Tutorial 3 - Linear Models

I hope this tutorial will help you master linear models. Don't hesitate to ask me questions if you need further clarification. That's what I'm here for.

Learning objectives:

Familiarize yourself with R's functions for creating linear models, as well as with data exploration and checking model application conditions.

I. DATA EXPLORATION

1. Getting started

- Download the dataset « pokemon_sub.csv ».

2. Exploration steps

Sampling effort

Use the table function on R to check for EACH qualitative variable that you have enough data per modality. Here we have only one qualitative variable : type1.

Here's an example:

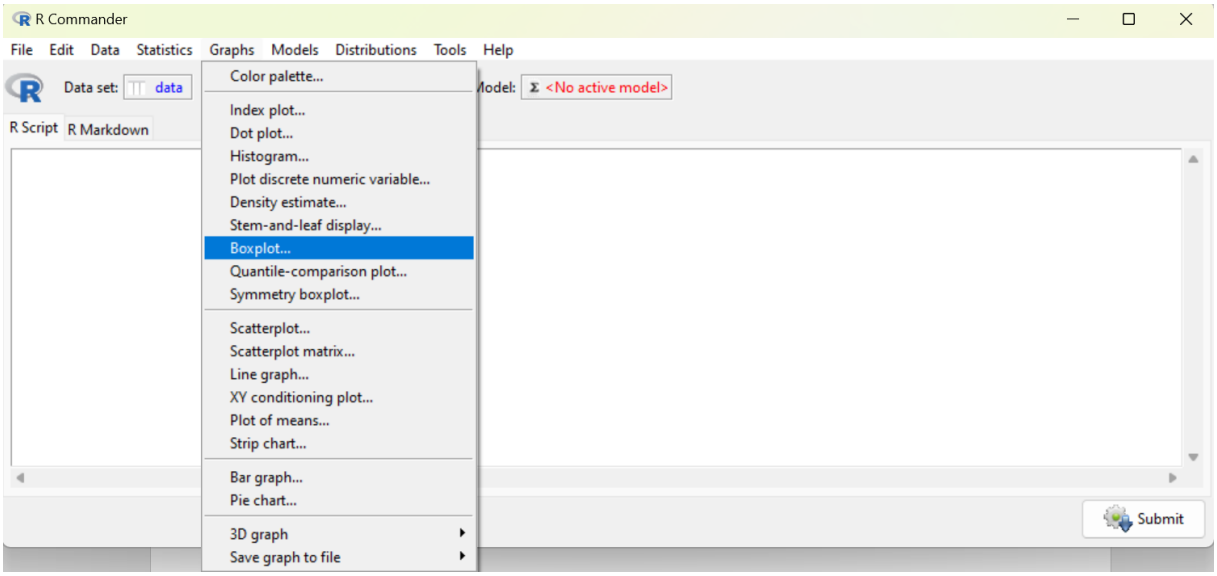
```
table(data$type1)
```

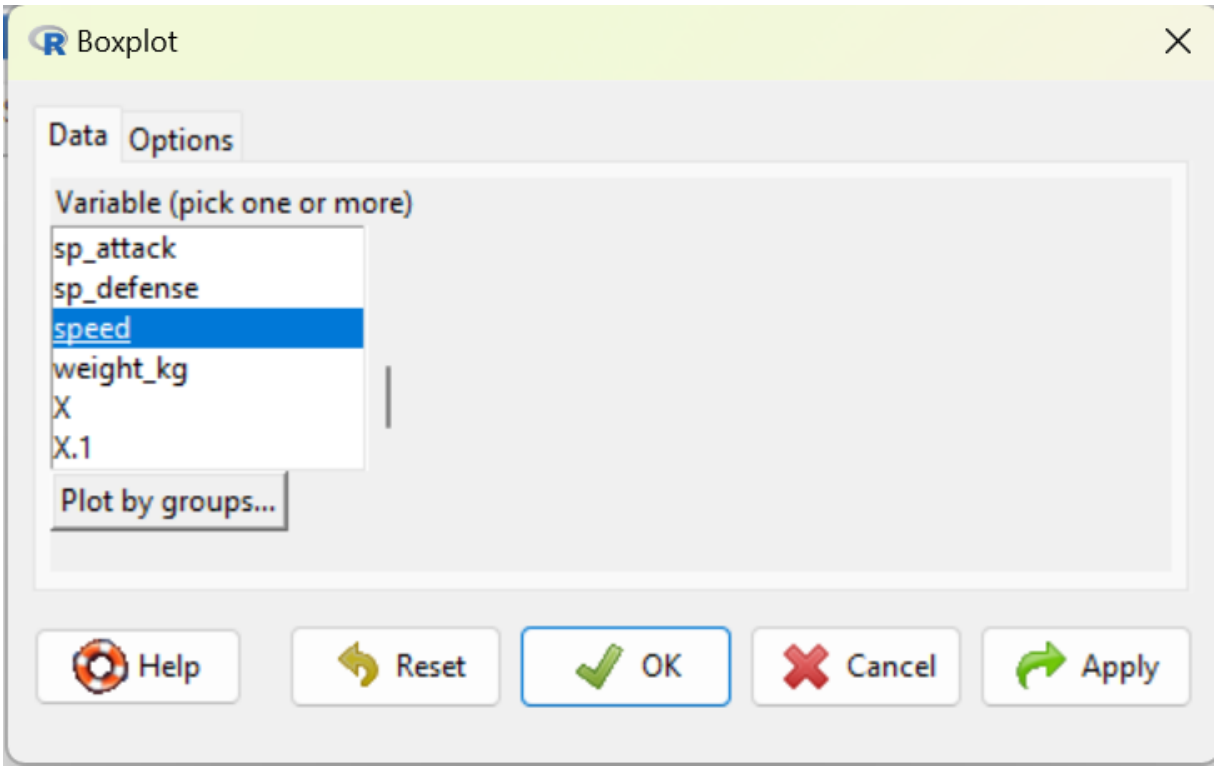
Outliers in quantitative variables

For the presence of extreme data, take each quantitative variable and make a boxplot.

For example : the variable speed.

BUTTON CLICK TEAM





And then OK.

CODING TEAM

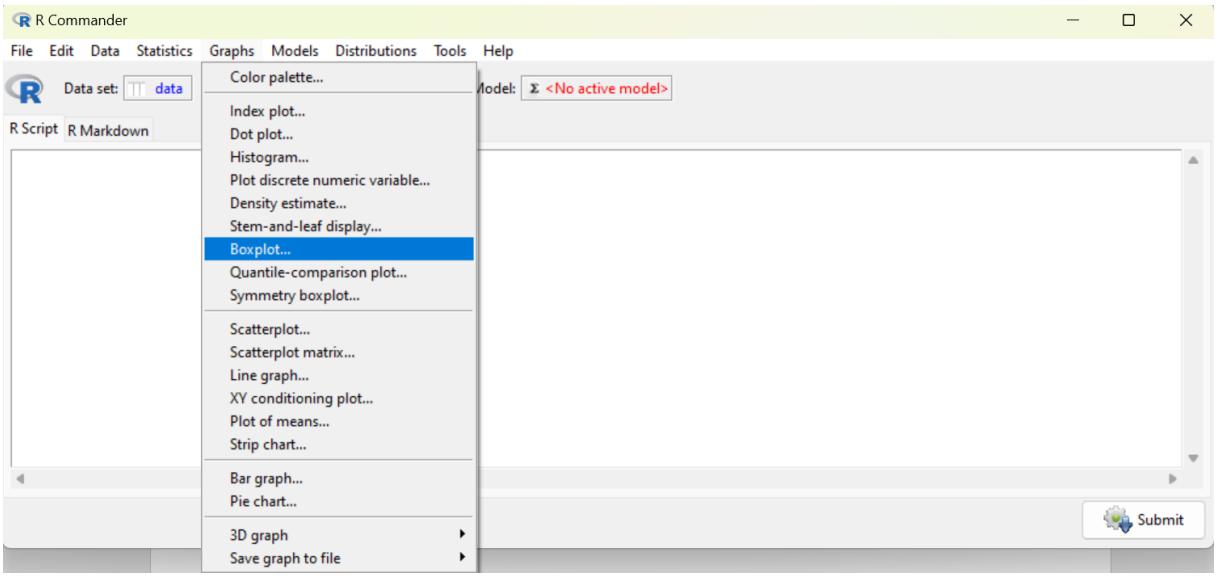
```
boxplot(data$speed)
```

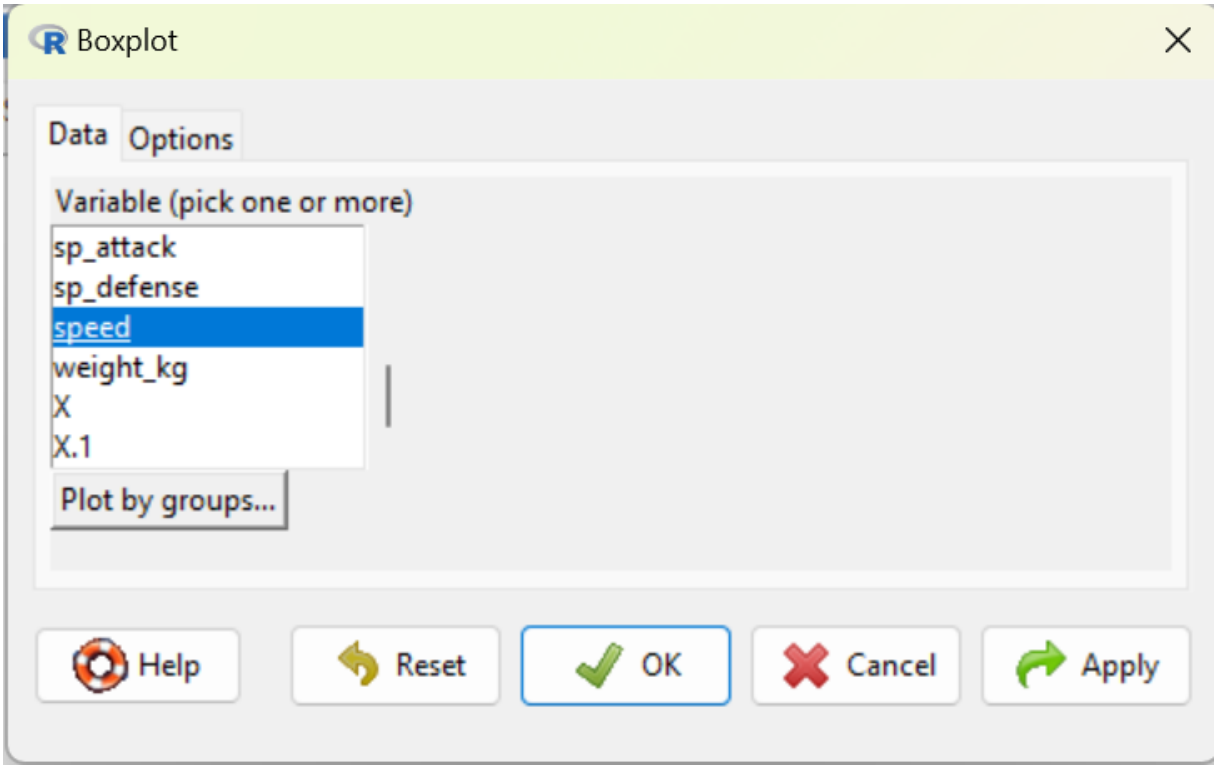
Study of the relationships between the response variable and the explanatory variables

i. When the explanatory variable is qualitative

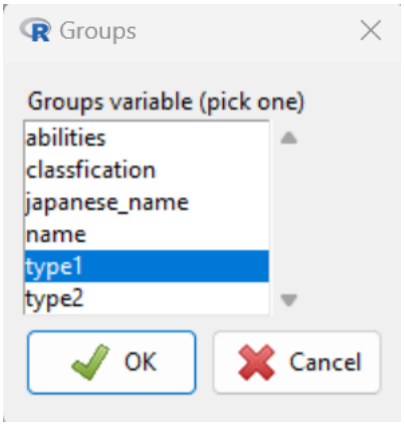
You can plot the response variable as a function of the qualitative explanatory variable.

BUTTON CLICK TEAM





The click on Plot by groups



Then OK.

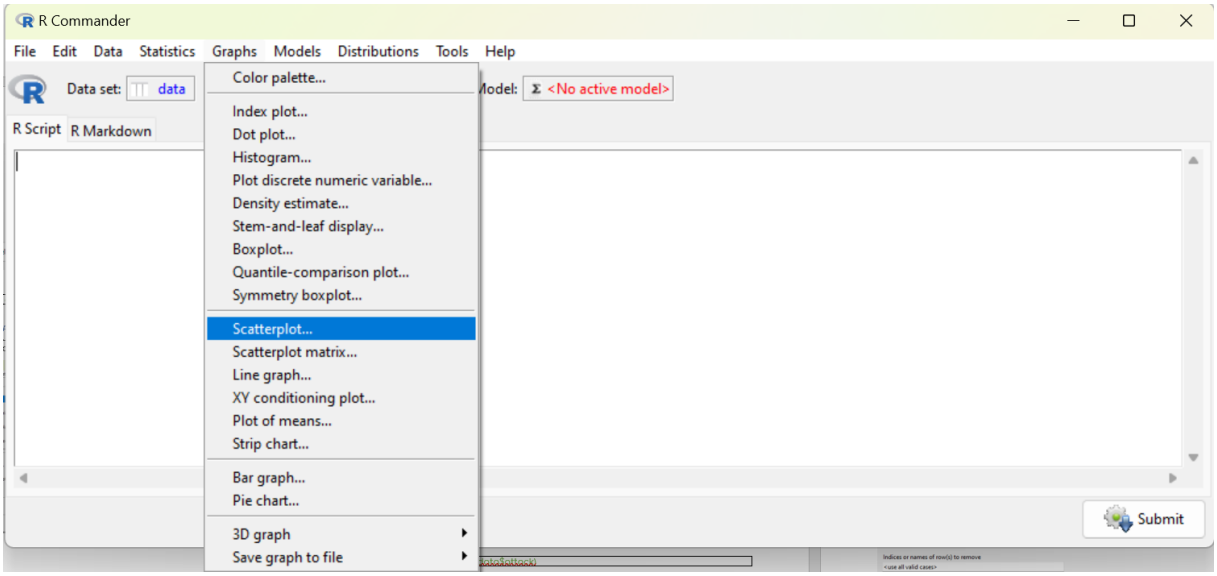
CODING TEAM

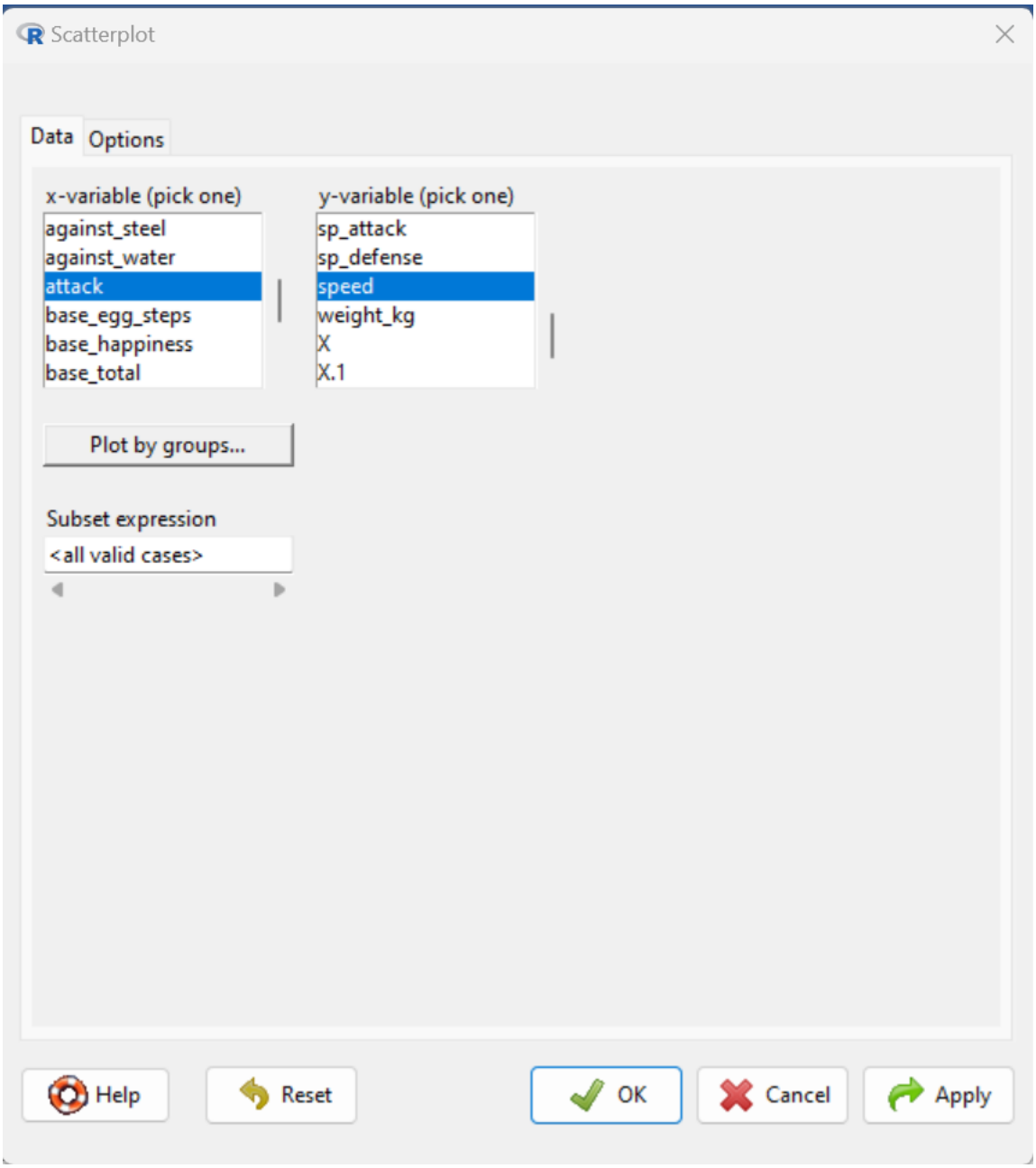
```
boxplot(data$speed ~ data$type1)
```

ii. When the explanatory variable is quantitative

You can plot the scatterplot (dotplot) of the response variable against the quantitative explanatory variable.

BUTTON CLICK TEAM





And then click on OK.

CODING TEAM

```
plot(data$speed ~ data$attack)
```

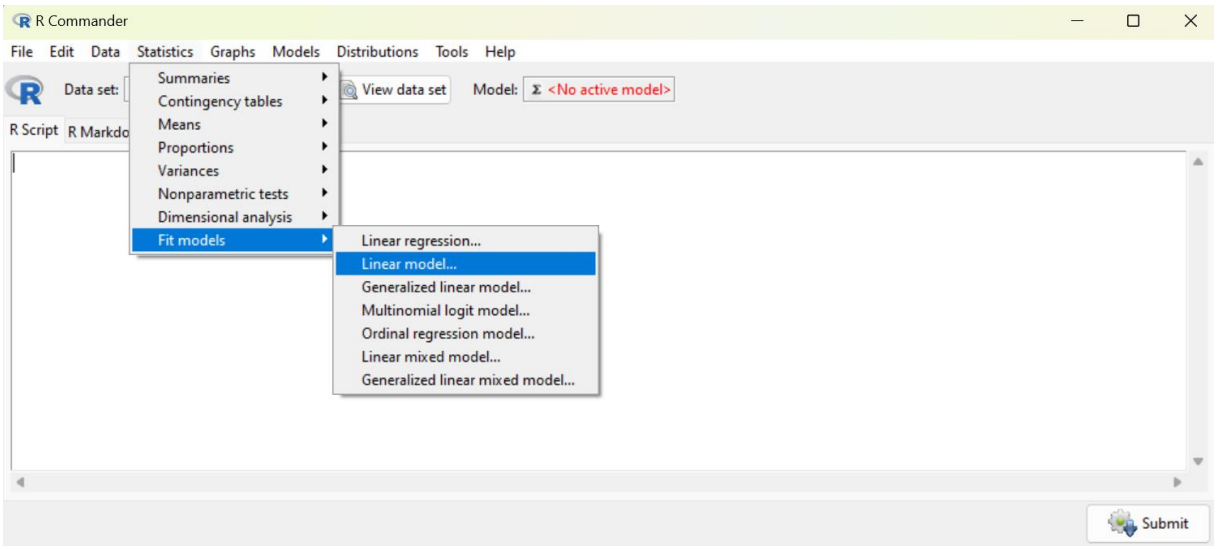
I let you do the same for the variable height.

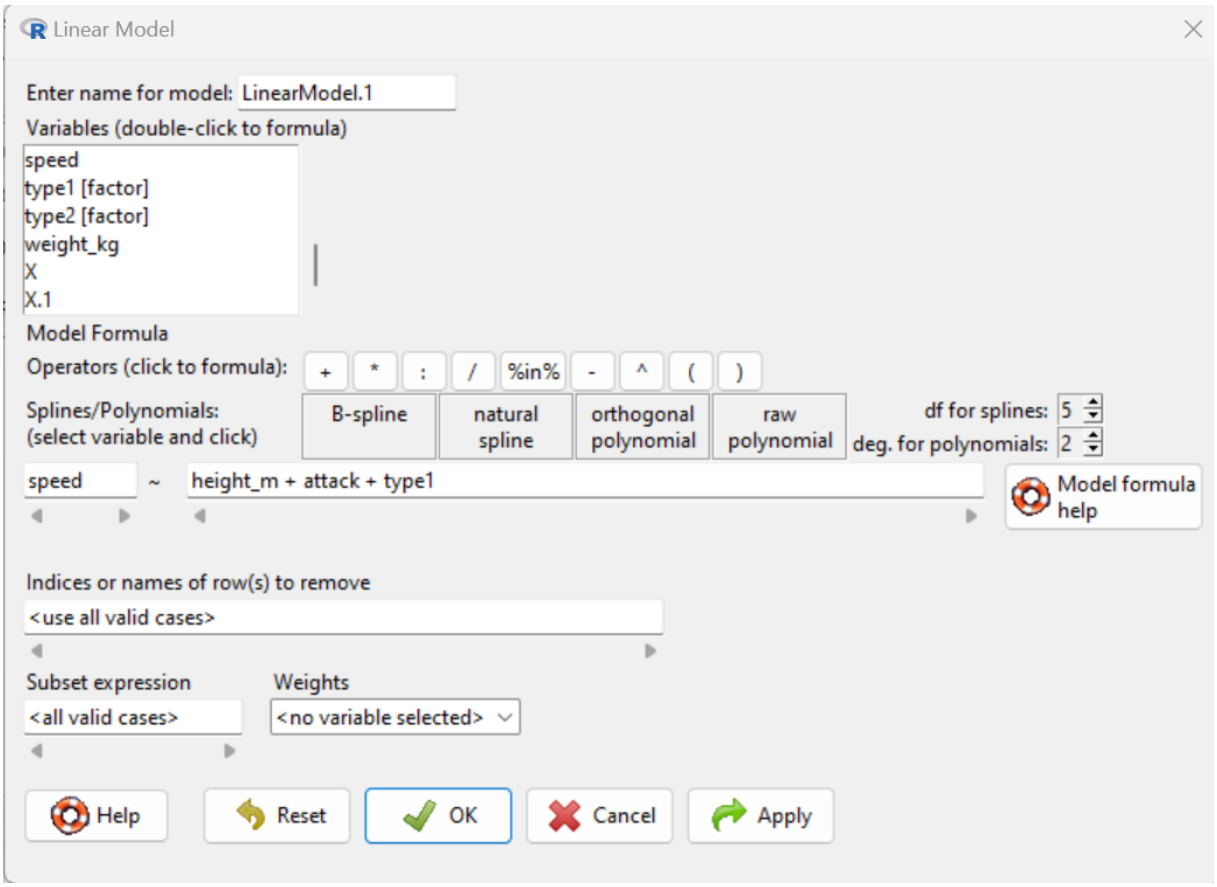
3. Run the model

Let's run the model to investigate the potential significant effects.

We write the model and run it

BUTTON CLICK TEAM





CODING TEAM

The lm() function takes as argument a formula of the form y ~ x, where y is the dependent variable and x is the independent variable.

```
mod <- lm(speed ~ height_m + attack + type1, data = data)
```

We retrieve the ANOVA table

The ANOVA table is what gives us access to the significativity of the variables/parameters.

We all switch to TEAM CODING.

Retrieve the ANOVA table using the Anova function (from the 'car' package).

If you have called your model 'mod', this gives :

```
Anova(mod)
```

I'll leave you to interpret the Anova table.

Recover the model summary

The model summary allows you to retrieve the model's R^2 (~ information, % variation of response variable explained by explanatory variables), as well as certain pairwise comparisons (for a qualitative explanatory variable) or the slope (for a quantitative expanatory variable).

Retrieve the summary table with the summary function.

CLICK BUTTON TEAM

In the previous procedure, R has already returned it to you.

CODING TEAM

```
summary(mod)
```

Interpret the model's R^2? What can we say about it?

Interpret the summary of the model? What can you say about it?

4. Run post-hocs tests

If the differences are significant, then we perform multiple comparison tests, also known as post-hoc tests.

Download the emmeans package.

Use the emmeans function to perform post-hoc/multiple comparison tests.

```
emmeans(mod,pairwise~type1)
```

5. Application condition check

You need to check that the model meets the conditions of application (linearity, normality, homoscedasticity and outliers).

We all switch to TEAM CODING.

```
par(mfrow=c(2,2)) #to get the four graphs on the same page
plot(mod) #graphs to check application conditions on the residuals
```

6. Adapt

If the application conditions are not right, adapt.

If you need to remove a specific line, you can do it on excel or directly on R. Here is the code to remove for example line 13

```
data2 <- data[-13, ]
```

II. BONUS

I'll leave you to do the same for the question of factors influencing happiness in Pokemon or test your new skills on another dataset.