Leverage (statistics)

In statistics and in particular in regression analysis, **leverage** is a measure of how far away the independent variable values of an observation are from those of the other observations.

High-leverage points are those observations, if any, made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation.[1]

Modern computer packages for statistical analysis include, as part of their facilities for regression analysis, various quantitative measures for identifying influential observations among these measures is partial leverage, a measure of how a variable contributes to the leverage of a datum.

Contents

Linear regression model

Definition

In the linear regression model, the leverage score for the i-th data unit is defined as:

$$h_{ii} = [\mathbf{H}]_{ii},$$

the i-th diagonal element of the projection matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}$, where $\mathbf{X}$ is the design matrix. The leverage score is also known as the observation self-sensitivity or self-influence,[2] as shown by

$$h_{ii} = \frac{\partial \widehat{y}_i}{\partial y_i},$$

where $\widehat{y}_i$ and y_i are the fitted and measured observation, respectively

Bounds on leverage

$$0 \leq h_{ii} \leq 1.$$

Proof

First, note that H is an idempotent matrix: $\boldsymbol{H^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = XI(X^\top X)^{-1}X^\top = H.}$ Also, observe that $\boldsymbol{H}$ is symmetric. So equating the ii element of H to that of H^2, we have

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \geq 0$$

and

$$h_{ii} \geq h_{ii}^2 \implies h_{ii} \leq 1.$$

Effect on residual variance

If we are in an ordinary least squares setting with fixed X, regression errors $\boldsymbol{\varepsilon_i}$, and

$$\begin{aligned} Y &= X\beta + \varepsilon \\ \operatorname{Var}(\varepsilon) &= \sigma^2 I \end{aligned}$$

then $\operatorname{Var}(e_i) = (1 - h_{ii})\sigma^2$ where $e_i = Y_i - \widehat{Y}_i$ (the i th regression residual).

In other words, if the model errors $\boldsymbol{\epsilon}$ are homoscedastic, an observation's leverage score determines the degree of noise in the model's misprediction of that observation.

Proof

First, note that $\boldsymbol{I - H}$ is idempotent and symmetric. This gives,

$$\operatorname{Var}(e) = \operatorname{Var}((I - H)Y) = (I - H)\operatorname{Var}(Y)(I - H)^\top = \sigma^2(I - H)^2 = \sigma^2(I - H).$$

Thus $\operatorname{Var}(e_i) = (1 - h_{ii})\sigma^2.$

Studentized residuals

The corresponding studentized residual—the residual adjusted for its observation-specific estimated residual variance—is then

$$t_i = \frac{e_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\widehat{\boldsymbol{\sigma}}$ is an appropriate estimate of $\boldsymbol{\sigma}$.

See also

- Projection matrix – whose main diagonal entries are the leverages of the observations
- Mahalanobis distance – a measure of leverage of a datum
- Cook's distance – a measure of changes in regression coefficients when an observation is deleted
- DFFITS
- Outlier – observations with extreme Y values

References

1. Everitt, B. S. (2002). *Cambridge Dictionary of Statistics* Cambridge University Press. ISBN 0-521-81099-X.
2. Cardinali, C. (June 2013). "Data Assimilation: Observation influence diagnostic of a data assimilation system" (https://www.ecmwf.int/sites/default/files/elibrary/2013/16938-observation-influence-diagnostic-data-assimilation-system.pdf) (PDF).

WIKIPEDIA

Cook's distance

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.[1] In a practical ordinary least squares analysis, Cook's distance can be used in several ways: to indicate influential data points that are particularly worth checking for validity; or 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

Definition

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.

For the algebraic expression, first define

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$ is the error term, $\boldsymbol{\beta} = [\beta_0\, \beta_1 \ldots \beta_{p-1}]^\mathsf{T}$ is the coefficient matrix, p is the number of covariates or predictors for each observation, and $\mathbf{X}$ is the design matrix including a constant. The least squares estimator then is $\mathbf{b} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$, and consequently the fitted (predicted) values for the mean of $\mathbf{y}$ are

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H} \equiv \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$ is the projection matrix (or hat matrix). The i-th diagonal element of $\mathbf{H}$, given by $h_i \equiv \mathbf{x}_i^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_i$,[4] is known as the leverage of the i-th observation. Similarly, the i-th element of the residual vector $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\,\mathbf{y}$ is denoted by e_i.

Cook's distance D_i of observation i $(\text{for } i = 1, \ldots, n)$ is defined as the sum of all the changes in the regression model when observation i is removed from it[5]

$$D_i = \frac{\sum_{j=1}^{n}\left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{ps^2}$$

where $\widehat{y}_{j(i)}$ is the fitted response value obtained when excluding i, and $s^2 \equiv (n-p)^{-1}\mathbf{e}^\top\mathbf{e}$ is the mean squared error of the regression model.[6] Equivalently, it can be expressed using the leverage[5]

$$D_i = \frac{e_i^2}{s^2 p}\left[\frac{h_i}{(1-h_i)^2}\right].$$

Detecting highly influential observations

There are different opinions regarding what cut-off values to use for spotting highly influential points. Since Cook's distance is in the metric of an F distribution with p and $n-p$ (as defined for the design matrix $\mathbf{X}$ above) degrees of freedom, the median point (i.e., $F_{0.5}(p, n-p)$) can be used as a cut-off.[7] Since this value is close to 1 for large n, a simple operational guideline of $D_i > 1$ has been suggested.[8] Note that the Cook's distance measure does not always correctly identify influential observations.[9]

Interpretation

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. 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

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

References

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_j, 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 (https://doi.org/10.2307%2F1268249). JSTOR 1268249 (https://www.jstor.org/stable/1268249). MR 0436478 (https://www.ams.org/mathscinet-getitem?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 (https://doi.org/10.2307%2F2286747). JSTOR 2286747 (https://www.jstor.org/stable/2286747). MR 0529533 (https://www.ams.org/mathscinet-getitem?mr=0529533).
4. Hayashi, Fumio (2000). *Econometrics* (https://books.google.com/books?id=QyIW8WUIyzcC&pg=PA21). Princeton University Press. pp. 21–23.
5. "Cook's Distance" (http://se.mathworks.com/help/stats/cooks-distance.html).
6. "Statistics 512: Applied Linear Models" (https://www.stat.purdue.edu/~jennings/stat514/stat512notes/topic3.pdf#page=9) (PDF). *Purdue University*.

7. Bollen, Kenneth A.; Jackman, Robert W. (1990). "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases". In Fox, John; Long, J. Scott. *Modern Methods of Data Analysis*. Newbury Park, CA: Sage. pp. 257–91 [p. 266]. ISBN 0-8039-3366-5.
8. Cook, R. Dennis; Weisberg, Sanford (1982). *Residuals and Influence in Regression* (https://books.google.com/books?id=MVSqAAAAIAAJ). New York, NY: Chapman & Hall. ISBN 0-412-24280-X.
9. Kim, Myung Geun (31 May 2017). "A cautionary note on the use of Cook's distance" (http://www.csam.or.kr/journal/view.html?doi=10.5351/CSAM.2017.24.3.317). *Communications for Statistical Applications and Methods*. **24** (3): 317–324. doi:10.5351/csam.2017.24.3.317 (https://doi.org/10.5351%2Fcsam.2017.24.3.317). ISSN 2383-4757 (https://www.worldcat.org/issn/2383-4757).

Further reading

- Atkinson, Anthony; Riani, Marco (2000). "Deletion Diagnostics" (https://books.google.com/books?id=X0dPBOJ_L4UC&pg=PA22). *Robust Diagnostics and Regression Analysis*. New York: Springer. pp. 22–25. ISBN 0-387-95017-6.
- Heiberger, Richard M.; Holland, Burt (2013). "Case Statistics" (https://books.google.com/books?id=co3gBwAAQBAJ&pg=PA312). *Statistical Analysis and Data Display*. Springer Science & Business Media. pp. 312–27. ISBN 9781475742848.
- Krasker, William S.; Kuh, Edwin; Welsch, Roy E. (1983). "Estimation for dirty data and flawed models". *Handbook of Econometrics*. **1**. Elsevier. pp. 651–698. doi:10.1016/S1573-4412(83)01015-6 (https://doi.org/10.1016%2FS1573-4412%2883%2901015-6).
- Aguinis, Herman; Gottfredson, Ryan K.; Joo, Harry (2013). "Best-Practice Recommendations for Defining Identifying and Handling Outliers" (https://www.researchgate.net/profile/Herman_Aguinis/publication/258174106_Best-Practice_Recommendations_for_Defining_Identifying_and_Handling_Outliers/links/004635276b1ff93ba8000000.pdf) (PDF). *Organizational Research Methods*. Sage. **16** (2): 270–301. doi:10.1177/1094428112470848 (https://doi.org/10.1177%2F1094428112470848). Retrieved 4 December 2015.