## Influential Cases

- Could one or two data points throw off entire regression (leading to very different coefficients)?
- Would deleting a certain case result in . . .
    - Very different regression coefficients?
    - A different line of best fit?

## Leverage vs. Influence

- **Leverage:** amount of *potential* each data point has to change the regression
- **Influence:** amount by which each data point *actually* changes the regression

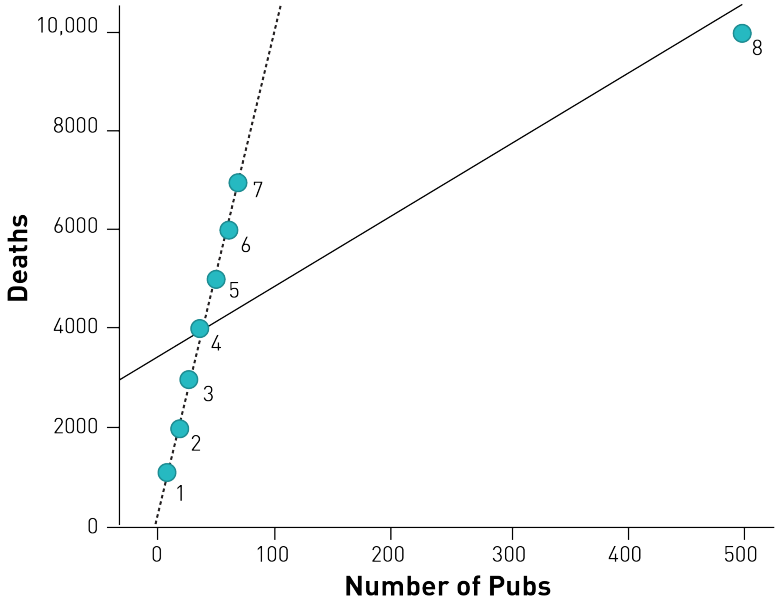
## Leverage

- A.k.a. "hat value"
- Written $h_{ii}$
- = a measure of how sensitive the regression is to a given data point *i*
    - Imagine moving a data point up and down on the Y-axis and observing how much the regression line follows it.
    - $h_{ii}$ = ratio of how much the regression line moves compared to how much you move the data point
    - Range = 0–1; 0 = no influence, 1 = complete influence
- A data point with high leverage does not necessarily influence the regression; i.e., that point could be in line with the rest of the data.

## Influence

- A data point needs a large residual to substantially change the regression coefficients.
- This means it lies far from the trend of the rest of the data.
- May indicate high influence
- **Cook's distance** measures the actual effect of deleting a given observation.
    - Each point with a large Cook's distance should be examined and potentially eliminated from the analysis.
    - General rule: Cook's distances >1 may be problematic (too much influence).

## Residuals vs. Influence




Example from Field et al. (2012). *Discovering Statistics Using R*

- Example: Eight different regions (boroughs) of London examined: number of deaths, number of pubs
  - The last case (Case 8) changes the line of best fit.
  - Interestingly, the residual for Case 8 is not large: the point is near the line.
  - Measures of influence (e.g., Cook's distance, hat value) indicate Case 8 has enormous influence.
  - Some cases can exert huge influence and still produce small residuals; thus it's important to look at both.

## Types of Residuals

- **Unstandardized** residuals
  - Measured in same units as outcome variable
  - Single model
  - Do not indicate which residual is "too large"
- **Standardized** residuals
  - Comparable across different regression models
  - Used to identify outliers

## Standardized Residuals: Rough Method

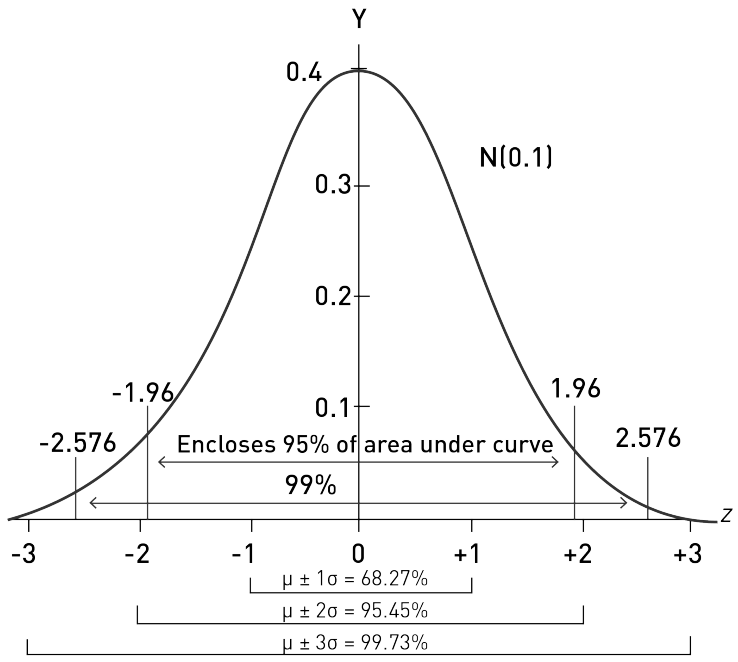
**Standardized Residuals**

$$r_i = e_i / (s\sqrt{1 - h_{ii}})$$

Residual — Standard error — Leverage (points with high leverage pull the regression with them, so they tend to have smaller residuals)

- Take the normal residuals and divide by their standard error.
- Results look a lot like *z* scores (distribution close to a normal curve).
  - But not exactly a normal curve
- We standardize residuals in order to compare them across different models.
- Also enables us to use a consistent rule for large residuals
  - Analogy: *z* scores at tail end of a *z* distribution, with critical values >2
  - Outliers are significant and merit attention.

## Dealing With Standardized Residuals




- Standardized residuals give us a rough estimate of what scores are outliers.
- Values >3 are very large, not likely due to chance.
- If 1% (or more) of cases have residuals >2.5, model contains too much error.
- If 5% (or more) of cases have residuals >2, model has too much error *and* represents our data poorly.

## Studentized Residuals: Precise Method

**Studentized Residuals**

$$t_i = e_i / (s_{-i} \sqrt{1 - h_i})$$

New standard error

- Studentized residual looks similar to standardized residual.
- Difference: Before computing standard error, we remove one data point.
- Doing this makes numerator and denominator independent.
- Studentized residual now follows a Student's *t* distribution.
  - This lets us apply precise tests to identify significant outliers.