Cook's distance

From Wikipedia, the free encyclopedia

In statistics, **Cook's distance** or **Cook's *D*** is a commonly used estimate of the influence of a data point when performing least squares regression analysis.[1] In a practical ordinary least squares analysis, Cook's distance can be used in several ways: to indicate data points that are particularly worth checking for validity; to indicate regions of the design space where it would be good to be able to obtain more data points. It is named after the American statistician R. Dennis Cook, who introduced the concept in 1977.[2][3]

Contents [hide]

Definition [edit]

Cook's distance measures the effect of deleting a given observation. Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Points with a large Cook's distance are considered to merit closer examination in the analysis. It is calculated as:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \ \mathrm{MSE}},$$

where:

- $\hat{Y}_j$ is the prediction from the full regression model for observation *j*;
- $\hat{Y}_{j(i)}$ is the prediction for observation *j* from a refitted regression model in which observation *i* has been omitted;
- p is the number of fitted parameters in the model;
- MSE is the mean square error of the regression model.

The following are the algebraically equivalent expressions (in case of simple linear regression):

$$D_i = \frac{e_i^2}{p \ \mathrm{MSE}} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right],$$

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{(-i)})^T (X^T X)(\hat{\beta} - \hat{\beta}^{(-i)})}{(1+p)s^2},$$

where:

h_{ii} is the leverage, i.e., the i-th diagonal element of the hat matrix $\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$;

e_i is the residual (i.e., the difference between the observed value and the value fitted by the proposed model).

Detecting highly influential observations [edit]

There are different opinions regarding what cut-off values to use for spotting highly influential points. A simple operational guideline of $D_i > 1$ has been suggested.[4] Others have indicated that $D_i > 4/n$, where n is the number of observations, might be used.[5]

A conservative approach relies on the fact that Cook's distance has the form W/p, where W is formally identical to the Wald statistic that one uses for testing that $H_0 : \beta_i = \beta_0$ using some $\hat{\beta}_{[-i]}$.[*citation needed*] Recalling that W/p has an $F_{p,n-p}$ distribution (with p and n-p degrees of freedom), we see that Cook's distance is equivalent to the F statistic for testing this hypothesis, and we can thus use $F_{p,n-p,1-\alpha}$ as a threshold.

Interpretation [edit]

Specifically D_i can be interpreted as the distance one's estimates move within the confidence ellipsoid that represents a region of plausible values for the parameters.[*clarification needed*] This is shown by an alternative but equivalent representation of Cook's distance in terms of changes to the estimates of the regression parameters between the cases where the particular observation is either included or excluded from the regression analysis.

See also [edit]

- Outlier
- Leverage (statistics)
- Partial leverage
- DFFITS
- Studentized residual

References [edit]

1. ^ Mendenhall, William; Sincich, Terry (1996). *A Second Course in Statistics: Regression Analysis* (5th ed.). Upper Saddle River, NJ: Prentice-Hall. p. 422. ISBN 0-13-396821-9. "A measure of overall influence an outlying observation has on the estimated β coefficients was proposed by R. D. Cook (1979). Cook's distance, D_i, is calculated..."
2. ^ Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression". *Technometrics* (American Statistical Association) **19** (1): 15–18. doi:10.2307/1268249. JSTOR 1268249. MR 0436478.
3. ^ Cook, R. Dennis (March 1979). "Influential Observations in Linear Regression". *Journal of the*

American Statistical Association (American Statistical Association) **74** (365): 169–174. doi:10.2307/2286747. JSTOR 2286747. MR 0529533.

4. ^ Cook, R. Dennis; Weisberg, Sanford (1982). *Residuals and Influence in Regression*. New York, NY: Chapman & Hall. ISBN 0-412-24280-X.
5. ^ Bollen, Kenneth A.; Jackman, Robert W. (1990). Fox, John; Long, J. Scott, eds. *Modern Methods of Data Analysis*. Newbury Park, CA: Sage. pp. 257–91. ISBN 0-8039-3366-5.

Further reading [edit]

- Atkinson, Anthony; Riani, Marco (2000). "Deletion Diagnostics". *Robust Diagnostics and Regression Analysis*. New York: Springer. pp. 22–25. ISBN 0-387-95017-6.
- Chatterjee, Samprit; Hadi, Ali S. (2006). *Regression analysis by example* (4th ed.). John Wiley and Sons. ISBN 0-471-74696-7.
- Heiberger, Richard M.; Holland, Burt (2013). "Case Statistics". *Statistical Analysis and Data Display*. Springer Science & Business Media. pp. 312–27. ISBN 9781475742848.
- Lorenz, Frederick O. (April 1987). "Teaching about Influence in Simple Regression". *Teaching Sociology* (American Sociological Association) **15** (2): 173–177. doi:10.2307/1318032. JSTOR 1318032.