We enter the following dataset in R, with two predictors - X_1 and X_2 - and a response variable, Y.

As always, we plot the data.

Slide 5: Fit a Multiple Linear Regression Model

Fit a Multiple Linear Regression Model

5 of 21

- Apply a multiple linear regression model:

```
lmex1 =lm(y~x1+x2 , data = ex1data )
summary ( lmex1 )
```

- Output:

Residuals:

Min	1Q	Median	3Q	Max
-153.184	-65.713	8.787	55.552	215.259

Coefficients:

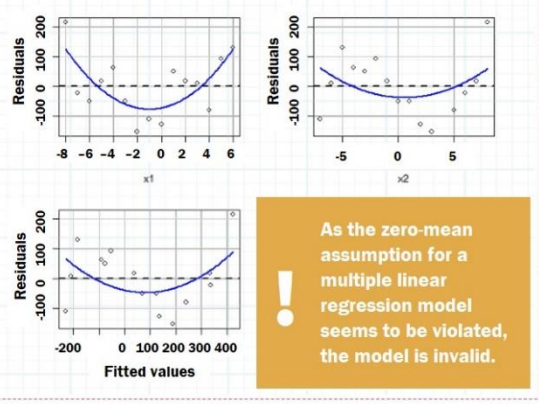
Min	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	54.203	28.898	1.876	0.0852
x1	-4.971	7.021	-0.708	0.4925
x2	41.346	6.498	6.363	3.49e-05***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 109 on 12 degrees of freedom
Multiple R-squared: 0.8112 Adjusted R-squared: 0.7797

Plotting the Residuals

```
library ( car)
residualPlots (lmex1 , ylab =" Residuals ", tests =F)
```



As the zero-mean assumption for a multiple linear regression model seems to be violated, the model is invalid.

We apply a multiple linear regression model obtaining the following summary table.

Careful reading of the output highlights the following:

The fit presents a very high $R^2 \simeq 81\%$.

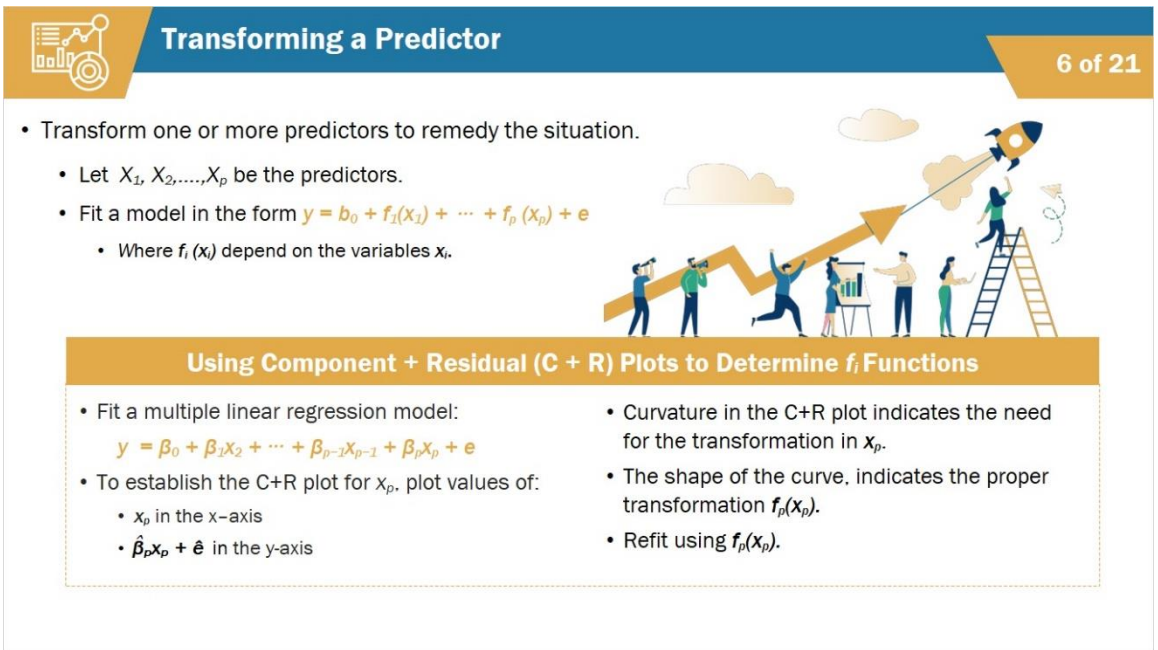
The coefficient $\hat{\beta}_1$ of the X_1, is not significant. In contrast $\hat{\beta}_2$ is highly significant.

There are very large (absolute) values on the residuals. Let us plot residuals against the values of the two predictors and the fitted values using the following command:

The plots are displayed on the screen.

The residuals present curvature. The zero-mean assumption for a multiple linear regression model, seems to be violated. The linear regression model is invalid.

Slide 6: Transforming a Predictor



A potential remedy in such a situation is to transform one or more predictors.

So, formally, let $X_1, X_2, \ldots, X_p$ be the predictors. We try fitting a model in the form:

$y = b_0 + f_1(x_1) + \ldots\ldots + f_p(x_p) + e$ where $f_i(x_i)$ are functions depending on the variable x_i.

But, how do we determine these f_i functions?

Click the tab for details of the C+R plots method. When you are ready, click "Next" to continue.

The answer to the question is from Component + Residuals plots (C+R plots).

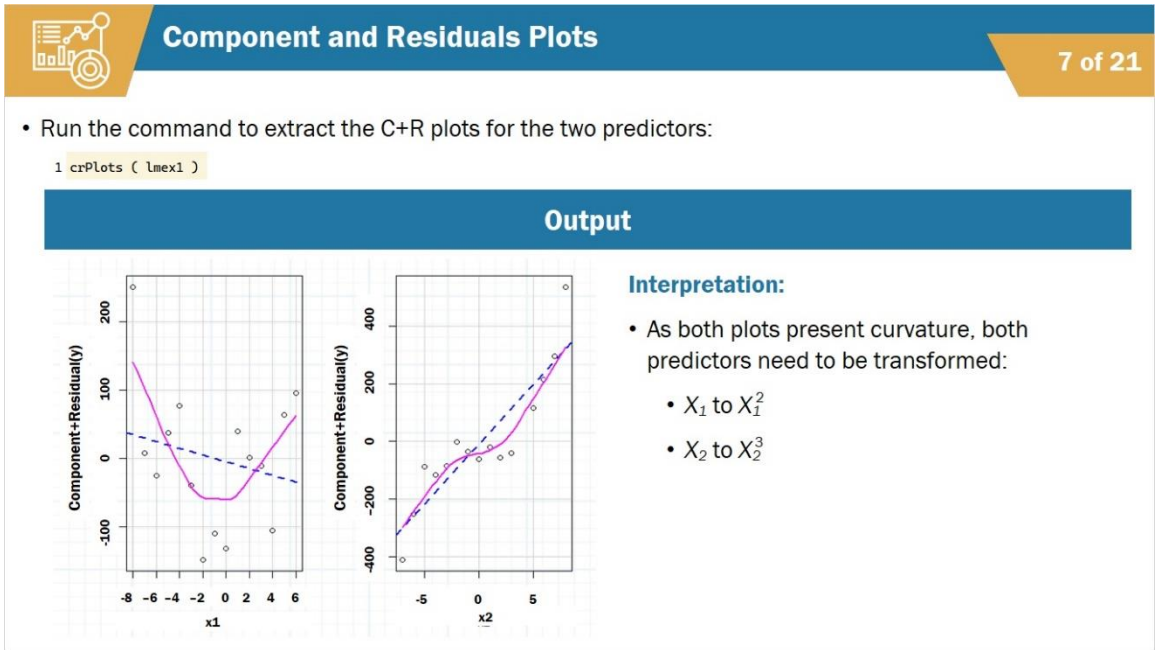
We first fit a multiple linear regression model as in the equation displayed here.

We then plot the values of xp in the x -axis and in the y -axis, the values of the component $\hat{\beta}$pxp, plus the residuals $\hat{e}$;. This is the C+R plot for the variable x_p.

Curvature in the C+R plot indicates the need for transformation in x_p. The shape of the curve, indicates the proper transformation $f_p(x_p)$.

We re-fit, using the transformed predictor $f_p(X_p)$.

Slide 7: Component and Residuals Plot



Let us see the use of C+R plots in our example.

Run the following command and extract C+R plots for the two predictors:

Let us interpret the output of the C+R plots obtained:

- Both plots present curvature. This means that we have to transform both predictors.
- The shape of the first plot reminds us of a parabola. This indicates that X_1 needs to be transformed as X_1^2 .
- The shape of the second plot indicates a transformation of the type X_2^3 .

Slide 8: Model with Transformed Variables

Model with Transformed Variables

8 of 21

- Run the following command to fit the model with the transformed variables:

```
lmex2 =lm(y~I(x1 ^2) +I(x2 ^3) ,data = ex1data )
summary ( lmex2 )
```

- Output:

Residuals:

Min	1Q	Median	3Q	Max
-3.1541	-0.9196	-0.1972	1.0269	4.8829

Coefficients:

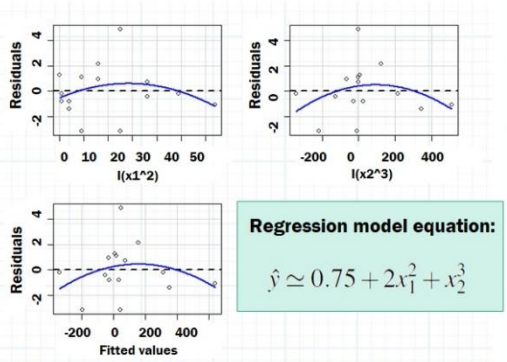
Min	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.749503	0.860322	0.871	0.401
I(x1^2)	2.014654	0.035446	56.837	5.81e-16***
I(x2^3)	0.998738	0.003275	304.921	<2e-16***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 2.18 on 12 degrees of freedom
Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999

Plotting the Residuals

```
library ( car)
residualPlots (lmex1 , ylab =" Residuals ", tests =F)
```



Regression model equation:

$$\hat{y} \simeq 0.75 + 2x_1^2 + x_2^3$$

We fit the model with the transformed variables as below:

We derive the following output.

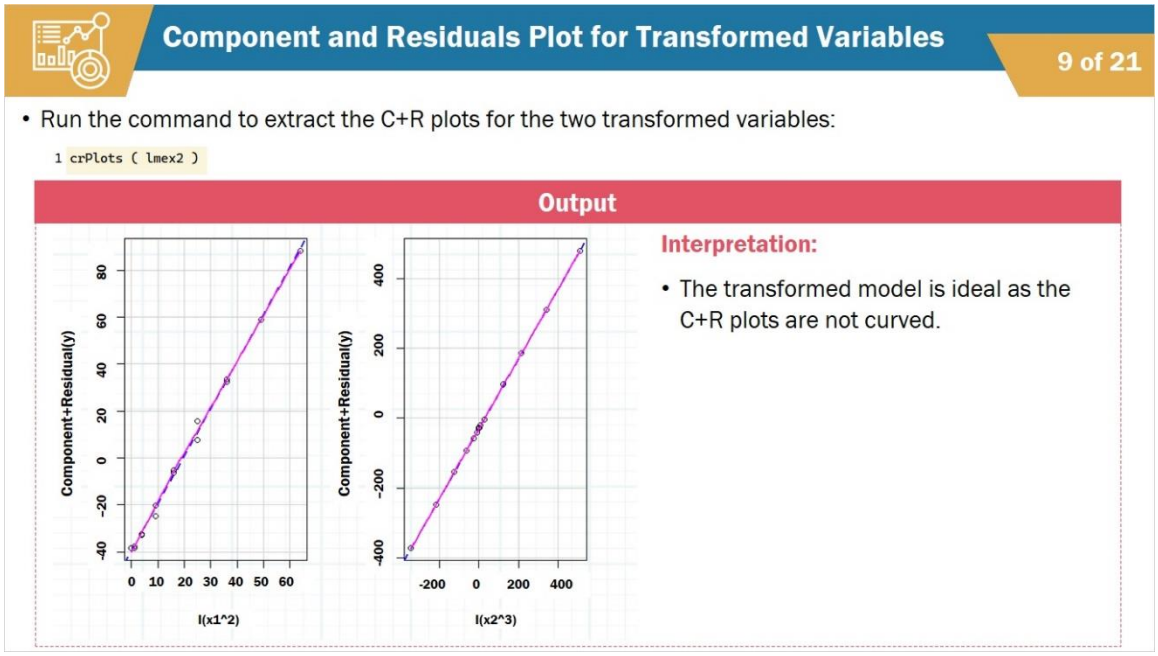
Interpreting the output of the newly obtained model:

- $R^2 \simeq$ 99.99% and the regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are highly significant!
- The values of the residuals are much smaller than before.
- The regression model equation is as follows:
  - $\hat{y} \simeq 0.75 + 2x_1^2 + x_2^3$.
- As the intercept is a bit small with a large standard error, it is not significant. We'd better exclude it from the model.

We plot the residuals versus the transformed variables and the fitted values.

The residuals of the new model are almost ideally distributed. They still look to be a bit curved, but with very small values and light curvature.

Slide 9: Component and Residuals Plot for Transformed Variables



We further extract the C+R plots of the transformed variables resulting in the following graphs:

C+R plots for the transformed variables are not curved at all. The transformed model is ideal.

Slide 10: Section 2: Transforming the Response

Slide 11: Scenario: Demonstrating Response Variable Transformation

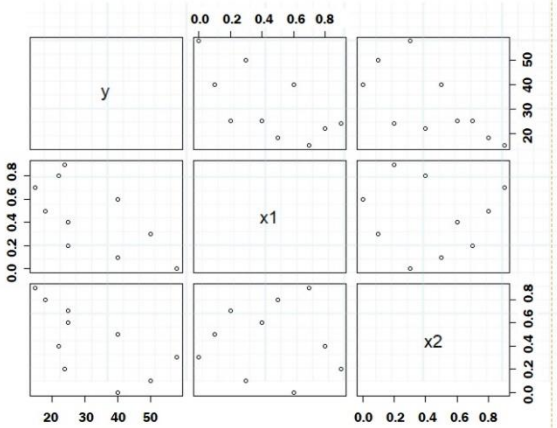
Scenario: Demonstrating Response Variable Transformation

11 of 21

- It can be helpful to transform the response variable rather than the predictors.
- Enter the dataset in R:

```
x1=c (0 ,0.1 ,0.2 ,0.3 ,0.4 ,0.5 ,0.6 ,0.7 ,0.8 ,0.9)
x2=c (0.3 ,0.5 ,0.7 ,0.1 ,0.6 ,0.8 ,0 ,0.9 ,0.4 ,0.2)
y=c (58 ,40 ,25 ,50 ,25 ,18 ,40 ,15 ,22 ,24)
ex2 . data = data . frame (y,x1 ,x2)
plot ( ex2 . data [, c ("y", "x1", "x2")])
```

Scatterplot



In some cases, it would be helpful to transform the response variable instead of the predictors. Let us study this scenario now.

We enter the following dataset in R, with two predictors (X_1,X_2) and a response variable Y and we extract their plot.

Slide 12: Create a Multiple Linear Regression Model

Apply a Multiple Linear Regression Model

12 of 21

- Apply a multiple linear regression model:

```
lmex2 =lm(y~x1+x2 , data = ex2 . data )
summary ( lmex2 )
```

- Output:

Coefficients:

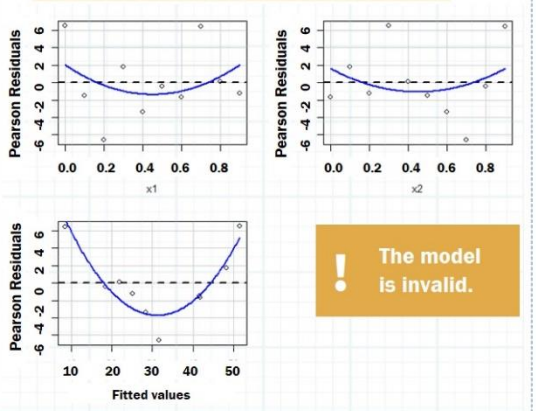
Min	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	61.434	3.617	16.986	6.01e-07***
x1	-32.864	5.092	-6.454	0.000349***
x2	-33.212	5.092	-6.523	0.000327***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 4.621 on 7 degrees of freedom
Multiple R-squared: 0.9202 Adjusted R-squared: 0.8975

Residuals

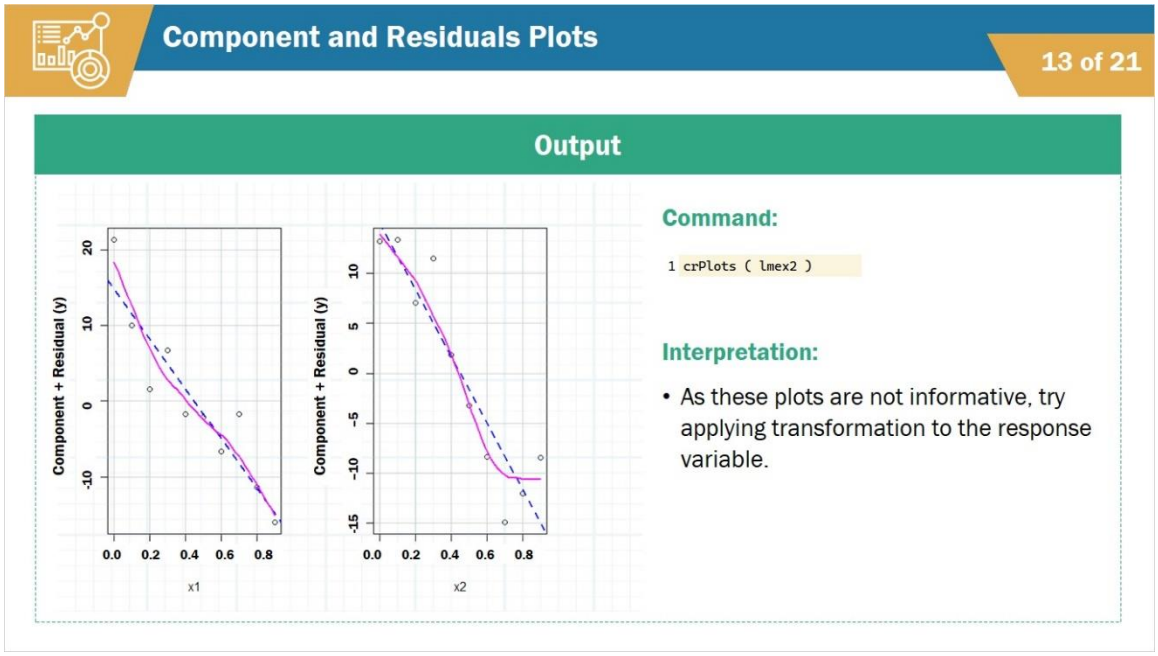
```
library ( car)
residualPlots (lmex2 , ylab =" Residuals ", tests =F)
```



We apply the multiple linear regression model with the usual command which presents high R^2 and significance in all coefficients as it appears in the summary table.

The model is invalid as you can see by the curvature of the residuals.

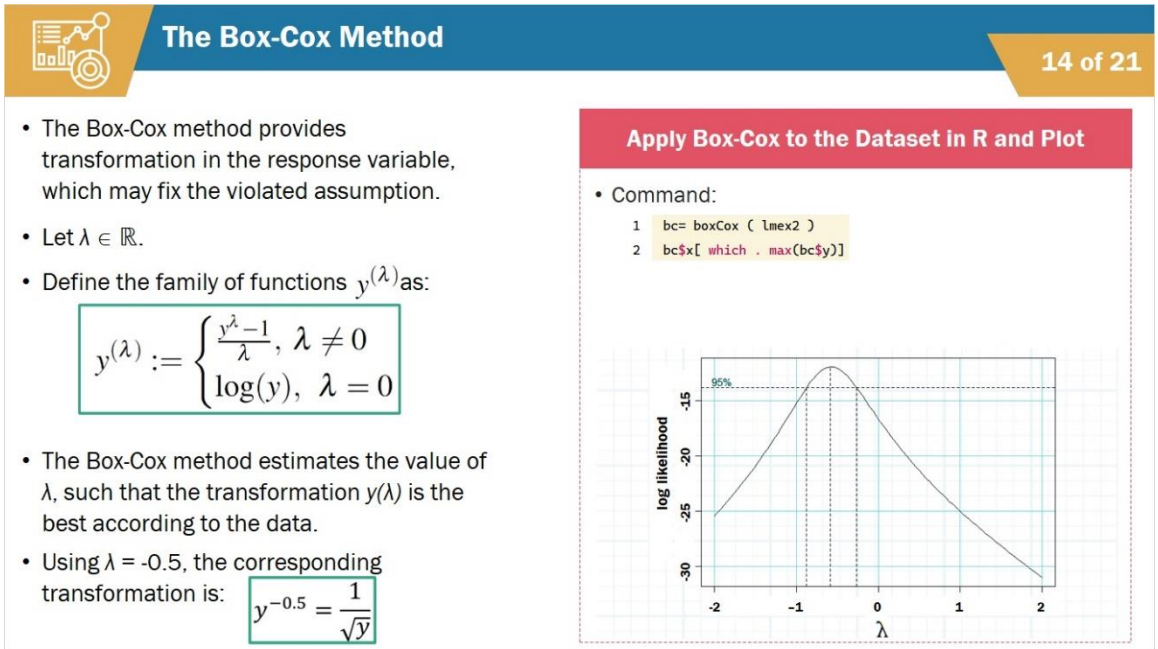
Slide 13: Component and Residuals Plots



The C+R plots here, are not that informative.

In such a case, we may try to apply a transformation to the response variable. The next method is the most commonly accepted one used for transforming the response.

Slide 14: The Box-Cox Method



The Box-Cox method provides the transformation in the response variable, which may fix the violated assumption.

Let $\lambda \in \mathbb{R}$. We define the family of functions $y^{(\lambda)}$: as below.

The Box-Cox method (which works via the likelihood), estimates the value of λ, such that the transformation $y^{(\lambda)}$ is the best according to the data.

Let us apply Box-Cox to our dataset. We use the following command in R.

R returns the center of the confidence interval for the parameter λ;

Here, it is around -0.59 and the plot of the log-likelihood, together with a 95% confidence interval is displayed on the slide.

From the plot, we can see the possible values of λ that could be used. Here we would prefer to use λ = -0.5 for simplicity, which corresponds to the transformation:

$y^{-0.5} = \frac{1}{\sqrt{y}}$ - which looks quite familiar.

We use this transformation and refit.

Slide 15: Refitting Transformation using the Box-Cox Method

Refitting Transformation Using the Box-Cox Method

15 of 21

- Run the following command in R:

```
lmboxcox =lm(y^( -1/2)~x1+x2 , data = ex2 . data )
summary ( lmboxcox )
```

- Output:

Residuals:

Min	1Q	Median	3Q	Max
-0.0055686	-0.0022233	-0.0007266	0.0023293	0.0060085

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(intercept)	0.101475	0.003167	32.04	7.46e-09***
x1	0.090435	0.004459	20.28	1.77e-07***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.004046 on 7 degrees of freedom
Multiple R-squared: 0.9926, Adjusted R-squared: 0.9905

Regression Equation

$$y^{1/2} \simeq 0.1 + 0.1x1 + 0.1x2$$

Or

$$y \simeq \frac{100}{(1 + x_1 + x_2)^2}$$

We run the next command in R, resulting in the ideal fit appearing in the summary table.

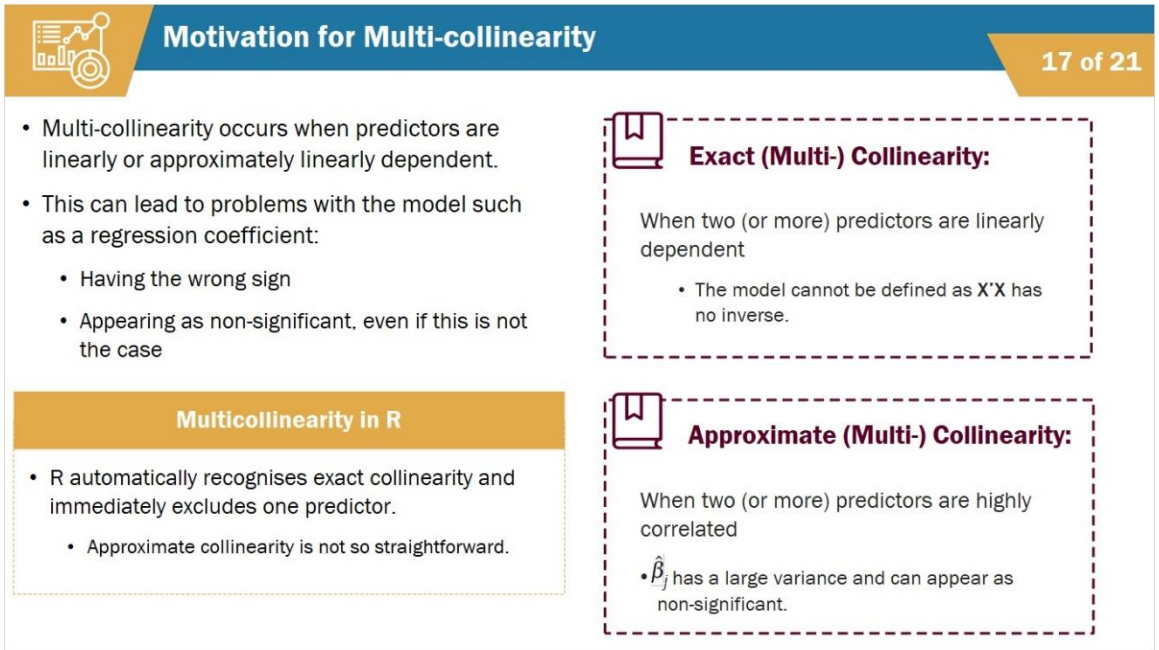
The new model presents $R^2 \simeq 99.26\%$, highly significant coefficients and almost zero residuals!

The regression equation is as follows: $y^{-1/2} \simeq 0.1 + 0.1\, x_1 + 0.1\, x_2$

or equivalently $y \simeq \frac{100}{(1+x_1+x_2)^2}$.

Slide 16: Section 3: Multi-collinearity

Slide 17: Motivation for Multi-collinearity



Assume that we are working on a model with predictors $X_1, X_2, \ldots, X_p$. In practice, the predictors are often linearly dependent or approximately linearly dependent. Such a situation is referred as multi-collinearity.

This could cause problems for the obtained model. For example, a regression coefficient could end up with the wrong sign or may appear to be non-significant, even if this is not the case.

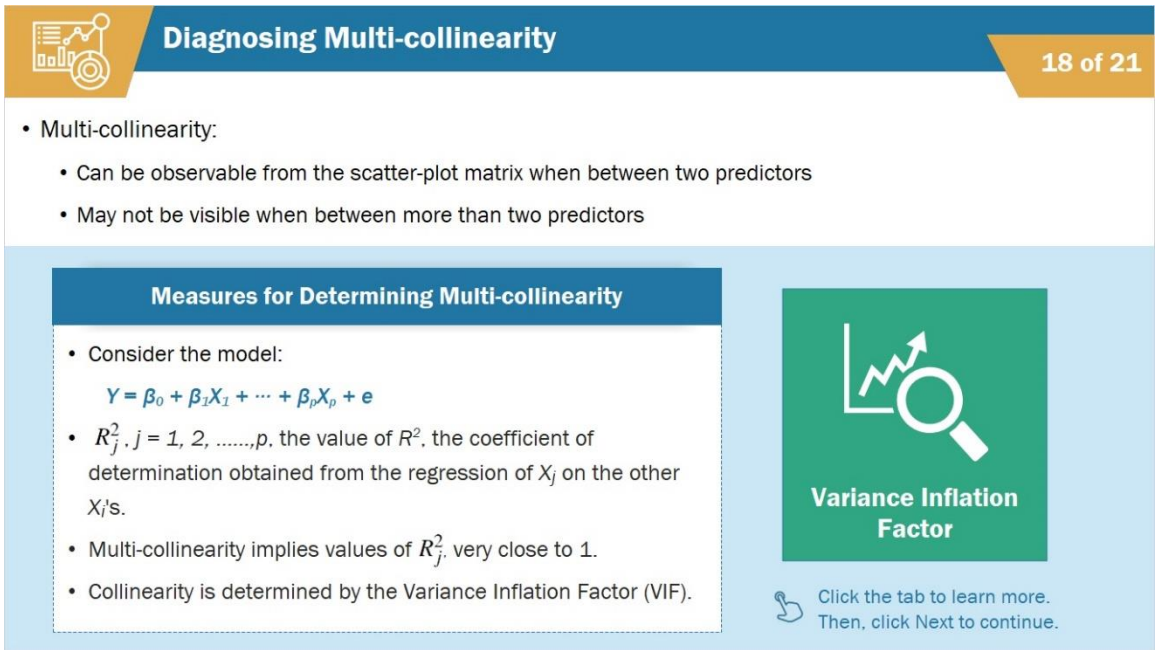
When two (or more) predictors are linearly dependent, we say that we have exact (multi-) collinearity. In this case, the model cannot be defined. The mathematical reason is that the matrix X'X has no inverse, so $\hat{\beta}$ is not defined

When two (or more) predictors are highly correlated, that is approximately linearly dependent, we say that we have approximate (multi-)collinearity. In this instance, $\hat{\beta}_j$ has large variance and can appear as non-significant.

R automatically recognises exact collinearity and immediately excludes one of the predictors.

For approximate collinearity, the situation is not so straightforward and demands careful handling.

Slide 18: Diagnosing Multi-collinearity



Let us look at how we diagnose multi-collinearity.

In simple cases, especially when it is between only two predictors, the collinearity may be already observable from the scatter-plot matrix. But when the multi-collinearity is between more than two predictors, it may not be visible.

We need measures for determining the collinearity. More formally: Consider the model

$Y = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_p X_p + e.$

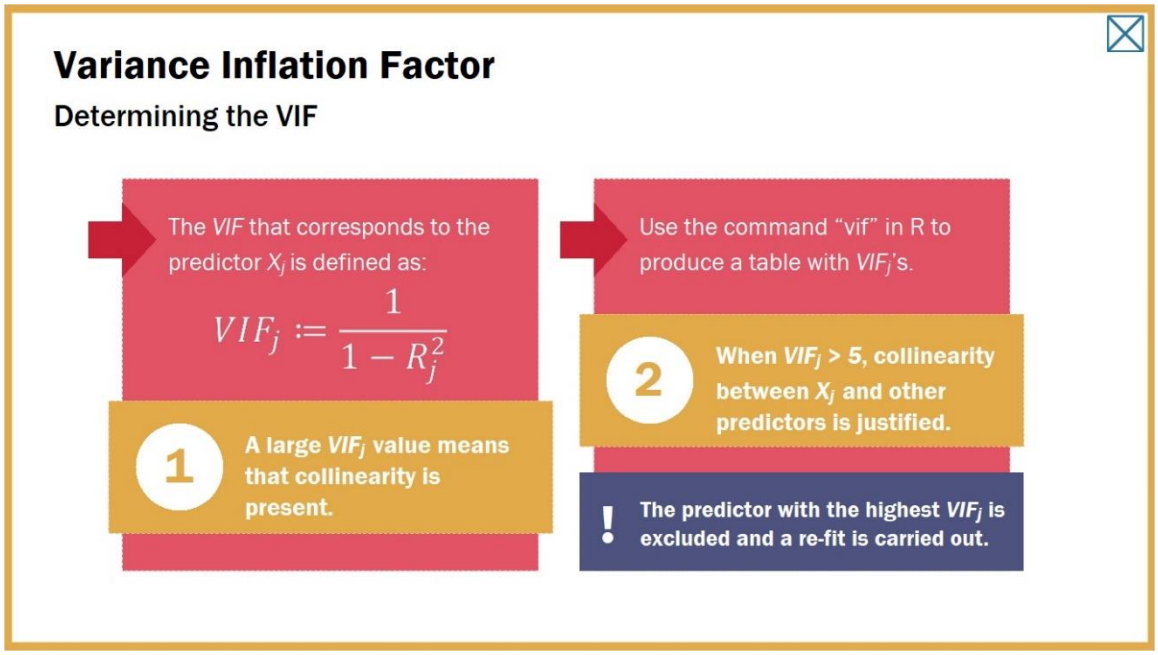
We denote this by R_j^2, j = 1, 2,.....,p, the value of R^2 , the coefficient of determination obtained from the regression of X_j on the other X_i 's. In other words, the percentage of variability explained by the regression of X_j on the other predictors.

Multi-collinearity implies values of R_j^2 , very close to 1, and this can be shown algebraically and lead to high variance on $\hat{\beta}_j$. The latter may imply non-significance on $\hat{\beta}_j$.

The measure for determining collinearity is the so-called Variance Inflation Factor (VIF).

Click the tab to learn more about the VIF. When you are ready, click “Next” to continue.

Tab 1: The Variance Inflation Factor



The VIF that corresponds to the predictor X_j, is defined as

$VIF_j := \frac{1}{1-R_j^2}$.

A large value of VIF_j means that the collinearity is present. As a consequence, the coefficient $\hat{\beta}_j$ can be just poorly estimated.

When you run the command “vif”, R produces a table with VIF_j 's,

When $VIF_j>5$, we justify the collinearity between X_j and other predictors.

In practice, we start by excluding the predictor with the highest VIF_j and re-fit.

Slide 19: Worked Example to Determine Multi-collinearity

Worked Example to Determine Multi-collinearity 19 of 21

Scatterplot and Summary | Check for Correlation | Obtain VIF | Exclude a Predictor

Introduction

- A company is interested in modelling their employees' performance in terms of sales, based on four skills.
- Their grades are the variables X_1, X_2, X_3, X_4.
- Y = The performance in terms of sales

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

Dataset of *n*=10 employees:

X1	X2	X3	X4	Y
10	50	60	30	211
20	30	100	20	262
30	90	80	60	344
40	10	10	20	78
50	40	90	50	320
60	100	30	80	307
70	70	70	70	345
80	60	50	70	302
90	20	40	60	267
100	80	20	90	302

Let us see all these elements presented in a worked example.

A company is interested in modelling the performance in terms of sales of its employees based on their skills. The company evaluates its employees on four skills and their grades (from 0 to 100) are the variables X_1,X_2,X_3,X_4 . Let Y be the performance in terms of sales. In a sample of n=10 employees, we have the following data:

Click the tabs to follow the steps required to determine multi-collinearity. When you are ready click "Next" to continue.

Tab 1: Scatterplot and Summary

Scatterplot and Summary (1/2)

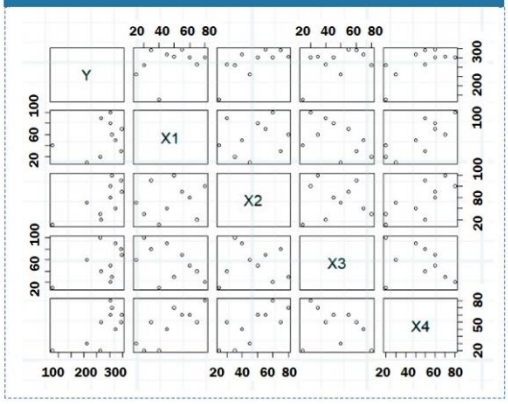
Command and Output

- Enter the data in R.
- Obtain the scatter-matrix using the following command:

```
dataset1 = data . frame (Y,X1 ,X2 ,X3 ,X4)
pairs ( dataset1 )
lm=lm(Y~., data = dataset1 )
summary (lm)
```

There is no collinearity between any two predictors graphically.

Scatter-matrix



We enter the data in R and act as always: We check the scatter-matrix and fit the multiple linear regression model.

From the scatter-matrix, we do not observe any collinearity between two predictors graphically.

Tab 1.1: Summary Observations

Scatterplot and Summary (2/2)

Summary Observations

- Output:

Coefficients:

Min	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.0387	11.8708	0.677	0.5284
X1	−0.3059	0.5514	- 0.555	0.6030
X2	−0.1615	0.4820	−0.335	0.7511
X3	1.9252	0.1080	17.823	1.02e-05 ***
X4	3.3742	0.9485	3.557	0.0163 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.739 on 5 degrees of freedom
Multiple R-squared: 0.9948,Adjusted R-squared: 0.9906

The fit appears in the summary table.

Only the predictors X_3 and X_4 appear to be significant. Also, the estimated values $\hat{\beta}_1$ and $\hat{\beta}_2$ appear to be negative, something that looks strange, given that we are talking about skills.

Tab 2: Check for Correlation

Check for Correlation

Obtain and Analyse the R Output

- To check for potential correlations between predictors, run the following command in R:

```
cor(dataset1)
```

- Output

	Y	X1	X2	X3	x4
Y	1.0000000	0.3293370	0.682132436	0.485266163	0.6921221
x1	0.3293370	1.000000	0.175757576	-0.490909091	0.7829501
x2	0.6821324	0.1757576	1.000000000	1.000000000	-0.2908100
x3	0.4852662	-0.4909091	-.0.006060606	1.000000000	-0.2908100
x4	0.6921221	0.7829501	0.738210084	-0.290810033	1.0000000

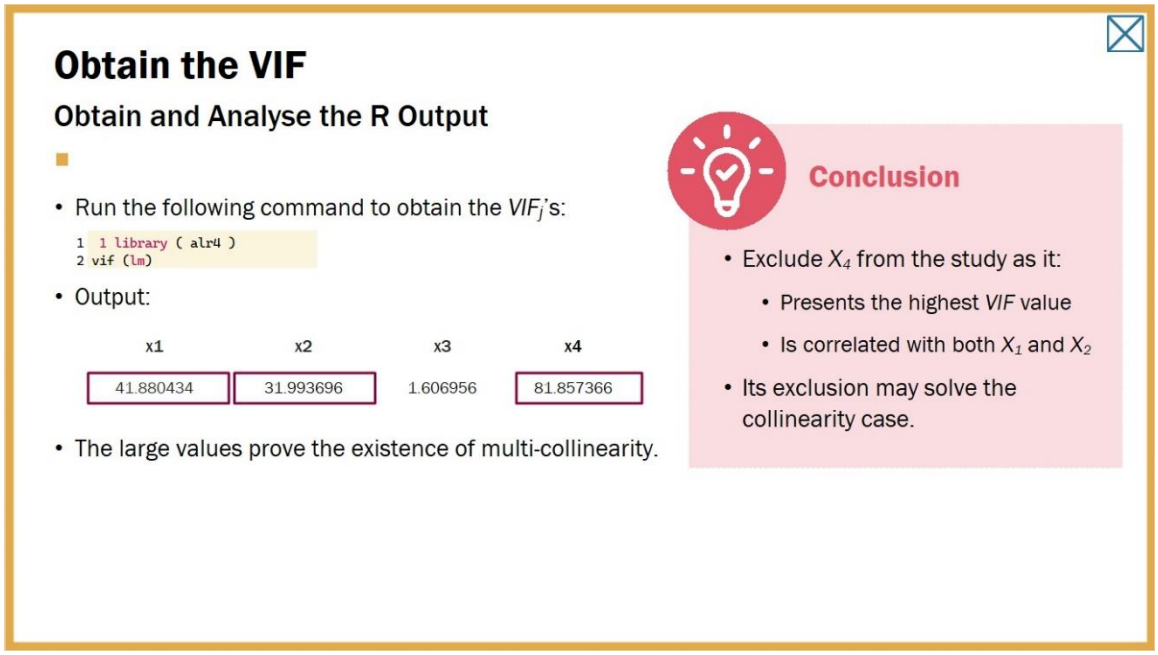
We check for potential correlations between the predictors. We run the next command in R.

The obtained table presents the correlations between the variables.

Here we observe that X4 is highly correlated with X1 and X2.

On the other hand, X1 and X2, present a very low level of correlation between each other.

Tab 3: Obtain VIF



Run the following command to obtain the VIF_j 's.

R returns the VIF_j's as in the next list.

The huge values of VIF_1, VIF_2 and VIF_4 (recall that the cut-off is the value 5 of VIF_j) prove the existence of multi-collinearity. We will exclude one of them.

As X_4 presents the highest value of VIF and, as we saw earlier, is correlated with both X_1 and X_2 , let's exclude it from the study. This exclusion may solve the collinearity case.

Tab 4: Exclude a Predictor

Exclude a Predictor (1/2)

Refit and Output

- Refit excluding X_4 using the command:

```
lm2 =lm(Y~X1+X2+X3 , data = dataset1 )
summary ( lm2)
```

- **Output:**

Residuals:

Min	1Q	Median	3Q	Max
-10.682	-9.440	-1.541	6.912	21.027

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−14.0292	17.3614	-0.808	0.45
X1	1.6234	0.1712	9.484	7.83e-05***
X2	1.5249	0.1491	10.225	5.10e-05***
x3	2.0850	0.1685	12.373	1.703-05***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.28 on 6 degrees of freedom
Multiple R-squared: 0.9815,Adjusted R-squared: 0.9723

Conclusion

- Always scan the predictor set to detect potential collinearities.

We refit as below excluding X_4 to get the following summary table.

Now all three predictors are significant meaning that Multi-collinearity was responsible for the bad fit we noticed before.

As a conclusion: we always have to scan the predictors' set to detect potential collinearities!

Tab 4.1: Fitting a Model Without an Intercept

Exclude a Predictor (2/2)

Fitting a Model Without an Intercept

- As the intercept suffers from low significance, fit the model without the intercept using the command:

```
lm3 =lm(Y~X1+X2+X3 +0, data = dataset1 )
summary ( lm3)
```

- Output:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
X1	1.5246	0.1168	13.05	3.61e-06 ***
X2	1.4869	0.1380	10.78	1.30e-05 ***
x3	1.9816	0.1069	18.54	3.29e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.94 on 7 degrees of freedom
Multiple R-squared: 0.9985,Adjusted R-squared: 0.9979

Conclusion

- Use the following model:

$$\hat{Y} \simeq 1.52X_1 + 1.49X_2 + 2X_3$$

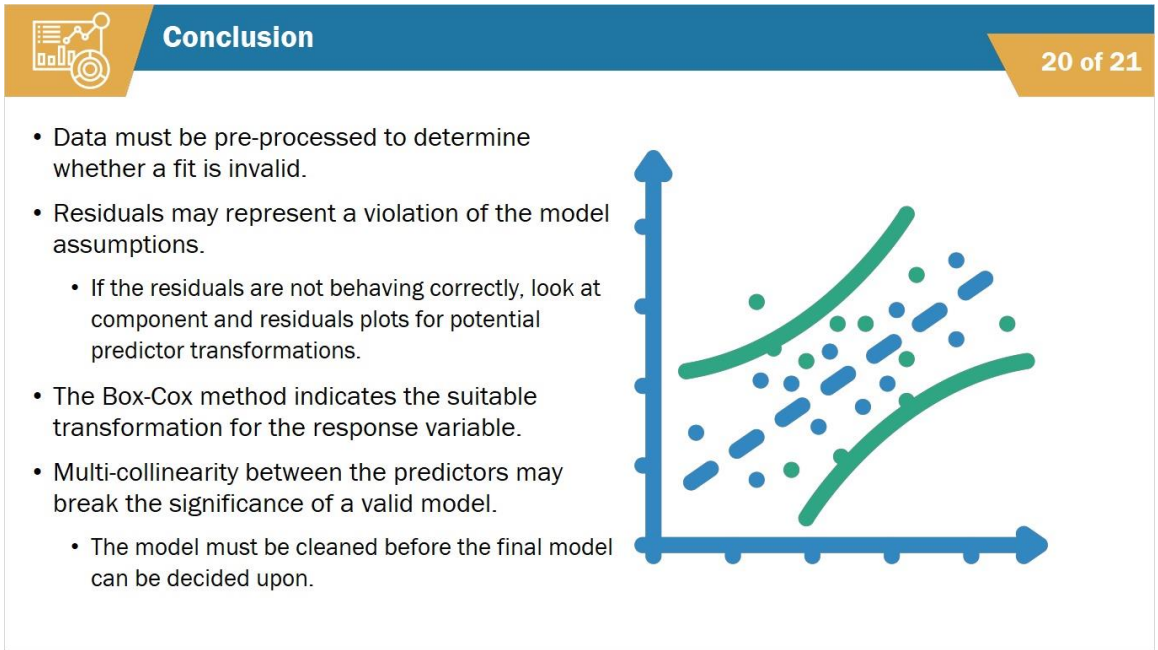
A final point to note in this example whilst not relevant to collinearity, but still worth mentioning here, is that the intercept still suffers from low significance. We can fit the

model without intercept using the following command. The new fit looks ideal in terms of the significances:

The model to be used is

$\hat{Y} \simeq 1.52X_1+1.49X_2+2X_3$.

Slide 20: **Conclusion**



Let us summarise the ideas of the presentation.

A fit may be invalid for many reasons, that's why we have to pre-process our data.

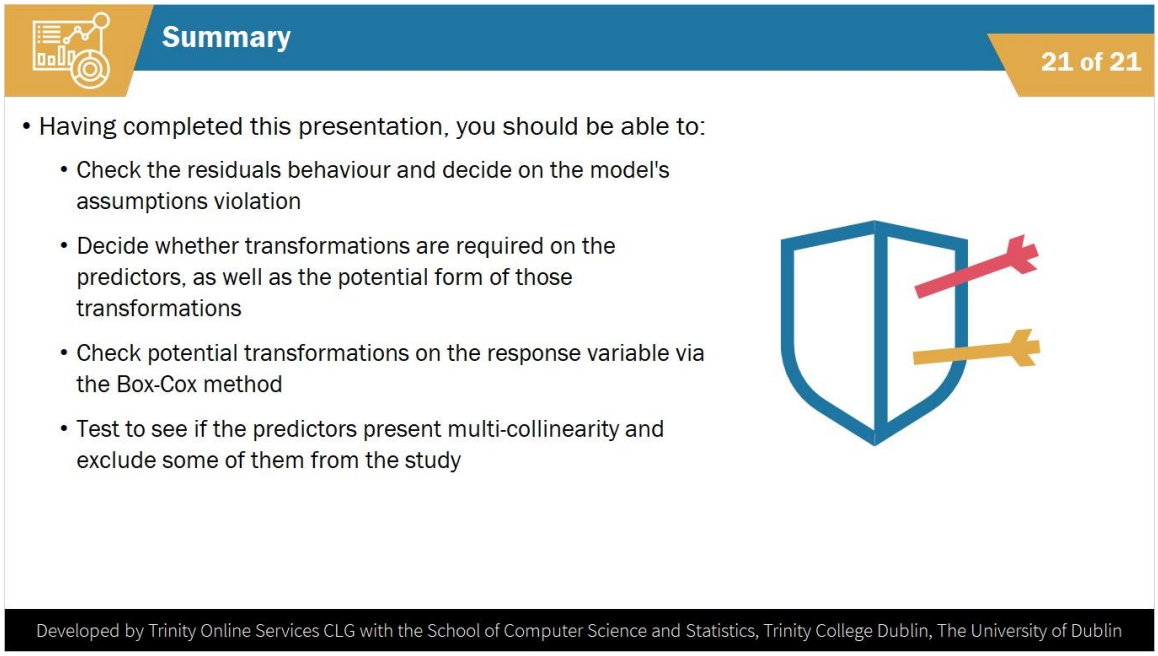
Residuals may represent a violation of the model assumptions. We have to check them before proceed.

In the case of residuals behaving wrongly, we can look at the C+R plots for potential transformations on the predictors.

Box-Cox method directly indicates the suitable transformation on the response variable.

The multi-collinearity between the predictors, may break the significance of a valid model. We have to clean the model with multi-collinearity before deriving and using the final model.

Slide 21: Summary



Having completed this presentation, you should be able to:

- Check the residuals behaviour and decide about the model's assumptions violation
- Decide whether transformations are required on the predictors, as well as the potential form of those transformations
- Check potential transformations on the response variable via the Box-Cox method
- Test to see if the predictors present multi-collinearity and exclude some of them from the study