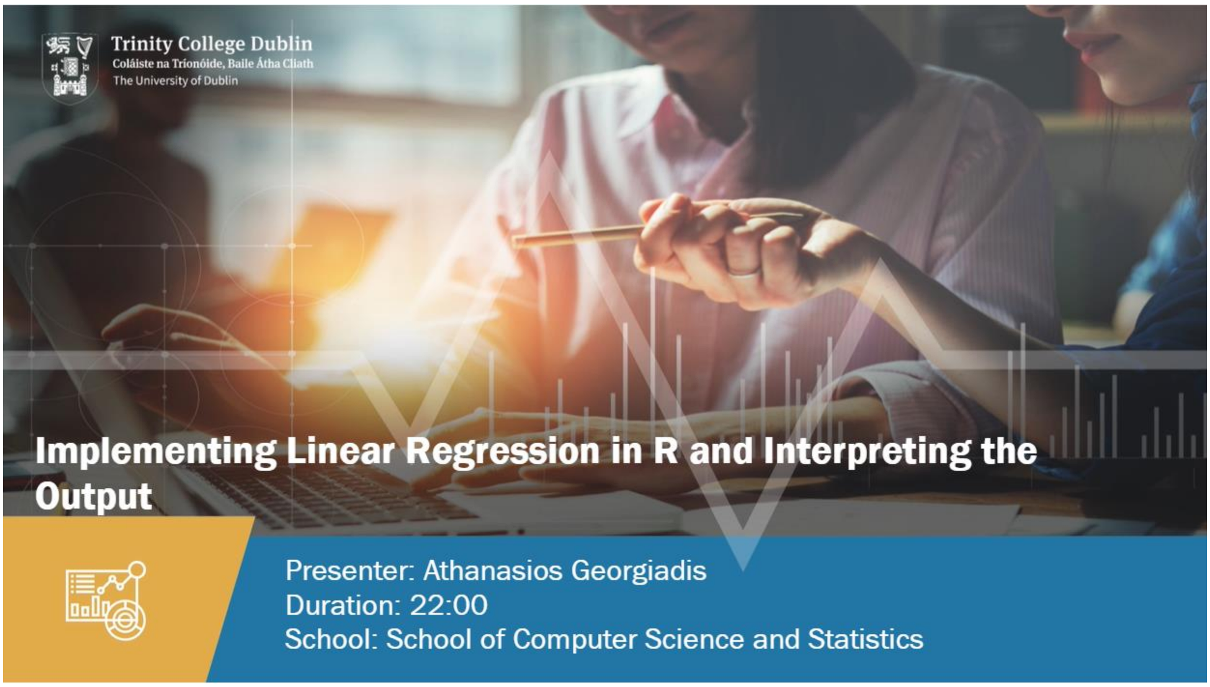
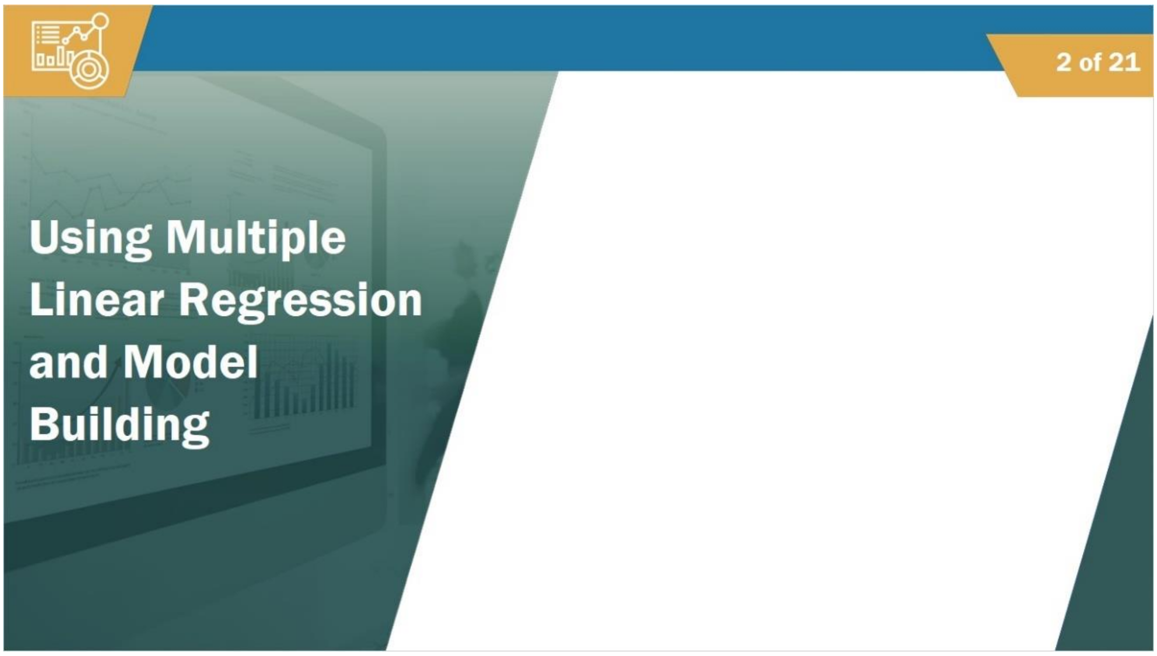
Session 3: Implementing linear regression in R and interpreting the output

Slide 1: ## Introduction

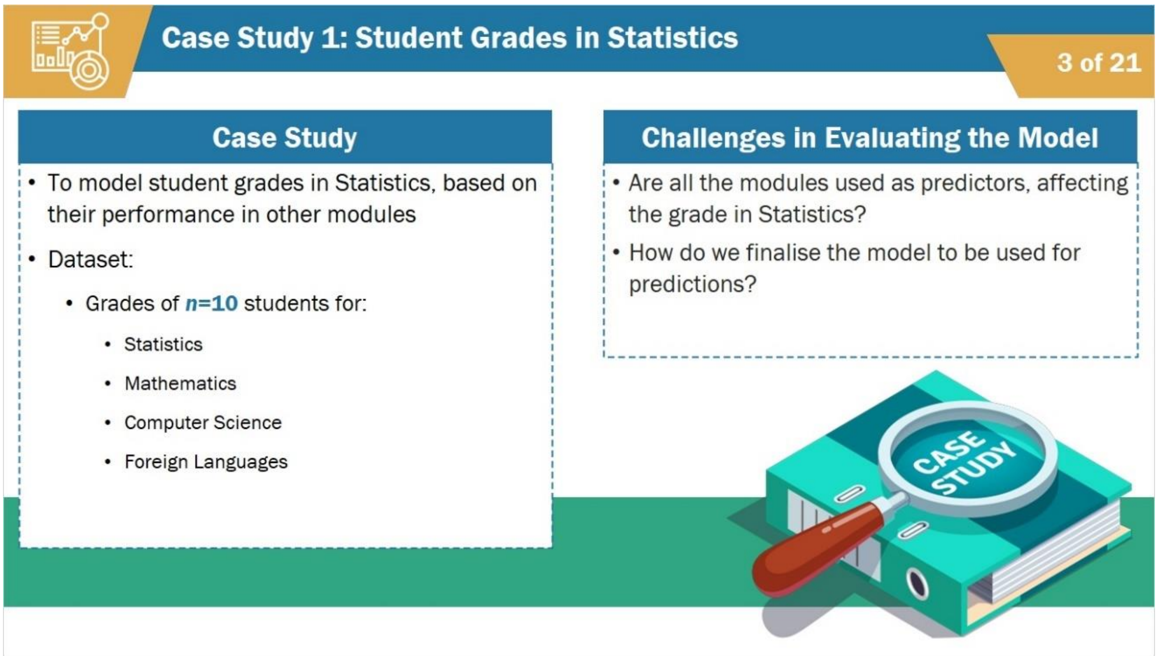


Hello and welcome to this presentation on Implementing Linear Regression in R and Interpreting the Output. My name is Athanasios Georgiadis and I'm the instructor for this presentation where we will look at two case studies and implement linear regression in R. In the first case study, we will implement multiple linear regression and model building. In the second, we'll implement polynomial regression.

Slide 2: ## Section 1: Case Study 1: Using Multiple Linear Regression and Model Building



Slide 3: **Case Study 1.: Student Grades in Statistics**



In our first case study, we would like to model Students' grades in Statistics, based on their performance in other modules. We have a dataset with the grades of n=10 Students in Statistics and some other modules; namely Mathematics, Computer Science and Foreign Languages.

We will use R to fit a multiple linear regression model. To this end, we would like to evaluate the obtained model.

Are all the modules used as predictors effecting the grade in Statistics?

How do we finalise the model to be used for predictions?

All these are just a few challenges that a model developer or user has to face when evaluating or creating a model..

In this presentation, we will address these challenges using a dataset and R language.

Slide 4: The Dataset

The Dataset

4 of 21

Grades for n = 10 students

Statistics (S)	Mathematics (M)	Computer Science (C)	Foreign Languages (L)
70	60	81	70
72	61	82	60
74	65	80	50
75	65	85	80
77	66	85	90
80	68	85	50
82	70	86	60
85	70	87	90
88	75	87	70
90	75	90	80

Take time to view the information on this slide.

We collect the grades of n=10 Students in the modules of Statistics (S), Mathematics (M), Computer Science (C) and Foreign Languages (L). The dataset is presented in this table.

Slide 5: Identifying Relationships Between Variables and Predictors

Identifying Relationships Between Variables and Predictors

5 of 21

- **Variable Y:** (Statistics grade) should be explained by the predictors:
  - X_1: Mathematics grade
  - X_2: Computer Science grade
  - X_3: Foreign Languages grade
- Potential relationships between the response variable and the predictors may be instinctively identified.
  - These relationships must be justified statistically.
- The data analyst may not be able to instinctively identify potential relationships in cases where their knowledge is limited.

Let's collect some initial thoughts.

The variable Y:= the grade in Statistics, should be hopefully explained by the predictors X_1:= grade in Mathematics, X_2:= grade in Computers and X_3:= grade in Languages.

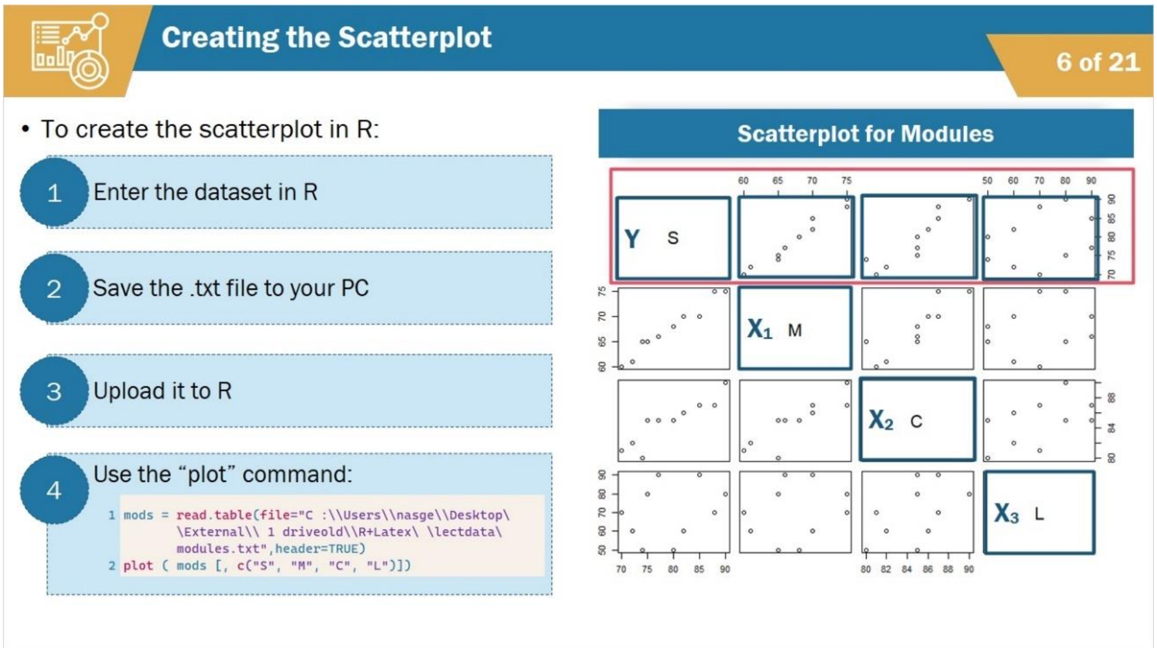
In the specific case study, we may have some experience or instinct about potential relationships between the response variable and the predictors. For example, we may expect the dependence of Y on X_1: the grade in Mathematics. On the other hand, the connection between Y and X_3 (the grade in Foreign languages), may be poor or non-existent.

These relationships must be justified statistically.

There may also be occasions when a data analyst has to deal with cases where his or her understanding of the data is limited and therefore, they do not instinctively recognise potential relationships.

Over the following slides, we are going to create a scatterplot and determine multiple linear regressions for different groups of variables to help determine which is the best to use.

Slide 6: Creating the Scatterplot



We start with the usual first step. We enter the dataset in R and view the scatterplot matrix. We save the .txt file in our PC, upload it in R and use the “plot” command shown and obtain the scatterplot.

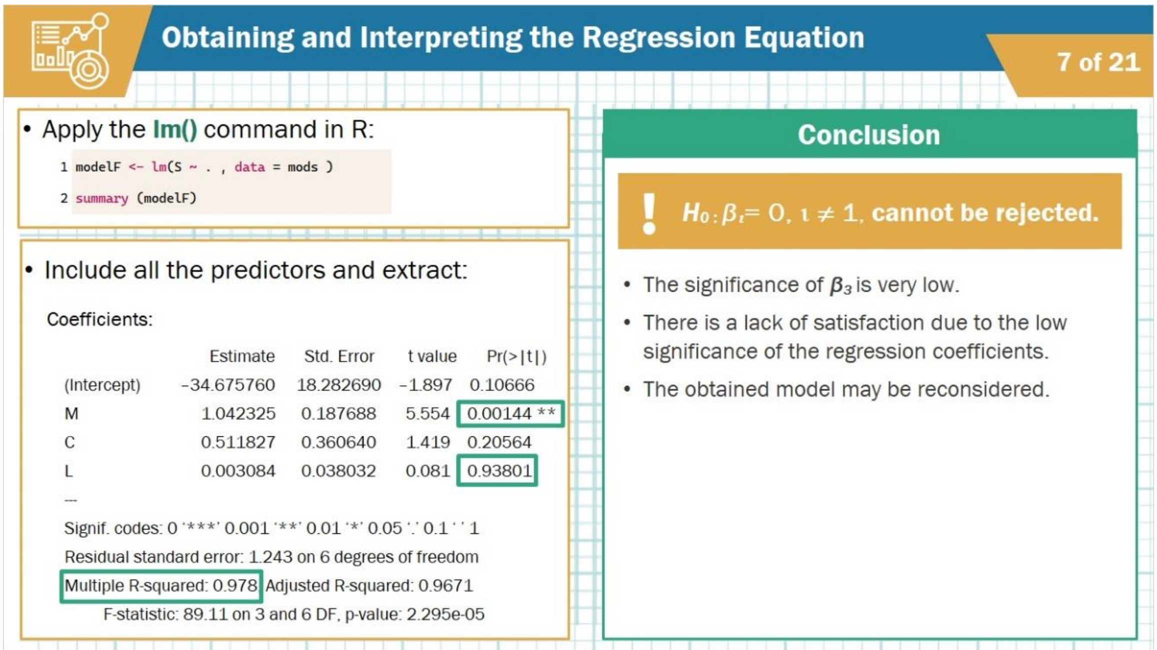
Let us read the scatter-matrix.

The scatter-matrix contains the response variable Y and the three predictors X_1, X_2, X_3.

Let’s focus on the first row. We can observe some linear dependence between the grade in Statistics Y and the grade in Mathematics X_1. Something similar holds for the grade X_2 in Computers too.

We can see a very sparse scatterplot between (X_3,Y). This supports our instinct, that Students' skills in Languages, may not affect their performance in Statistics.

Slide 7: Obtaining and Interpreting the Regression Equation



We apply the "lm()" command in R as follows:

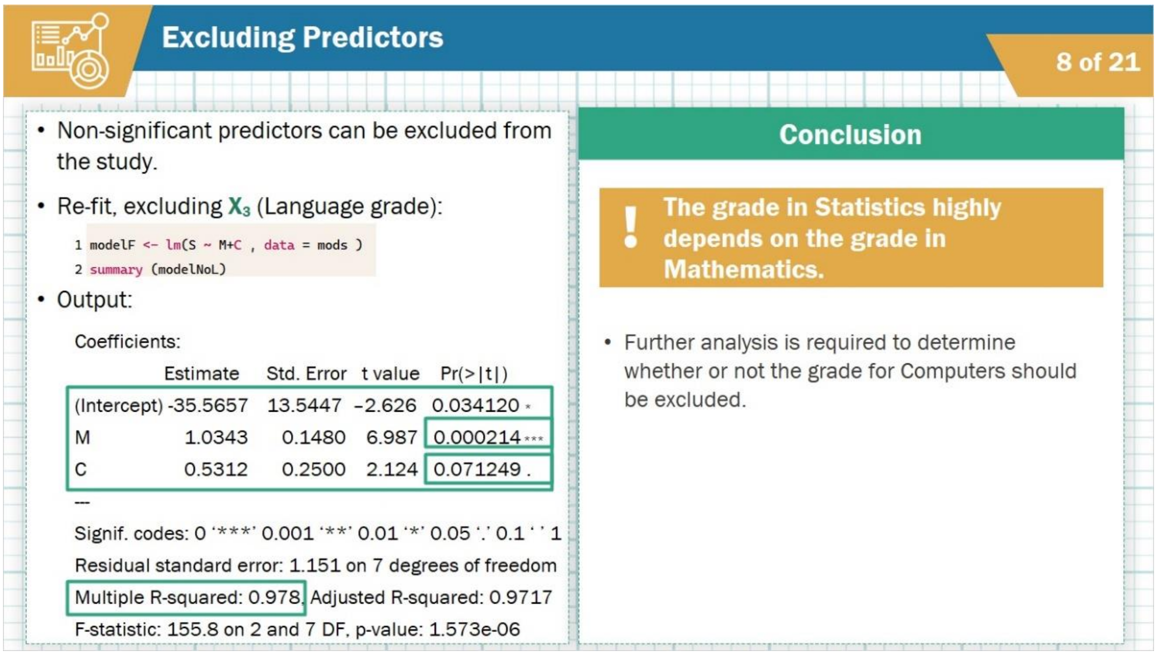
We start by including all the predictors that we have at our disposal. We then extract the following summary table.

Let's interpret the output carefully.

- The fit presents a very high R^2 around 97.8%.
- The coefficient β_1 of the X_1, is very significant. The p-value is around 1%, which is very low.
- There doesn't seem to be anything significant in the rest of the coefficients.
- Rigorously this means that the hypothesis H_0: $\beta_i = 0$, cannot be rejected, when "i" is not 1.
- The significance of β_3, which corresponds to the grade in Language, is really low.

Overall we are not completely satisfied, because of the low significance of the regression coefficients. We may reconsider the obtained model.

Slide 8: **Excluding Predictors**

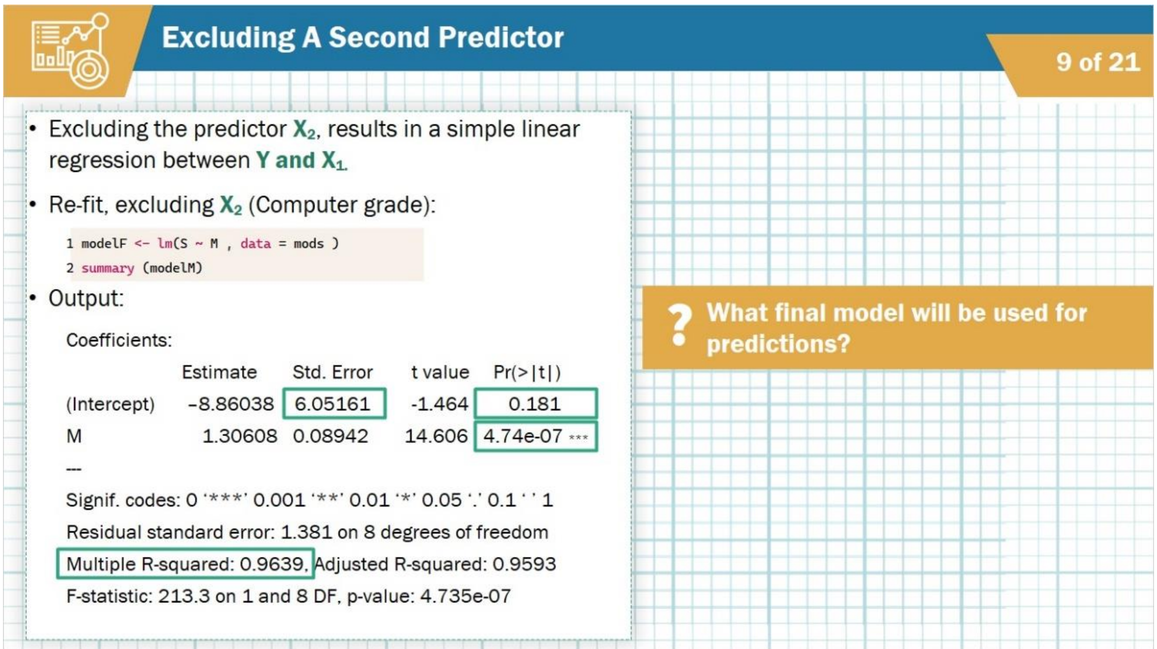


A predictor which ends up being non-significant could be excluded from the study. In our example, we re-fit excluding X_3 (the grade in Languages) and we obtain the following summary table in R.

Interpreting the output, we have the following:

- The fit still presents a very high $R^2 \simeq 97.8\%$.
- All the coefficients appear to be significant.
- The coefficient β_1 of the X_1, is the most significant. This confirms statistically our initial speculations that the grade in Statistics, highly depends on the grade in Mathematics.
- The p-value for the test H_0: $\beta_2 = 0$, which corresponds to the grade in Computers, is around 7%, which is a low one, but still leaves the door open for further analysis. Is it a good idea to exclude the predictor X_2?

Slide 9: Excluding a Second Predictor



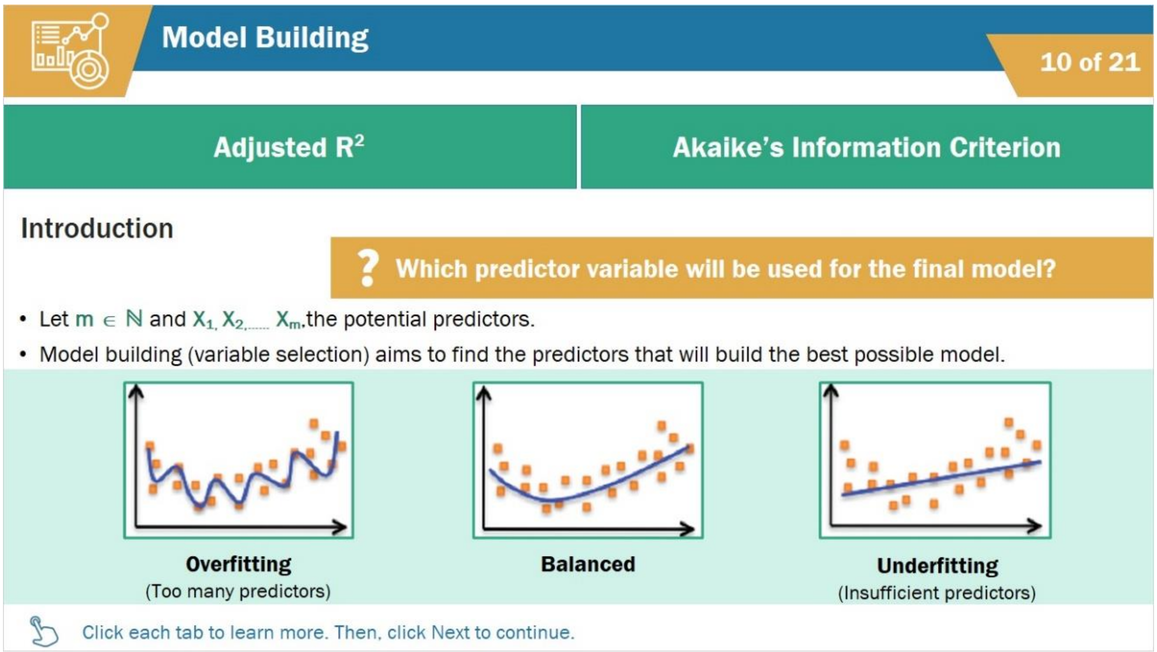
We re-fit excluding X_2. Therefore, we have a simple linear regression between Y and X_1.

Interpreting the output of the SLR model obtained, we observe that:

- The fit still presents a very high $R^2 \simeq 96.4\%$.
- The coefficient β_1 of the X_1, is very significant. The intercept presents a lower significance. More carefully, there is a large standard error for it.

The last two obtained models look satisfactory. But what is the final model that we will use for predictions?

Slide 10: Model Building

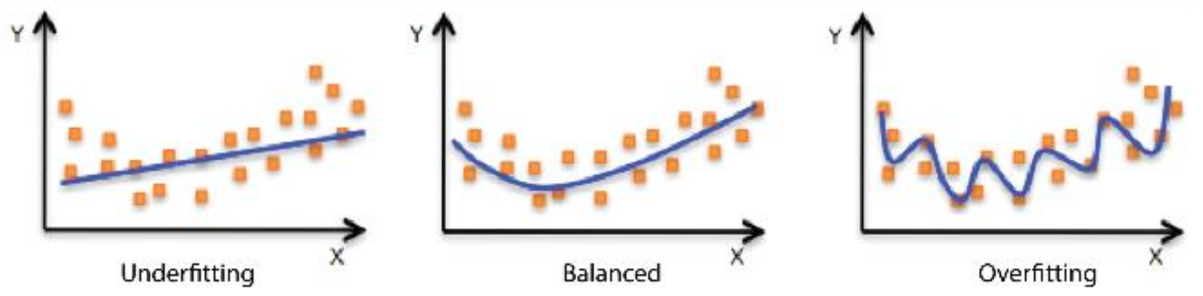


Nowadays, datasets can be easily found. In most cases, there are many candidates for predictor variables available. The natural question is which of them should be used for our final model?

Formally, let m ∈ ℕ and $X_1, X_2, \ldots, X_m$, the potential predictors.

Model building or Variable Selection, aims to find the predictors that will build the best possible model.

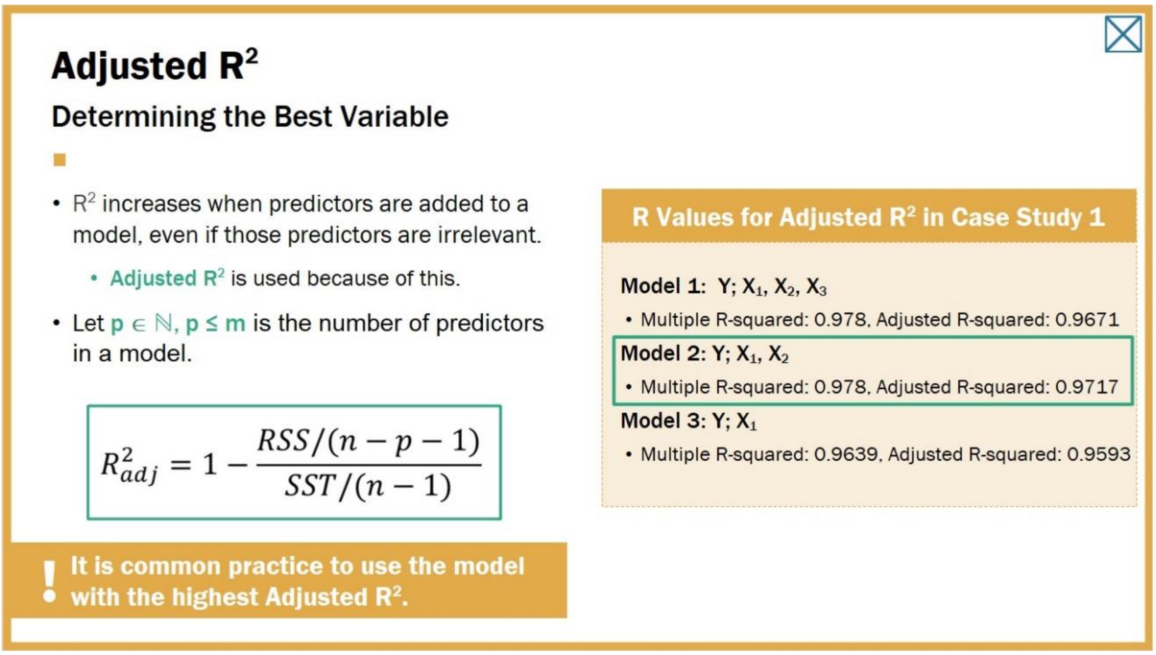
Including too many predictors in the model, is referred to as over-fitting while the opposite case is called under-fitting.



There are several measures that help us arrive at the "best" model. We will look at two measures that are most commonly used: Adjusted R^2 and Akaike's Information Criterion.

Click the tabs to learn about each of these measures. When you are ready, click "Next" to continue.

Tab 1: **Adjusted R^2**



We start with the so-called Adjusted R^2.

When we add predictors to a model, R^2 increases - even if the predictors are completely irrelevant.

For this reason, we use the so-called Adjusted R^2. Let $p \in \mathbb{N}$, $p \leq m$, be the number of the predictors included in a model. R provides the value of Adjusted R^2 in the summary table.

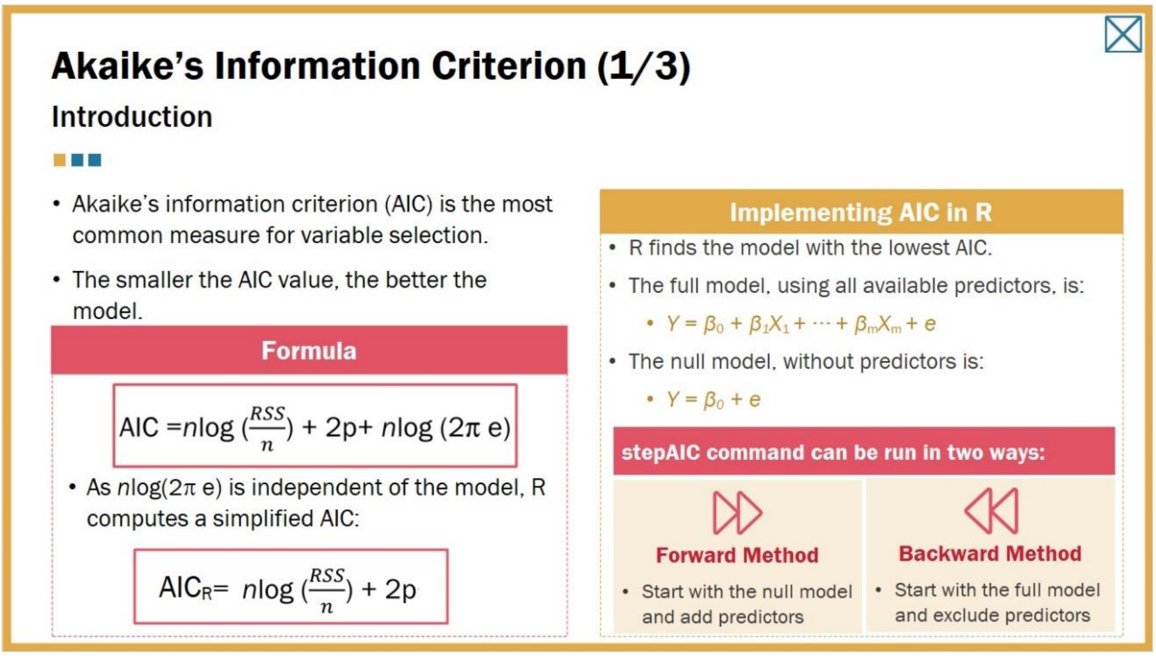
The precise value of R^2_{adj} is as follows:

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

It is common practice to use the model with the highest Adjusted R^2 .

If we go back to our case study on the student grades, the value for the Adjusted R^2 was featured in the output beside the R^2 value for all three of the models we ran. Those values are displayed on your screen now. Between the three models we examined, the one with the highest Adjusted R^2 is the model including predictors X_1, X_2. Although, we can observe small differences in Adjusted R^2 between the three models.

Tab 2: Akaike's Information Criterion



The most common measure for variable selection is <u>Akaike's</u> *Information Criterion* (AIC).

Its definition passes via the likelihood and it is based on balancing how well the model fits with a penalty for model complexity. *The smaller the value of AIC, the better the model.*

Precisely, for a model with *p* predictors, in a sample of *n*, the AIC formula is

AIC =$n$log $(\frac{RSS}{n})$ + 2p+ $n$log (2π e)

The last term, $n$log (2π e), is independent of the model under consideration, therefore R computes the simplified version of AIC, by excluding this term as below.

Let us now implement AIC in R.

R finds the model with the lowest possible AIC and this is the one we finally work with.

Let's look at some of the terminology.

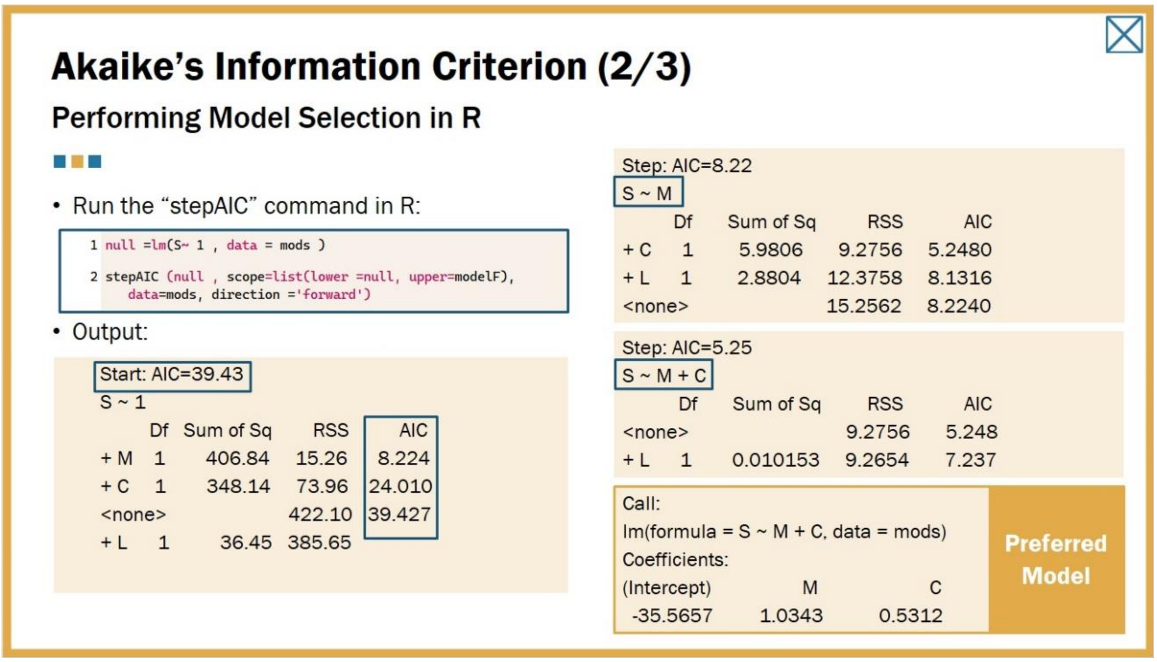
We refer to the model:

$Y=\beta_0+\beta_1X_1+\ldots+\beta_mX_m+e$, as the full model (using all the available predictors) and to the model $Y=\beta_0+e$ as the null model (using no predictors).

There are two methods of finding AIC in R:

- The first is known as the forward method. stepAIC command in R starts with the null model and adds predictors until it finds the model with the lowest AIC.
- The second is the backward method, which is performed in the opposite way; starting from the full model and excluding predictors.

Tab 2.1: Performing Model Selection



Let us perform the Model selection in our case study.

We run the command stepAIC in R, deriving a large table that I have split in two to allow for better commenting on it.

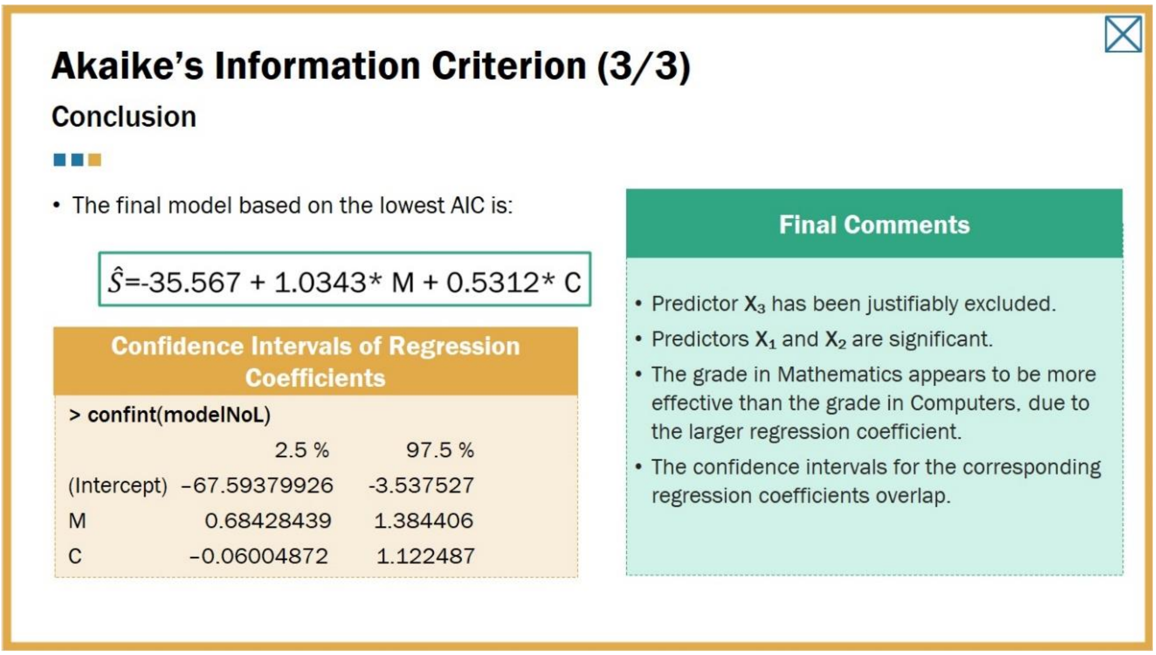
R starts with the null model because we chose to work with a forward method. This model presents AIC = 39.43. Next R gives a new AIC for every single predictor, if we add new predictors to the model. You can see these highlighted on the screen.

R then adds the predictor which leads to the new model with the smallest AIC and continues the algorithm in the same way.

R continues this way until the addition of more predictors cannot offer a lower AIC.

R concludes to the preferred model and presents its coefficients at the end of the table.

Tab 2.2: Conclusion



The *final model* based on the lowest-AIC criterion is:

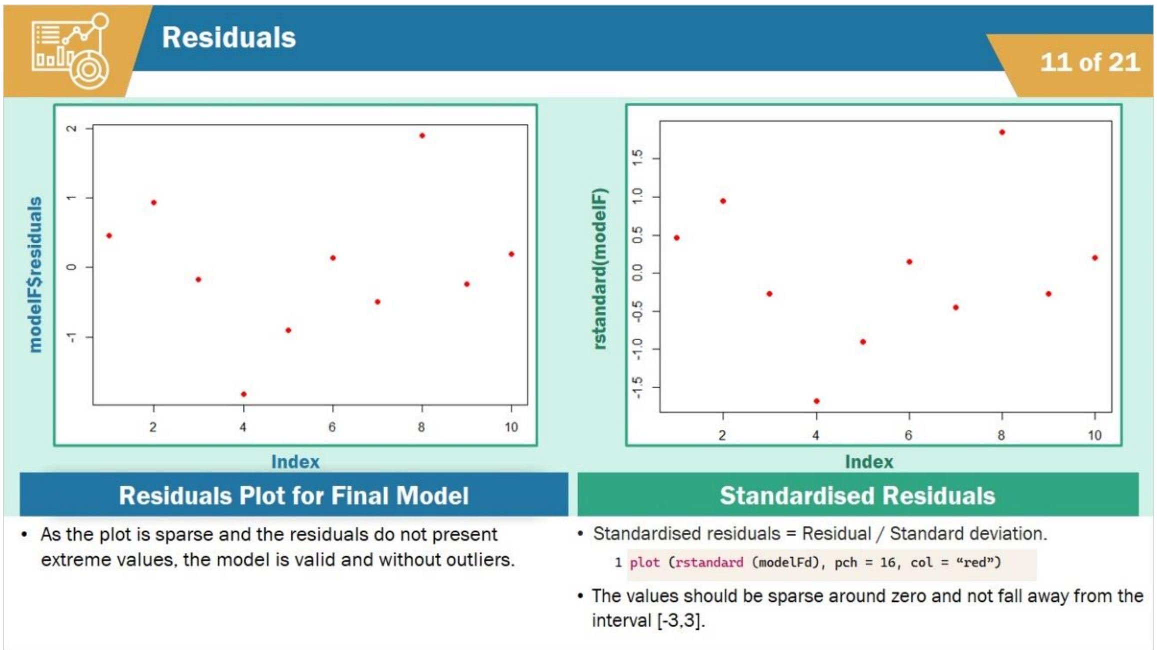
$\hat{S}$=-35.567 + 1.0343* M + 0.5312* C, where S, M and C are the grades in Statistics, Mathematics and Computers respectively.

Having the final model in hand, we can find for example, the confidence intervals of the regression coefficients presented here.

Let us further comment on the final regression model:

- The predictor X_3 has been justified statistically to be excluded.
- The predictors X_1 and X_2 are significant.
- The grade in Mathematics appears to be more effective than the grade in Computers because of the larger regression coefficient.
- The confidence intervals for the corresponding regression coefficients, overlap.

Slide 11: **Residuals**

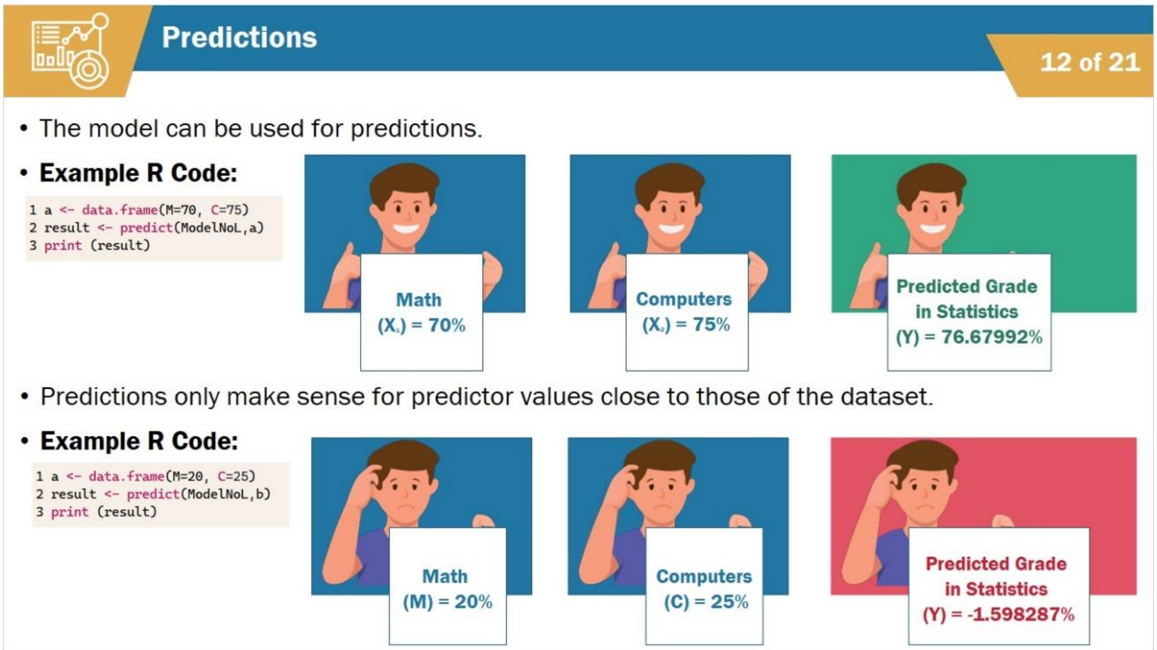


Let us extract the Residuals' plot for the final model, to validate it. As we can see, the plot is sparse, as it should be and the residuals do not present extreme values. This means that the model is valid and without outliers.

When we divide the residual by its standard deviation, we obtain the standardised residuals. Their values should be sparse around zero and not fall away from the interval [-3,3].

Here we can see that the standardised residuals are behaving properly, which validates the final model.

Slide 12: **Predictions**



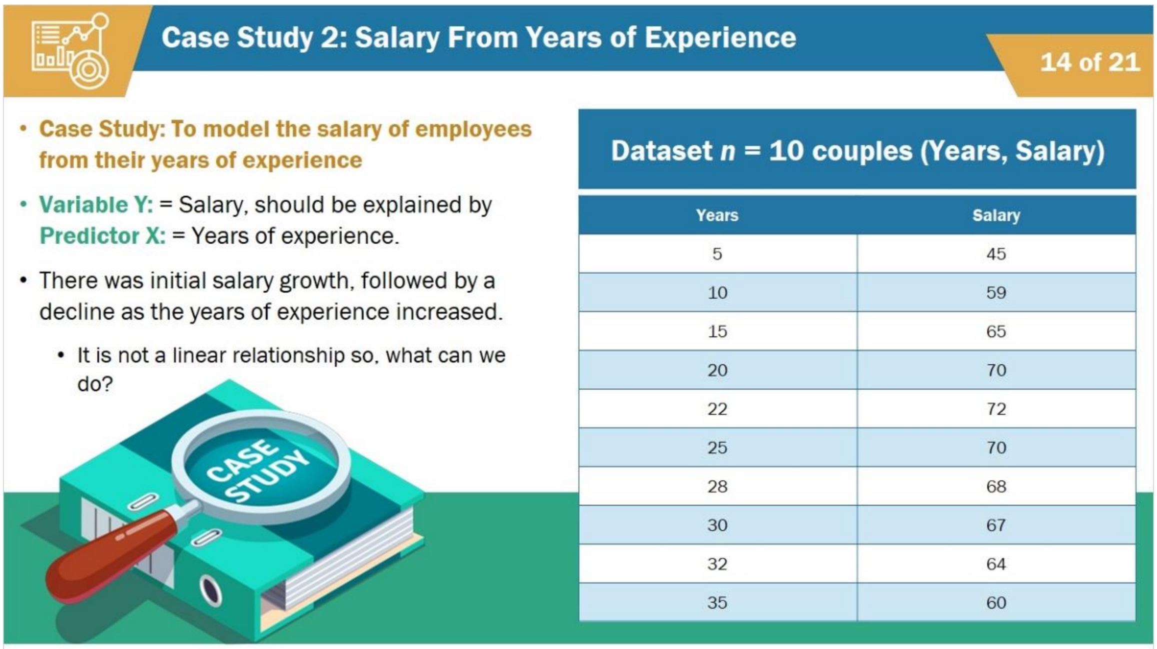
In practice, we can use the obtained model for predictions. For example, having a student with grades X_1=70 (in Math) and X_2=75 (in Computers), the model predicts a grade in Statistics as: around 76.7.

As a remark here, we have to mention that the predictions make sense for values of the predictors, “close” to those of the dataset. For example, if we want to predict the grade in Statistics, of a Student with grades M=20 and C=25 (far from the observed data values), we end up with an estimated grade in Statistics of around -1.6.

This negative grade in Statistics, makes no sense of course. The inconsistency is because we used predictors’ values, that are far from the original dataset. This has to be avoided.

Slide 13: Section 2: Case Study 2: Using Polynomial Regression

Slide 14: **Case Study 2: Salary From Years of Experience**



In our second case study, we would like to model the Salary of employees from their years of experience. We collected a dataset of n=10 couples in the form (Years,Salary). The dataset is the following:

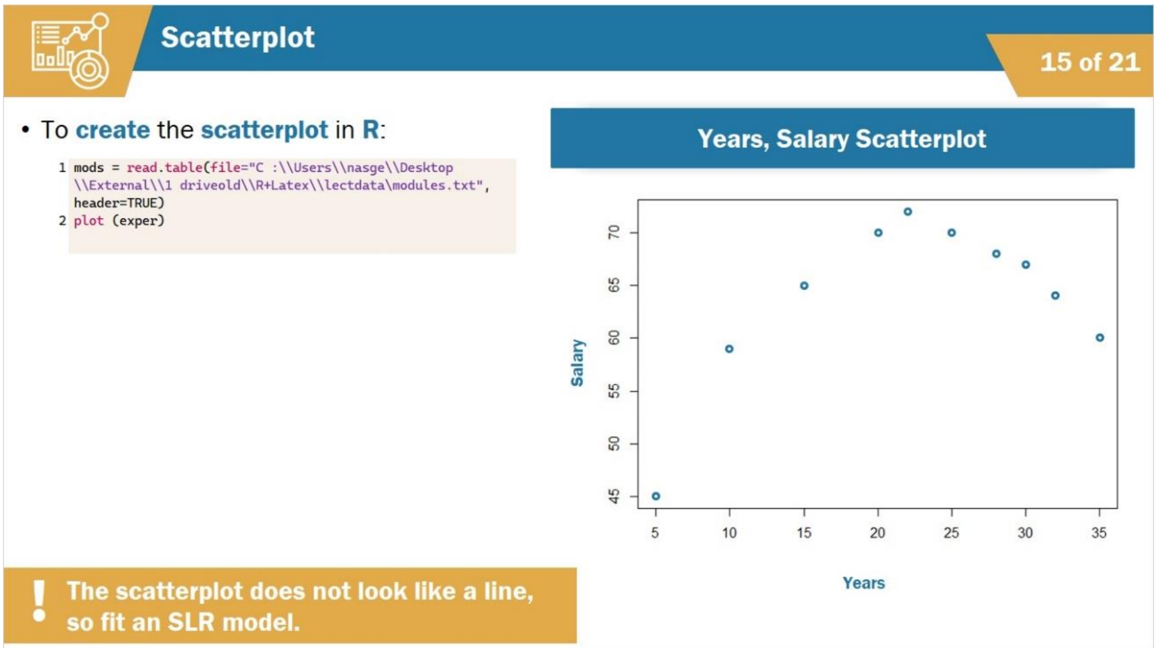
Let us collect some initial thoughts.

The variable Y:= the Salary, should be explained by the predictor X:= the years of experience. We have only one predictor variable at our disposal.

From the dataset, we may observe an initial growth of the Salary and later, a decline when more years of experience are added. This is, of course, not a case with a linear relationship.

So what can we do?

Slide 15: Scatterplot

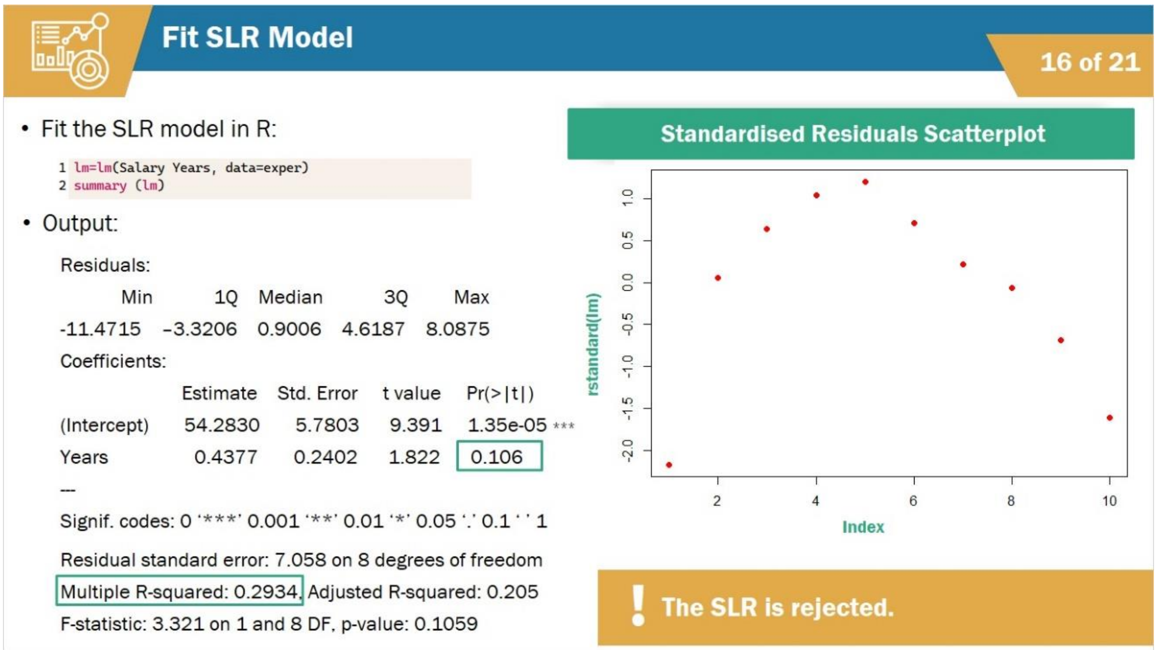


We act as always: we enter the dataset in R and view the scatterplot. We save the .txt file in our PC, upload it in R and use the “plot" command.

The scatterplot does not look like a line.

What if we fit a SLR model?

Slide 16: Fit SLR Model



We fit the SLR model in R, deriving this summary table. We ask R to plot the standardised residuals too, obtaining this graph.

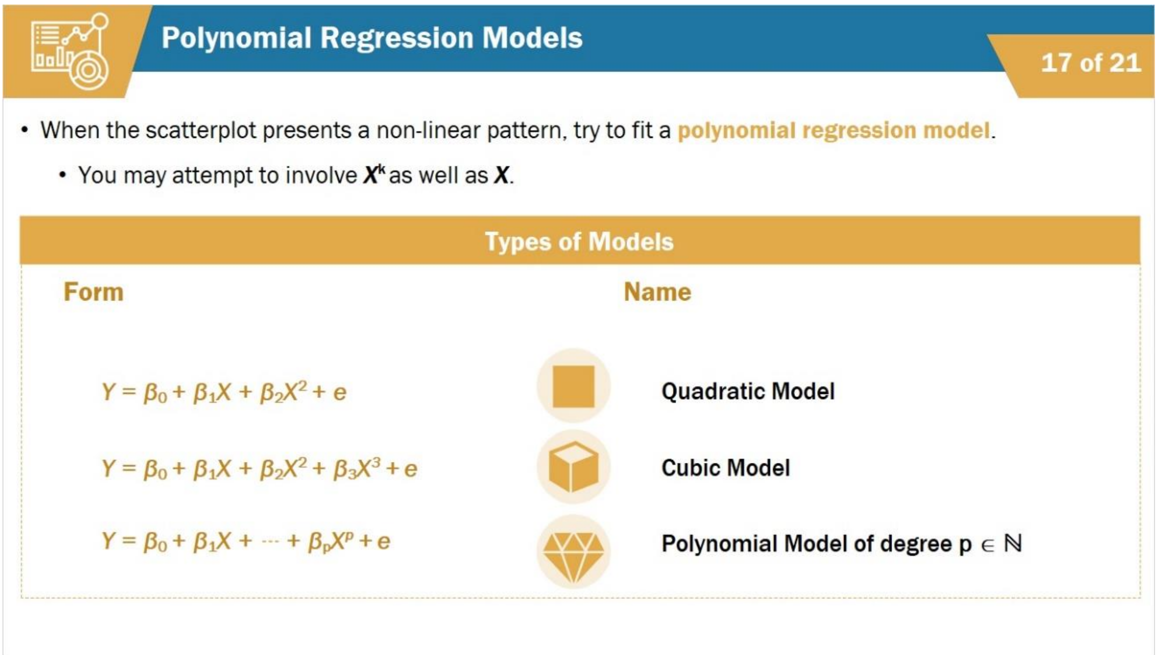
Let us comment on the SLR:

- The fit presents a disappointing $R^2 \simeq 29\%$.
- The predictor Years, the only one we have at our disposal, appears to be non-significant.
- The residuals present a very large range and the standard error is large.
- The standardised residuals are not sparse. They present a pattern. Not an appropriate residuals' curve.

For all these reasons, The SLR is rejected.

What can we do?

Slide 17: Polynomial Regression Models

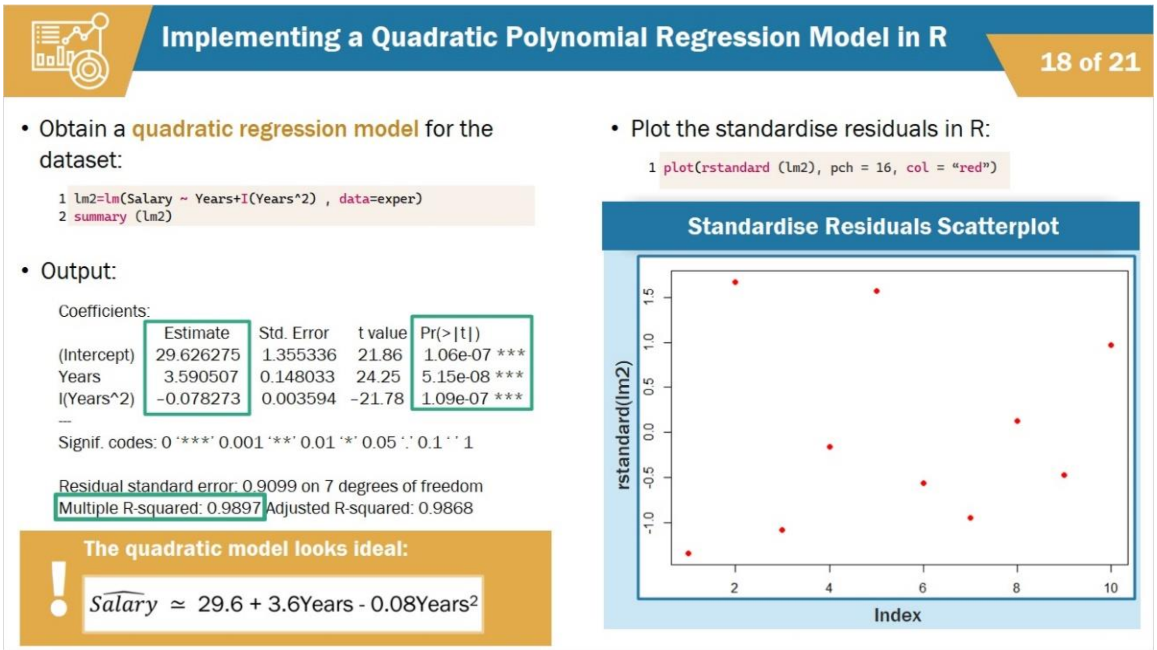


When the scatterplot presents a non-linear pattern, we may try to fit a polynomial regression model. Except with the predictor X, we may attempt to involve its powers X^k too.

A model in the form of $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$, is referred to as a quadratic model.

A model of the form $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e$, is referred to as a cubic model and more generally, a model $Y = \beta_0 + \beta_1 X + \ldots + \beta_p X^p + e$, is referred to as a polynomial model of degree (or order) p.

Slide 18: Implementing a Quadratic Polynomial Regression Model in R



We use R as follows to obtain a quadratic regression model for the Salary-Experience dataset. Note that in the lm() command, we add a new predictor Years^2. Then R runs a multiple regression with predictors X_1:=Years and X_2:=Years^2. We derive the following summary table.

We then ask R to provide us with the plot of the standardised residuals and get the next graph.

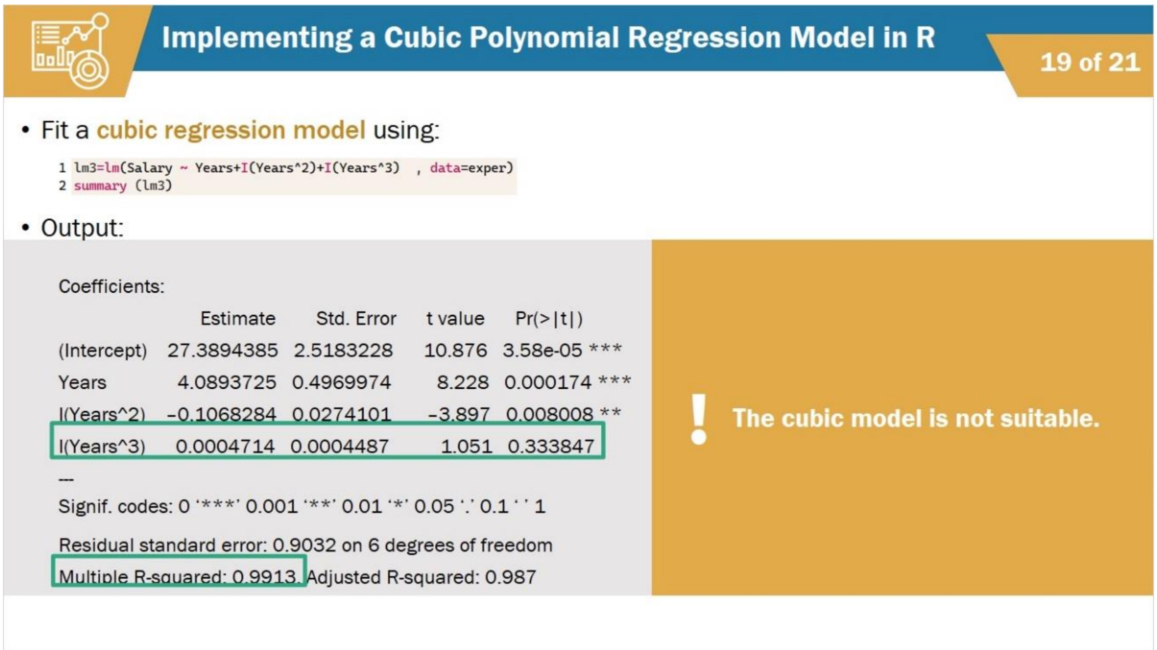
Let's comment on the polynomial fit.

- The summary table presents an ideal $R^2 \simeq 99\%$ and high significance in all the regression coefficients.
- The residuals are sparse around zero - an optimal residuals' plot.

The *quadratic model* looks ideal and has the form:

$\widehat{Salary} \simeq$ 29.6 + 3.6Years - 0.08Years^2

Slide 19: Implementing a Cubic Polynomial Regression Model in R

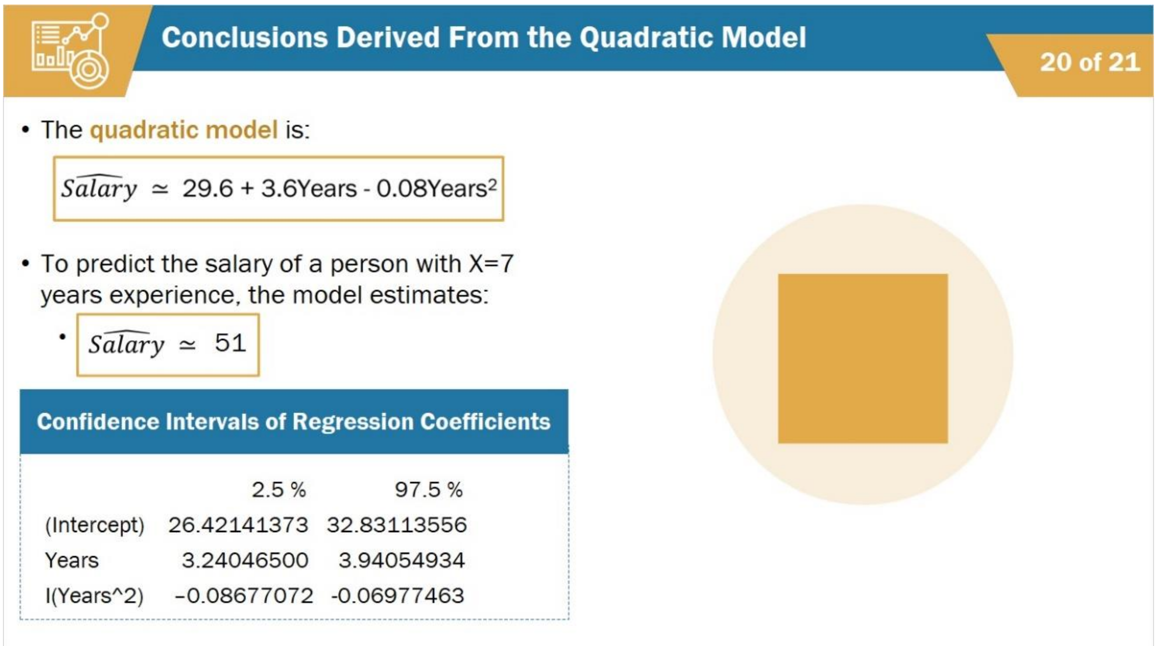


Let us attempt to fit a cubic polynomial regression model with this command. We obtain the following table.

We can see that the R^2 is very large, as before, but the coefficient β_3 is not significant. It may well be zero. We do not use the cubic model.

We will stay with the quadratic one obtained before. We can now use it for predictions.

Slide 20: Conclusions Derived from the Quadratic Model

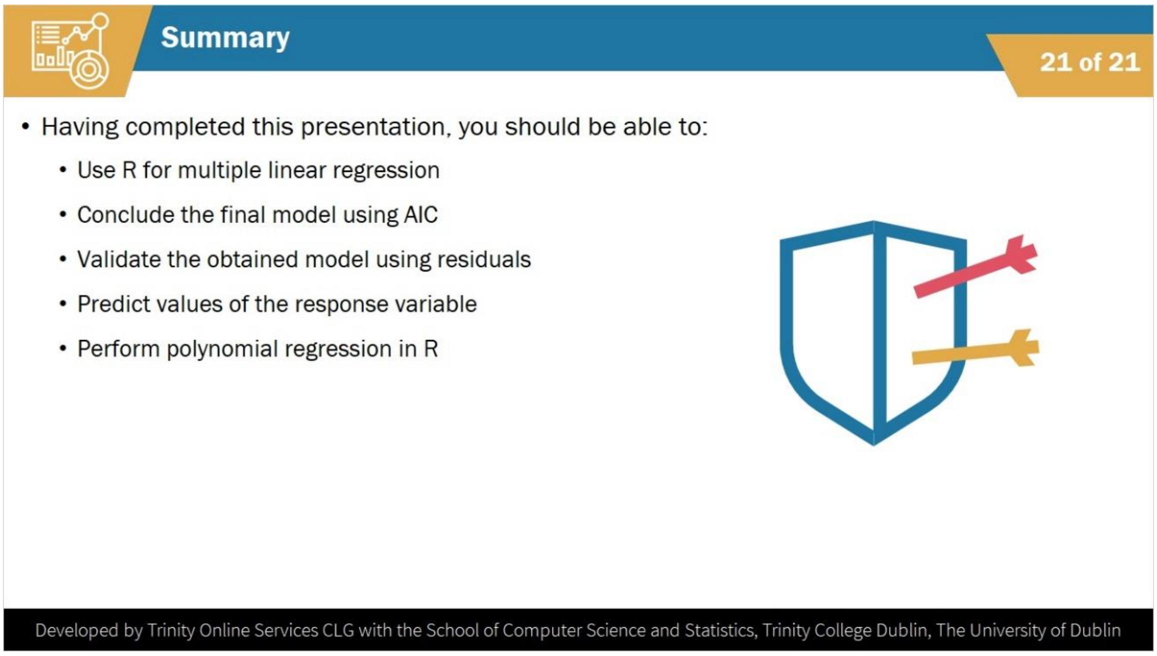


For a prediction of the salary of a person with X=7 years of experience, the model estimates:

$\widehat{Salary} \simeq 51$

The confidence intervals for the regression coefficients are listed here.

Slide 21: Summary



Having completed this presentation, you should be able to:

- Use R for multiple linear regression
- Conclude the final model using AIC
- Validate the obtained model using residuals
- Predict values of the response variable
- Perform polynomial regression in R