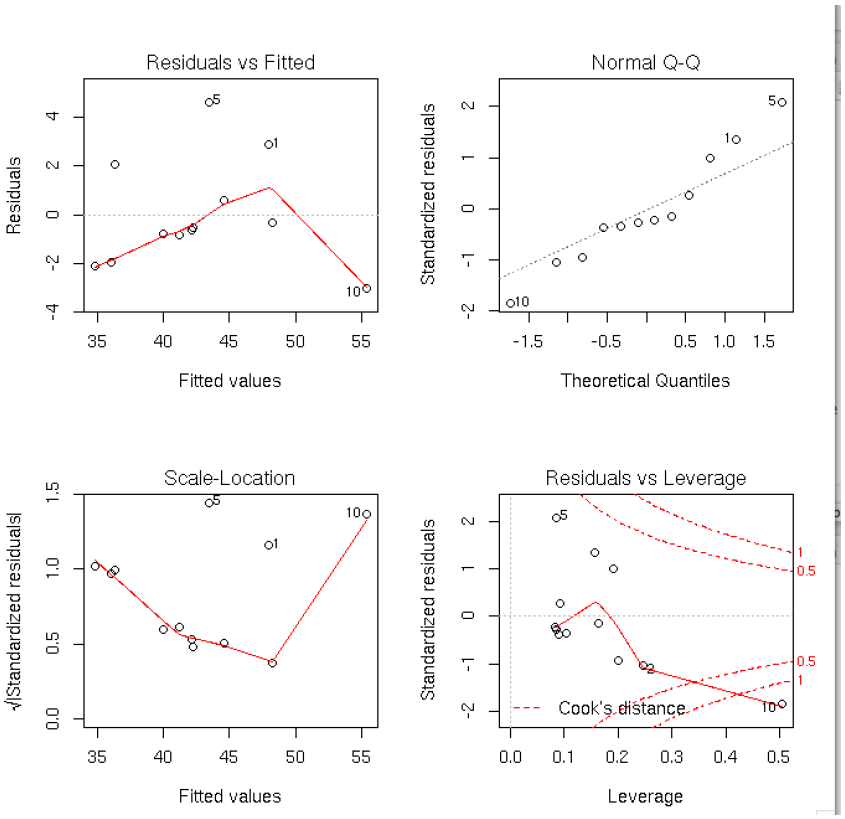
Understanding Diagnostic Plots for Linear Regression Analysis

After running a regression analysis, you should check if the model works well for data.
To do so, you have to pay attention to regression results, such as slope coefficients, p-values, or R2 . They will tell us how well a model represents given data.
Residuals could show how poorly a model represents data.

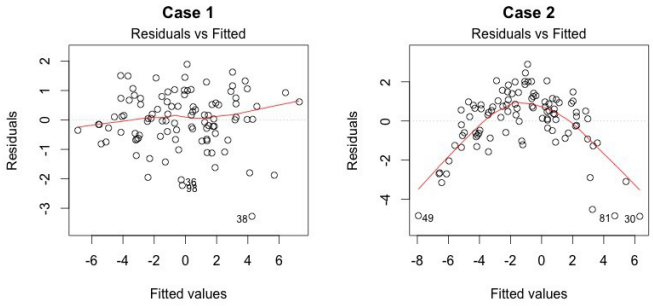
use a plot() to an lm object after running an analysis. Then R will show you four diagnostic plots one by one. For example:

```
#Q3.2.: LINEAR REGRESSION
plot(Cd_muscle~Cd_foie,sent2)#Relation lineaire observ?e
reg<-lm(Cd_muscle~Cd_foie,sent2)
x11();par(mfrow=c(2,2));plot(reg)
```



[1] a plot of residuals against fitted values

To interpret residuals vs. fitted plot for verifying the assumptions of a linear model.
The plot should indicate that the residuals and the fitted values are uncorrelated, as they should be in a homoscedastic linear model with normally distributed errors. On the contrary, the plot souldn't indicate dependency between the residuals and the fitted values, suggesting a different model.
This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.

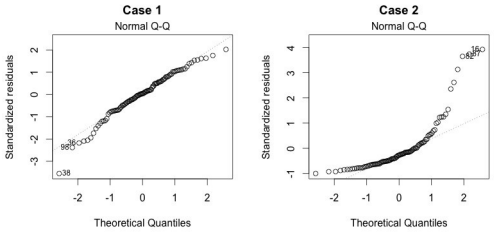


[2] a Normal Q-Q plot,

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

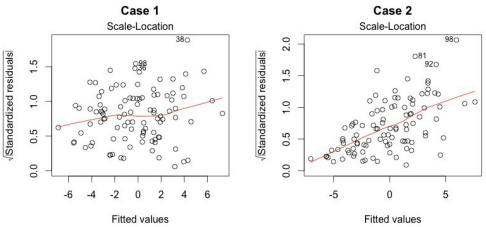
What are "quantiles"? These are often referred to as "percentiles". These are points in your data below which a certain proportion of your data fall.

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.

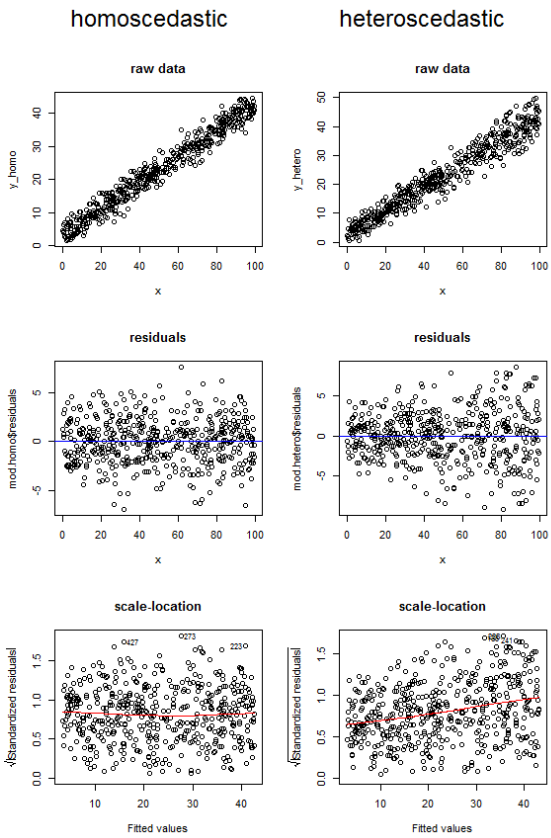


[3] a Scale-Location plot of sqrt(| residuals |) against fitted values,

It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points. It Should be a homogeneous cloud, that is not taller on one side than the other



Consider the plots below, which compare how homoscedastic vs. heteroscedastic data might look in these three different types of figures. Note the funnel shape for the upper two heteroscedastic plots, and the upward sloping lowess line in the last one.



[4] a plot of Cook's distances versus row labels

« En statistiques, la distance de Cook est couramment utilisée pour estimer l'influence d'une donnée lors de l'utilisation de méthode des moindres carrés. La distance de Cook mesure l'effet de la suppression d'une donnée. Les données avec d'importants résidus (Données aberante) et/ou fort effet de levier peuvent fausser le résultat et la précision d'une régression. Les points ayant une distance de Cook importante sont considérées comme méritant un examen plus approfondi dans l'analyse. »[1]

« In statistics, Cook's distance or Cook's D is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Cook's distance measures the effect of deleting a given observation. Points with a large Cook's distance are considered to merit closer examination in the analysis. »[2]

Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.

1" http://data.library.virginia.edu/diagnostic-plots/

2" http://data.library.virginia.edu/diagnostic-plots/