SMDE FIRST ASSIGNEMENT (20% OF THE FINAL MARK, INDIVIDUAL)

THIRD QUESTION: DEFINE A LINEAR MODEL FOR AN ATHLETE IN THE 1500 M (25% OF THE FIRST ASSIGMENT).

To start: load the package RCmdrPlugin.FactoMinerR.

Load the data “decathlon” located in the package.

The data represents a data frame with observations for different athletes.

What is the linear expression that better predicts the behavior of an athlete for 1500m?

Explore different expressions describing the power and the features of each one of them.

Justify your answers.

Remember to test the assumptions of the linear model.

Now use the expression to **predict** the behavior for a specific athlete.

Analyze and explain the results obtained.

Is the model accurate? What do you expect?

Justify your answers.

ANNEX: THE DISTINCTION BETWEEN CONFIDENCE INTERVALS, PREDICTION INTERVALS AND TOLERANCE INTERVALS.

When you fit a parameter of a model, the accuracy or precision can be expressed as (i) confidence interval, (ii) prediction interval or (iii) tolerance interval. Assume that the data really are randomly sampled from a Gaussian distribution.

Confidence intervals tell you about how well you have determined the mean. If you do this many times, and calculate a confidence interval of the mean from each sample, you'd expect about 95 % of those intervals to include the true value of the population mean. The key point is that the confidence interval tells you about the likely location of the true population parameter.

Prediction intervals tell you where you can expect to see the next data point sampled. Collect a sample of data and calculate a prediction interval. Then sample one more value from the population. If you do this many times, you'd expect that next value to lie within that prediction interval in 95% of the samples. The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean.

Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus data scatter. So a prediction interval is always wider than a confidence interval.