Cook's distance

Souvik Roy

October 20, 2022

1 Introduction

We have already learnt about DFFITS, where we consider the influence of a certain y_i on the corresponding predicted value $\hat{y}_i$ (notice the same indices i). Here, we delete the point y_i, and try obtain $\hat{y}_i$ based on the remaining $n-1$ points. But this approach has some problems. One of them is that it sheds no light on how dropping one point, say y_j, influences the predicted value for y_k, namely $\hat{y}_k$, when $k \neq j$. On the other hand, if we are interested in doing this for all the pairs of points, then summarizing the results obtained after doing so turns out to be a challenging task. The **Cook's distance** helps us especially in this regard.

2 Cook's distance

The *ith Cook's distance*, D_i, is defined to be

$$D_i = \frac{\sum_j (\hat{y}_j - \hat{y}_j(i))^2}{(p+1)\hat{\sigma^2}},$$

where, as usual,

- $\hat{y}_j(i)$ is the predicted value of the jth observation when the original ith point y_i, has been excluded,
- $\hat{y}_j$ is the original predicted value for y_j,
- $\hat{\sigma^2}$ is the estimator for σ^2 over the **whole** data and **not** after excluding the point y_i,
- and $p+1$ is the total number of columns in the matrix X, which corresponds to p many regressors, and one intercept term. The '$+1$' might be dropped if there is no intercept term.

It is worthwhile to remember that the Cook's distance is not for the ANOVA setup, instead it is mainly useful for the multiple regression setup, as all the variables are continuous. Upon calculating the D_i for each i, we stack them up as a vector $\vec{D}$, and that can be used to summarize the influence of each point on all the other points once it is removed. In particular, if any of the components of $\vec{D}$ is large, then we suspect that particular point to have a strong influence on the quality of the fit.