San José State University

Math 261A: Regression Theory & Methods

Diagnostics for Leverage and Influence

Dr. Guangliang Chen

This lecture is based on the following part of the textbook:

- Sections 6.1 – 6.7

Outline of the presentation:

- **Outliers** (and their leverage and influence)
- **Influence measures**
- **How to handle influential points**

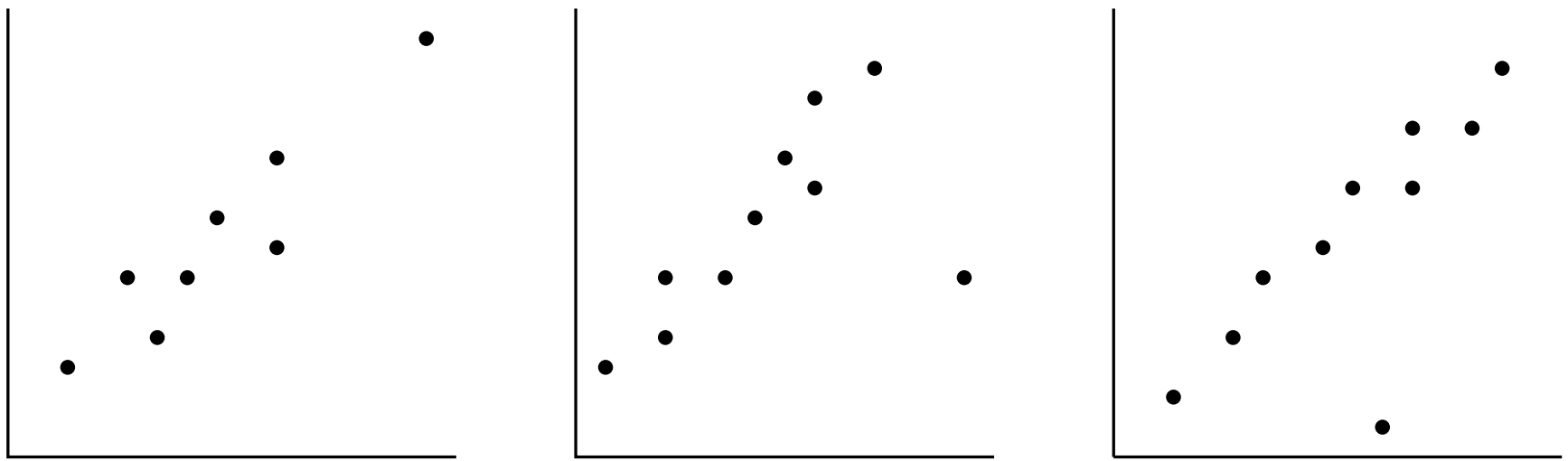
What are outliers?

Outliers are **extreme observations** that are considerably different from the majority of the data:

- They can be caused by recording errors (in which case, they should be removed), or
- they are just cases that cannot be explained by the model but carry important information.

An outlier may be outlying in x-space, y-space, or both.

Depending on their location in space, outliers can have moderate to severe effects on the regression model.



Def 0.1 (Leverage and influence points)**.**

- A **leverage point** is an observation that has an unusual predictor value (very different from the bulk of the observations).
- An **influence point** is an observation whose removal from the data set would cause a large change in the estimated reggression model coefficients.

A leverage point may have no influence if the observation lies close to the regression line.

A point has to have at least some leverage in order to be influential.

Outliers may be detected by examining

- **Scaled residuals** (magnitudes > 3 implies potential outliers)
- **Residual plots** (also pay attention to magnitudes of residuals)
- **Normal quantile plots** (large departures from straight line)

Outliers must be treated with caution (we cannot simply remove them unless we are sure that they are indeed caused by errors).

This chapter is an extension and consolidation of some of these issues.

Leverage

The location of points in x space determines their leverage on the regression model, which is measured by the diagonal elements h_{ii} of the hat matrix:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

It can be shown that

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

and in the special setting of simple linear regression $(k = 1)$ that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Remark.

- Large diagonal elements of $\mathbf{H}$ reveal leverage observations that are potentially also influential;
- The average value of h_{ii} is $\bar{h} = (k+1)/n$, because

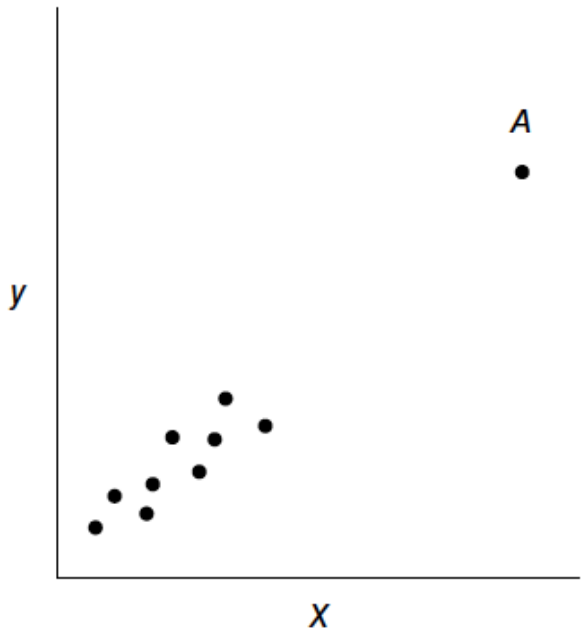
$$\sum_{i=1}^{n} h_{ii} = \text{trace}(\mathbf{H}) = \text{trace}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) = \text{trace}(\mathbf{I}) = k+1.$$

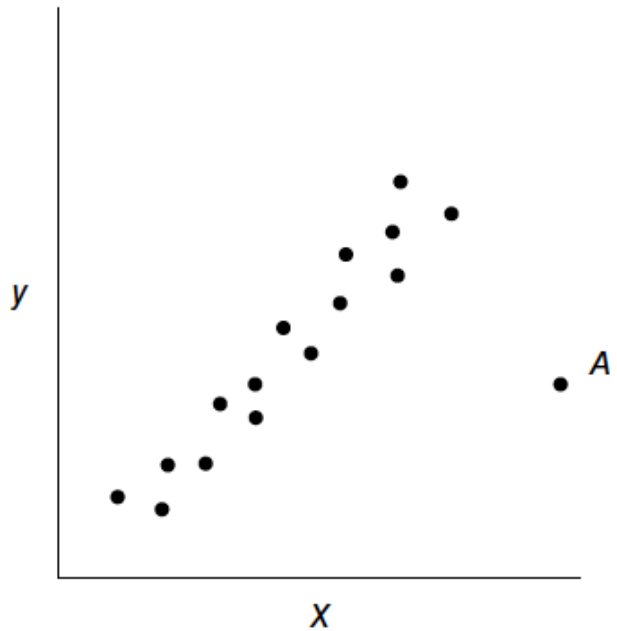
- It is traditionally assumed that any observation $\mathbf{x}_i$ with

$$h_{ii} > 2\bar{h} = 2(k+1)/n$$

is remote enough to be considered a leverage point.

Not all leverage points are influential, unless they have large residuals.

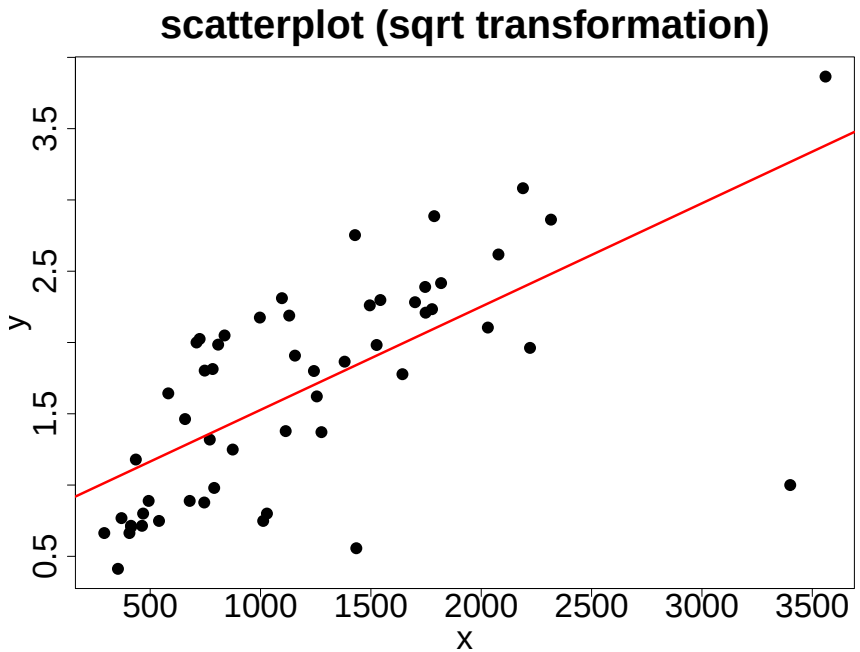


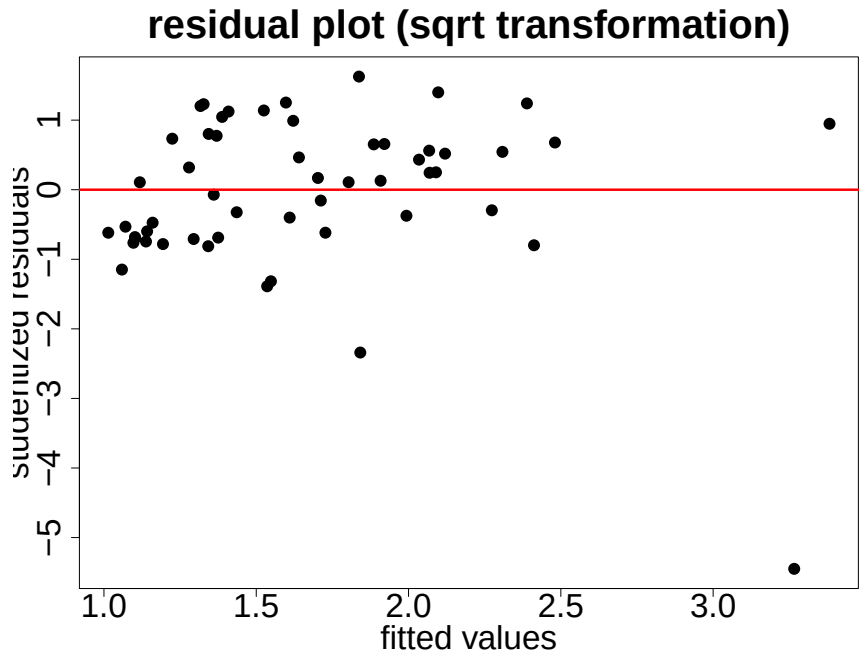


Observations with large values of h_{ii} and large residuals are likely to be influential.

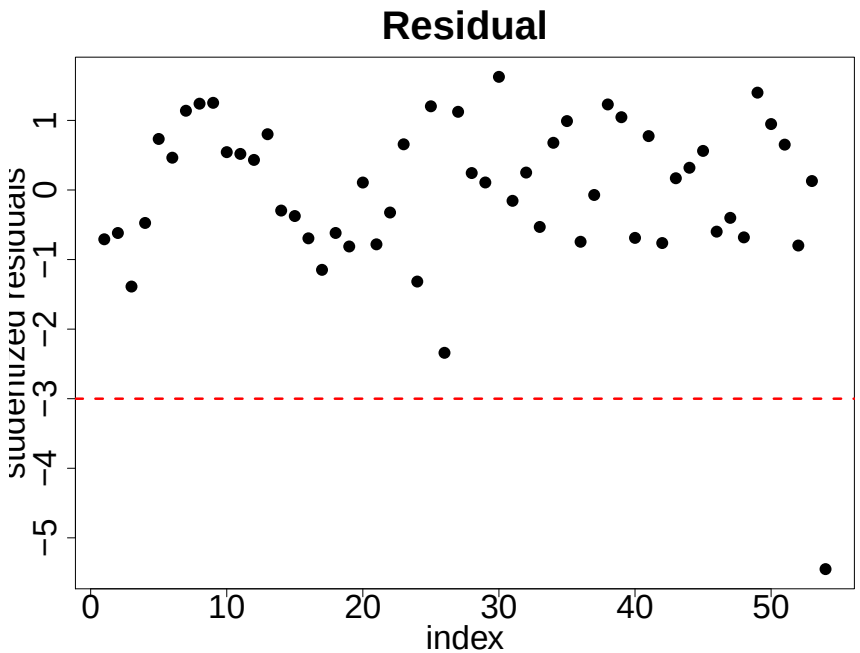
Example: ElectricUtility Data

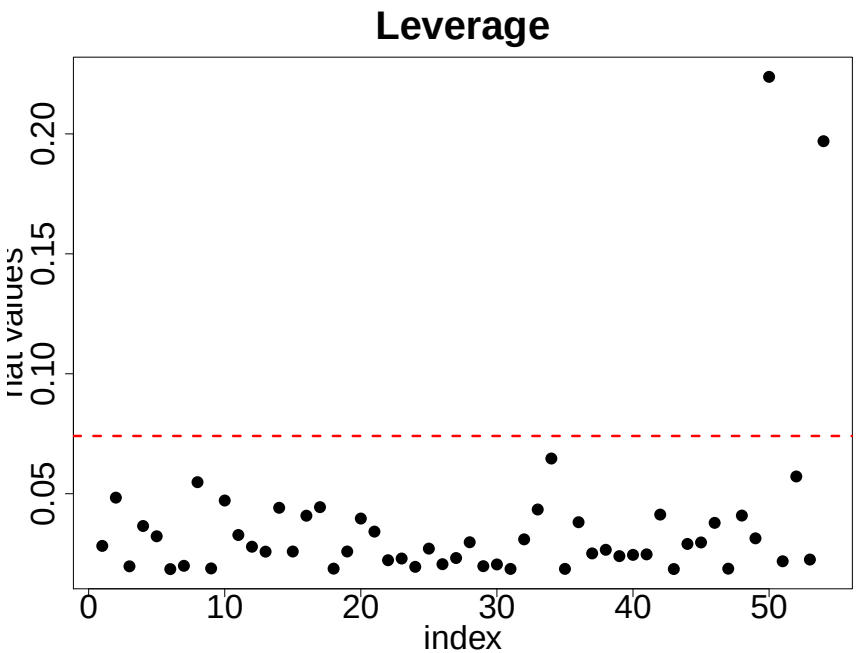
Consider this data set again, but with the square-root transformed response and an extra observation (in the bottom right corner).

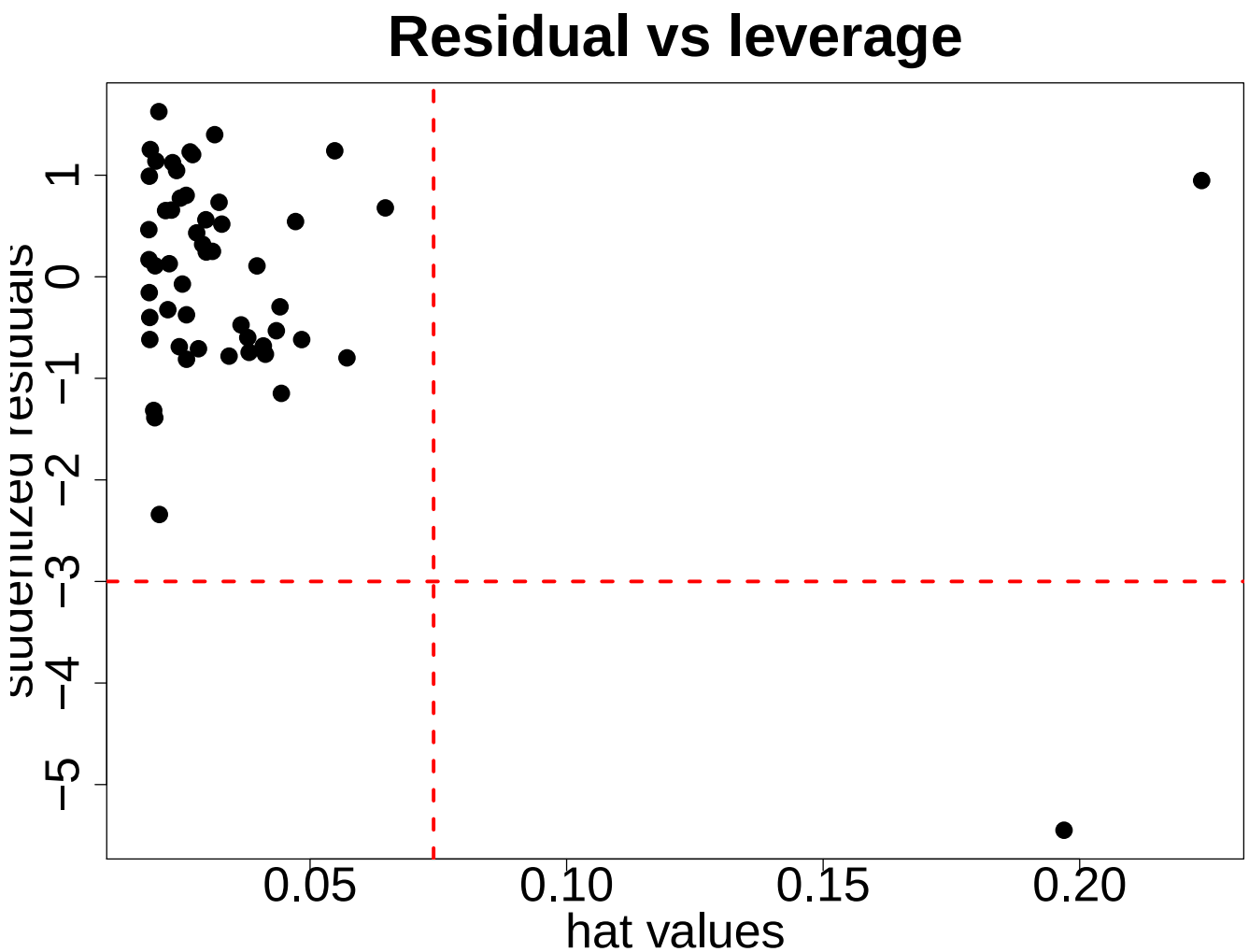




Diagnostics for influence points






Residual vs leverage
Standardized residuals
1
0
−1
−2
−3
−4
−5
0.05
0.10
0.15
0.20
hat values

Cook's distance measure

Cook [1977, 1979] suggested a way to consider both the location of the point in the x space and the response variable in measuring its influence:

$$D_i = \frac{\|\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}}\|^2}{p \cdot MS_{Res}} = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p \cdot MS_{Res}}$$

where

- $\hat{\boldsymbol{\beta}}$ (and $\hat{\mathbf{y}}$): least-squares estimate of regression coefficients based on all n points (and corresponding fitted values)
- $\hat{\boldsymbol{\beta}}_{(i)}$ (and $\hat{\mathbf{y}}_{(i)}$): least-squares estimate obtained by deleting the ith point (and corresponding fitted values)

Remark.

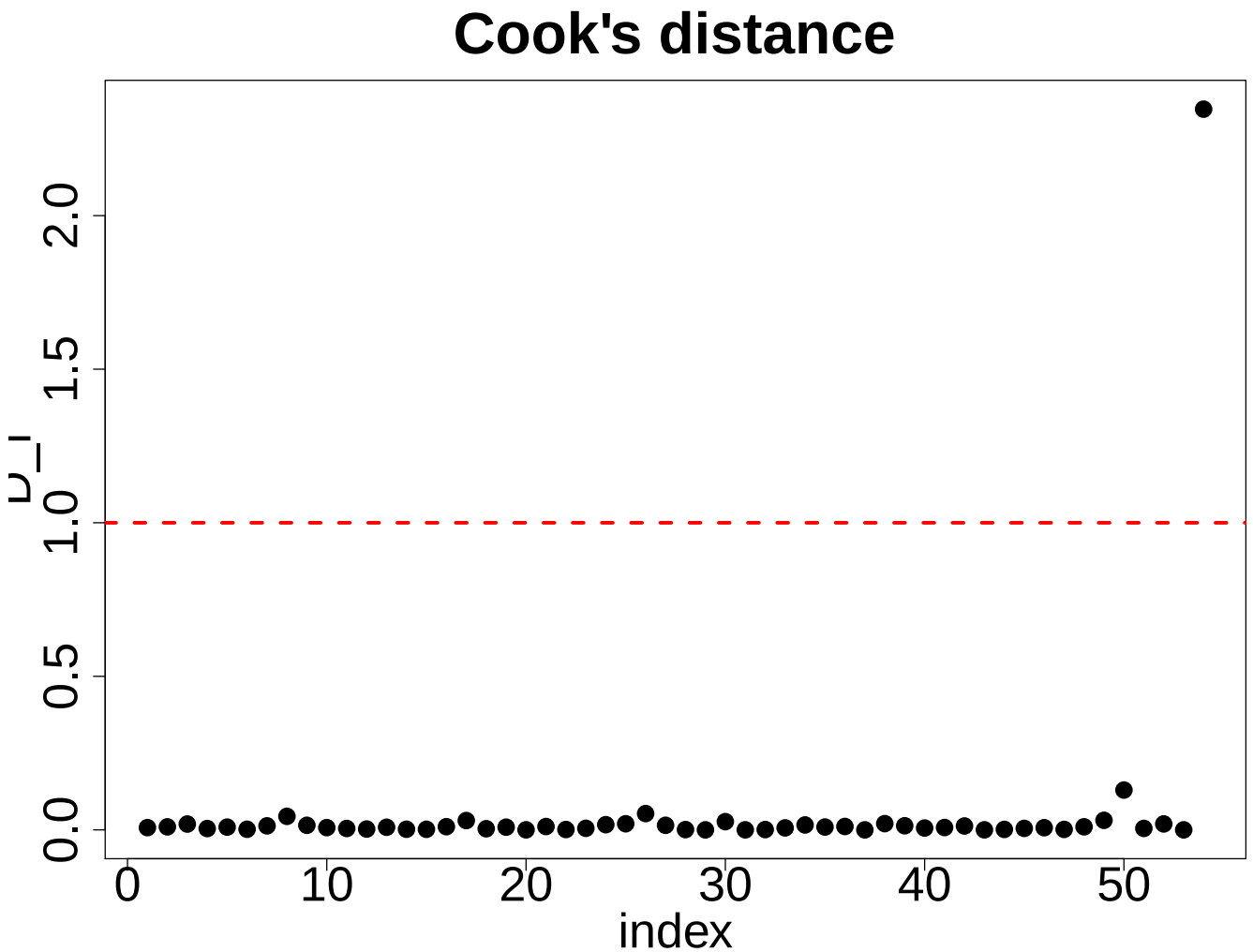
- It can be shown that for each $1 \leq i \leq n$,

$$D_i = \frac{r_i^2}{k+1} \cdot \frac{h_{ii}}{1-h_{ii}}$$

  where r_i is the ith studentized residual, and $\frac{h_{ii}}{1-h_{ii}}$ can be shown to be the distance from $\mathbf{x}_i$ to the centroid of the remaining data.

- D_i is made up of a component that reflects how well the model fits the ith observation y_i and a component that measures how far that point is from the rest of the data.

- **We usually consider points for which $D_i > 1$ to be influential.**


Cook's distance
D_i
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50
index

DFFITS and DFBETAS

Two other deletion diagnostics (besides Cook's distance) are

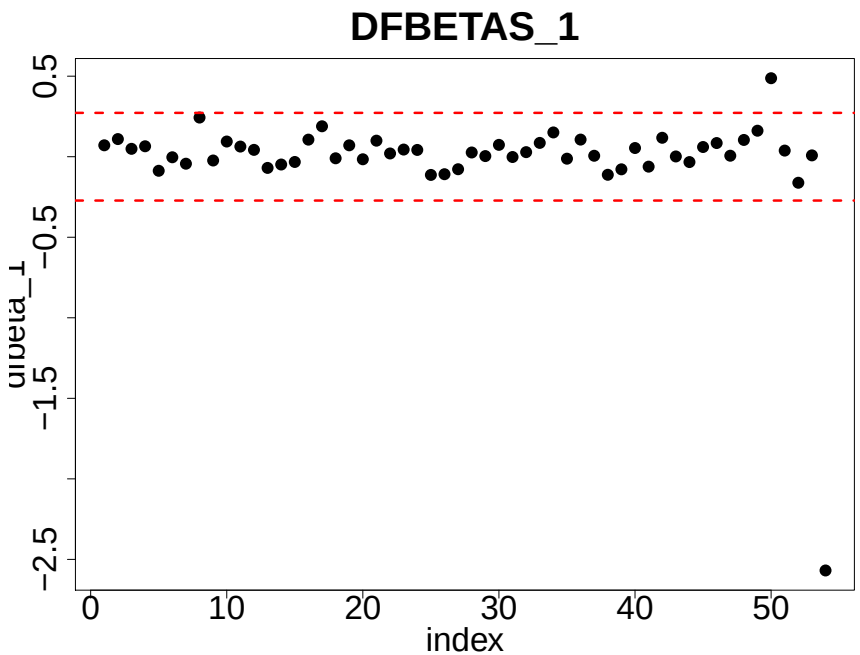
$$\text{DFBETAS}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S^2_{(i)} C_{jj}}}, \qquad \text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}$$

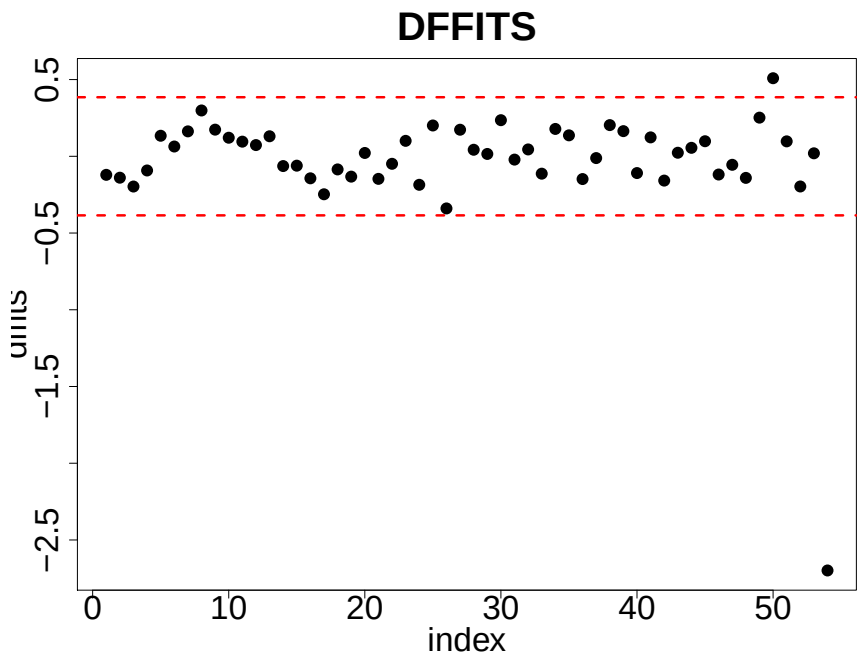
where C_{jj} is the jth diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$, and $S^2_{(i)}$ is the estimate of σ^2 without the ith observation.

Remark. $\text{DFBETAS}_{j,i}$ is a measure for how much the ith observation influences the value of $\hat{\beta}_j$, while DFFITS_i is a measure for how much the ith observation influences the value of $\hat{y}_i$.

The ith observation needs to be examined if

$$|\text{DFBETAS}_{j,i}| > 2/\sqrt{n} \text{ (for some } j\text{), or } |\text{DFFITS}_i| > 2\sqrt{p/n}$$





A measure of model performance

Effects on the precision of regression coefficient estimates is measured by

$$\text{COVRATIO}_i = \frac{\det(\text{Var}(\hat{\boldsymbol{\beta}}_{(i)}))}{\det(\text{Var}(\hat{\boldsymbol{\beta}}))} = \frac{\det\left((\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}S^2_{(i)}\right)}{\det\left((\mathbf{X}'\mathbf{X})^{-1}MS_{Res}\right)}, \quad i = 1, \ldots, n$$

Interpretation (*note that determinant of the covariance matrix is a scalar measure of precision*):

- If $\text{COVRATIO}_i > 1$, removing the ith observation degrades precision (and including it improves the precision of estimation);
- If $\text{COVRATIO}_i < 1$, removing the ith observation improves precision (and including it degrades the precision of estimation)

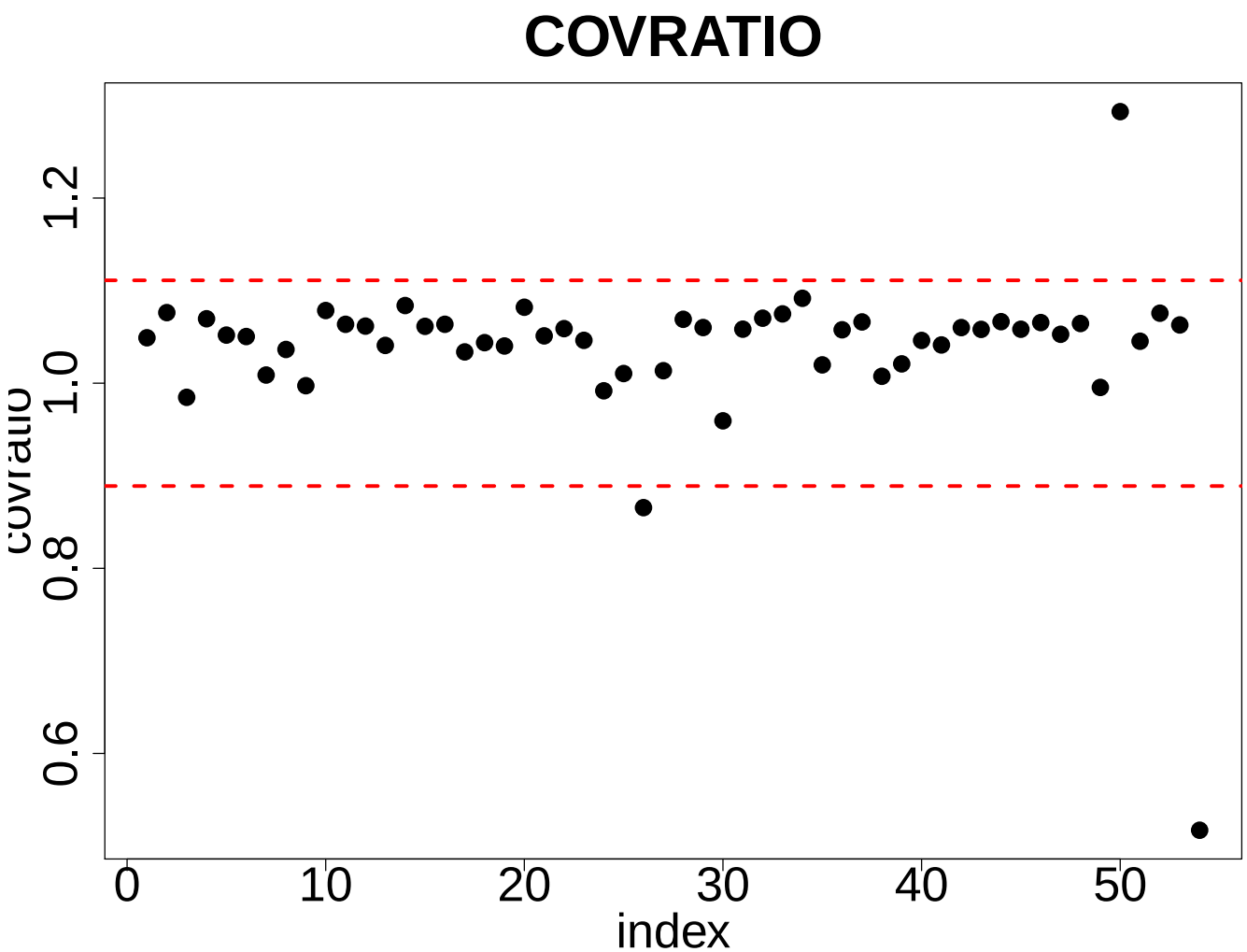
It can be shown that

$$\text{COVRATIO}_i = \left(\frac{S^2_{(i)}}{MS_{Res}}\right)^{k+1}\left(\frac{1}{1-h_{ii}}\right)$$

Remark.

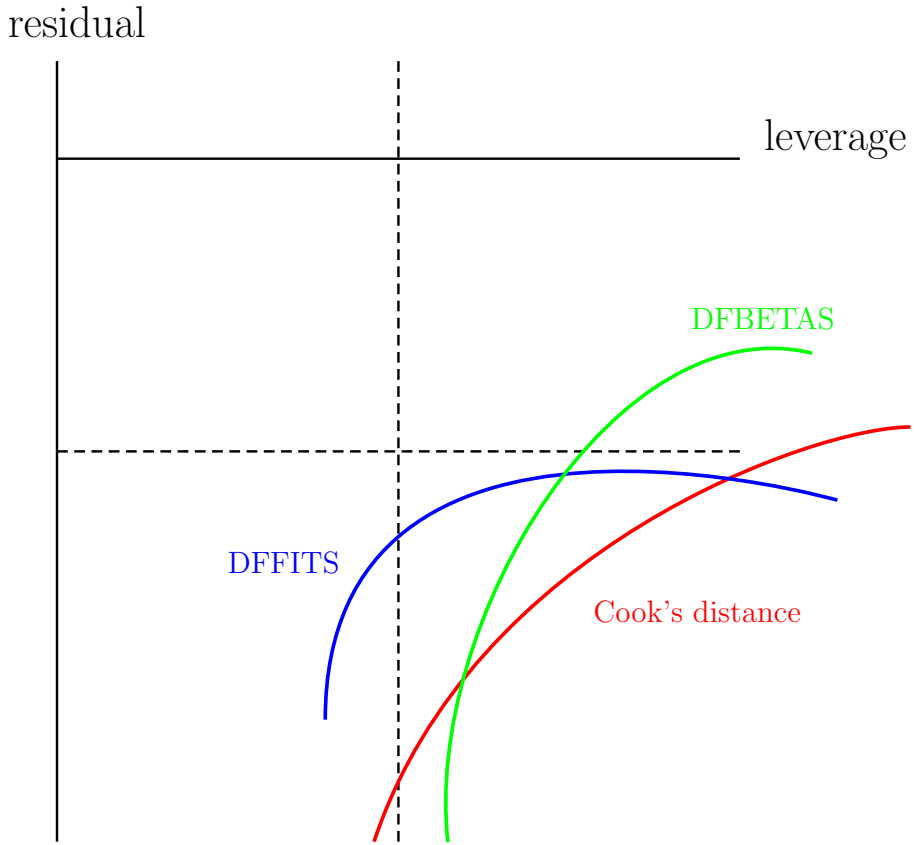
- A high leverage point will make COVRATIO_i large, thus improving the precision of estimation;
- If the ith observation is an outlier, $\frac{S^2_{(i)}}{MS_{Res}} < 1$ (so that COVRATIO_i is small). Thus, removing it improves precision.
- For large samples, the ith point should be considered influential if

$$|\text{COVRATIO}_i - 1| > 3(k+1)/n.$$


COVRATIO
covratio
0.6
0.8
1.0
1.2
0
10
20
30
40
50
index

A comparison of the different influence measures

- **Cook's distance**: overall influence on all fitted values (and regression coefficients)
- **DFBETAS**: influence on individual regression coefficients
- **DFFITS**: influence on individual fitted values
- **COVRATIO**: influence on precision of coefficients



Groups of influential observations

The discussion of which points constitute leverage or influential points can be extended to groups of two or more points in a similar way.

For example, Cook's distance measure can be extended to assess the **simultaneous influence** of a group of m points - simply leave out all m points simultaneously and re-calculate the regression parameters.

There can be situations in which several data points are jointly influential, while individual points are not.

Other authors suggest the use of cluster analysis to find groups of similar observations in a multivariate problem.

Treatment of influential observations

Once influential observations are identified, we need to analyze them carefully to see if they are valid observations or errors:

- Remove them if they are indeed errors
- Retain them if they are actually valid observations
- A compromise: Use robust estimation techniques, e.g., change the fitting error of a linear regression prolem to

$$\sum |e_i| \quad \longleftarrow \sum e_i^2$$

List of useful R functions

- rstudent()
- hatvalues()
- cooks.distance()
- dfbetas()
- dffits()
- covratio()
- influence.measures() ⟵ computes all the above

Summary

- **Concepts**: Leverage and influence
- **Influence measures** (of the observations on the model):
  - Cook's distance: overall influence on fitted values (and regression coefficients)
  - DFBETAS: influence on individual regression coefficients
  - DFFITS: influence on individual fitted values
  - COVRATIO: influence on precision of regression coefficients