Res. Lett. Inf. Math. Sci. (2002), **3**, 127-128
Available online at http://www.massey.ac.nz/~wwiims/research/letters/

A Teaching Note on Cook's Distance – A Guideline

Barry McDonald
Institute of Information and Mathematical Science, Massey University at Albany, Auckland, N.Z.
B.McDonald@massey.ac.nz

Abstract
Cook's Distance (D_i) is used for assessing influence in regression models. The usual criterion is that a point is influential if D_i exceeds the median of the $F_{p,n-p}$ distribution, where p is the number of regression parameters. The practice developed here at Massey Albany is to teach the guideline $D_i>0.7$ for $p=2$, $D_i>0.8$ for $p=3$, $D_i>0.85$ for $p>3$, where $n\geq 15$. It is not known if this guideline is used elsewhere.

Teaching Guideline
Cook's Distance (D_i) is an influence measure based on the difference between the regression parameter estimates $\hat{\beta}$ and what they become if the ith data point is deleted, $\hat{\beta}_{-i}$, say.

There are numerical rules for assessing Cook's D_i but the rules tend to be rough guidelines, and textbook authors differ in their advice. The most common criterion quoted around the world appears to be to declare the ith point influential if D_i exceeds the median of the $F_{p,\,n-p}$ distribution, where p is the number of regression coefficients (including the intercept) and n the number of data. This guideline is justified on a mixture of theoretical and practical grounds. However $F_{p,n-p}(0.5)$ is seldom tabulated, and needs to be computed for each situation, which seems unnecessary effort for a rough-and-ready guideline.

To simplify matters, Chatterjee, Hadi and Price (2000) quote $D_i>1$ as an operational guideline. However this generally diverges from $F_{p,n-p}(0.5)$ except for very small and overparameterised datasets, as illustrated by the plotted medians in Figure 1. One could argue that this is not a really a problem - for example Chatterjee *et al*. point out that the important thing is to graph the Cook's D_i values, to see whether any one or two points have a much bigger D_i than the others. If there are one or two points with relatively high D_i points, then we would be inclined to suspect these higher points of being influential even if no points break the $F_{p,n-p}(0.5)$ rule or the $D_i>1$ rule. And the same *relative size* interpretation applies even if many of D_i values exceed the chosen guideline, but one or two stand out well in excess of the others.

Nevertheless I argue that it is one thing to have a numerical guideline that has to be interpreted with thought, but quite another thing to have a guideline that one knows diverges markedly from the common one used around the world, except in the case of rather unusual p and n. This seems particularly important in a teaching situation, where students may slavishly follow a numerical rule. Hence I developed a compromise guideline for the Regression courses at Massey University. Specifically, I teach that Cook's distance should be assessed by the following guideline, based on Figure 1.

For datasets with $n>15$, we can consider points as influential:
if $D_i > 0.7$ for $p=2$, (one predictor)
if $D_i > 0.8$ for $p=3$, (two predictors)
and if $D_i > 0.85$ for $p>3$, (more than predictors).

For smaller datasets, a guideline could be obtained from the Figure or by direct calculation, but this would rarely if ever by needed. The guideline simplifies the unnecessary precision of the $F_{p,n-p}(0.5)$ rule but avoids the one-size-fits-all approach when clearly one size does not fit all situations.

It is not known if this guideline is used elsewhere in the world. The purpose of this note is therefore to submit it for scrutiny and possible adoption.

Reference

Chatterjee, S, Hadi, A.S., Price, B. *Regression Analysis by Example , 3rd edition*, Wiley, New York, p 104, (2000)

Figure 1.