Help on Cook's distance method

Cook's distance, is used in Regression Analysis to find influential outliers in a set of predictor variables. In other words, it's a way to identify points that negatively affect your regression model. The measurement is a combination of each observation's leverage and residual values; the higher the leverage and residuals, the higher the Cook's distance. Technically, Cook's distance is calculated by removing the i_{th} data point from the model and recalculating the regression. It summarizes how much all the values in the regression model change when the i_{th} observation is removed. The formula for Cook's distance is as following:

$$D_i = \frac{\Sigma_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{(p+1)\hat{\sigma}^2}$$

where, the nominator is the sum of squared errors when excluding observation i and p is the number of variables in the regression model. Several interpretations for Cook's distance exist: A general rule of thumb is that observations with a Cook's D_i of more than 3 times the mean, μ, is a possible outlier. An alternative interpretation is to investigate any point over 4/n. Screen example with p = 3 and cutoff level of 4 (times the mean, μ) is depicted below: