Influential Observations in Linear Regression

Author: R. Dennis Cook

Sunil Dhaka

Indian Institute of Technology, Kanpur

sunild@iitk.ac.in

Linear and Non-Linear Models

May 3, 2022

Table of Contents

Definitions

- An **outlier** is a data point whose response y does not follow the general trend of the rest of the data.

Definitions

- An **outlier** is a data point whose response y does not follow the general trend of the rest of the data.
- A data point has **high leverage** if it has ”extreme” predictor x values.

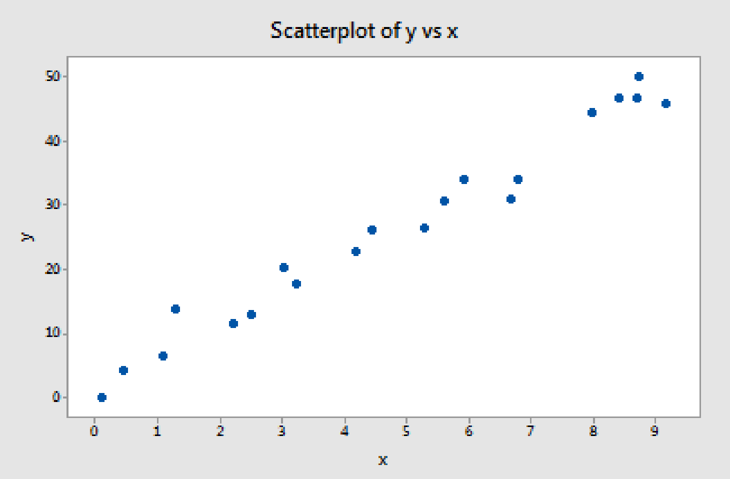
Definitions

- An **outlier** is a data point whose response y does not follow the general trend of the rest of the data.
- A data point has **high leverage** if it has ”extreme” predictor x values.
- A data point is **influential** if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.

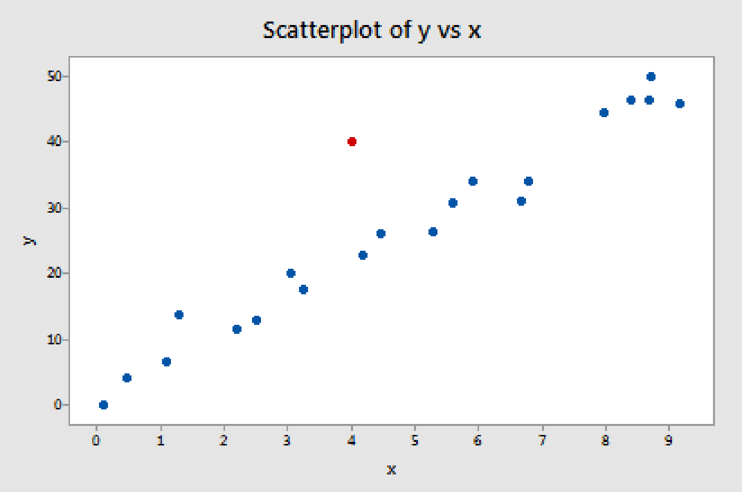
Definitions

- An **outlier** is a data point whose response y does not follow the general trend of the rest of the data.
- A data point has **high leverage** if it has ”extreme” predictor x values.
- A data point is **influential** if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.
  - Outliers and high leverage data points have the potential to be influential, but we generally have to investigate further to determine whether or not they are actually influential.
  - A data point is influential or not depends on the observed value of reponse and predictor of point

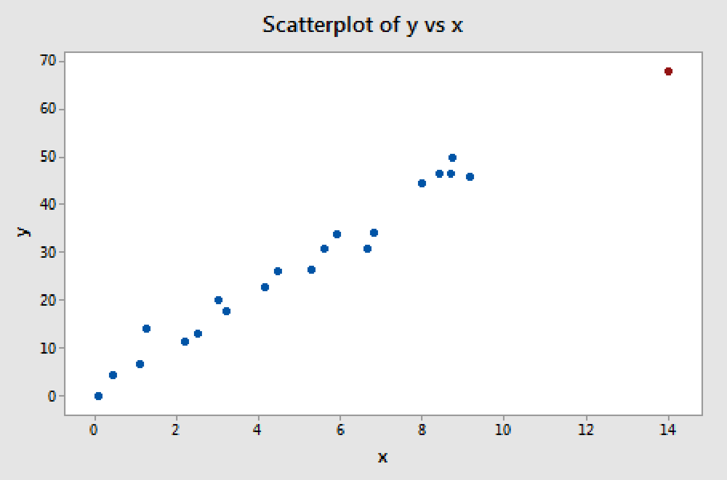
Example I



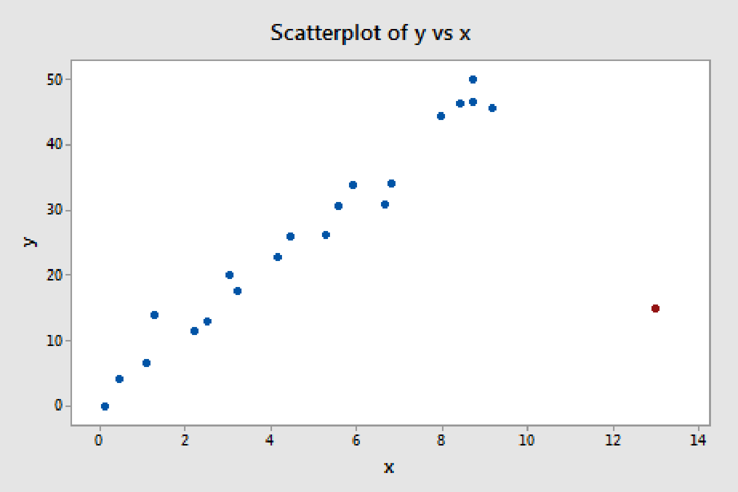
Example II



Example III



Example IV



Remarks

- the easy situation occurs for SLR, which can be visualized in 2D scatter plots

Remarks

- the easy situation occurs for SLR, which can be visualized in 2D scatter plots
- do not have that luxury in the case of MLR

Remarks

- the easy situation occurs for SLR, which can be visualized in 2D scatter plots
- do not have that luxury in the case of MLR
- we have to rely on various measures to help us determine whether a data point is an outlier, high leverage, or both

Remarks

- the easy situation occurs for SLR, which can be visualized in 2D scatter plots
- do not have that luxury in the case of MLR
- we have to rely on various measures to help us determine whether a data point is an outlier, high leverage, or both
- then need to see if the points are actually influential

Remarks

- the easy situation occurs for SLR, which can be visualized in 2D scatter plots
- do not have that luxury in the case of MLR
- we have to rely on various measures to help us determine whether a data point is an outlier, high leverage, or both
- then need to see if the points are actually influential
- after that have to decide whether to include or exclude such observations

Remarks

- the easy situation occurs for SLR, which can be visualized in 2D scatter plots
- do not have that luxury in the case of MLR
- we have to rely on various measures to help us determine whether a data point is an outlier, high leverage, or both
- then need to see if the points are actually influential
- after that have to decide whether to include or exclude such observations
    - must have a good, **objective** reason for deleting data points, then justify it with results

Remarks

- the easy situation occurs for SLR, which can be visualized in 2D scatter plots
- do not have that luxury in the case of MLR
- we have to rely on various measures to help us determine whether a data point is an outlier, high leverage, or both
- then need to see if the points are actually influential
- after that have to decide whether to include or exclude such observations
  - must have a good, **objective** reason for deleting data points, then justify it with results
  - common sense and knowledge about the situation matters

Setup and notations I

- consider standard full rank model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

- estimated coefficients

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- hat matrix

$$V = (v_{ij}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- predicted values: $\hat{\mathbf{Y}} = (y_i) = \mathbf{X}\hat{\beta} = \mathbf{VY}$

$$Var(\hat{\mathbf{Y}}) = \sigma^2\mathbf{V}$$

- residuals: $\mathbf{R} = (r_i) = (\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{I} - \mathbf{V})Y$

$$Var(\mathbf{R}) = \sigma^2(\mathbf{I} - \mathbf{V})$$

Detecting high leverage points I

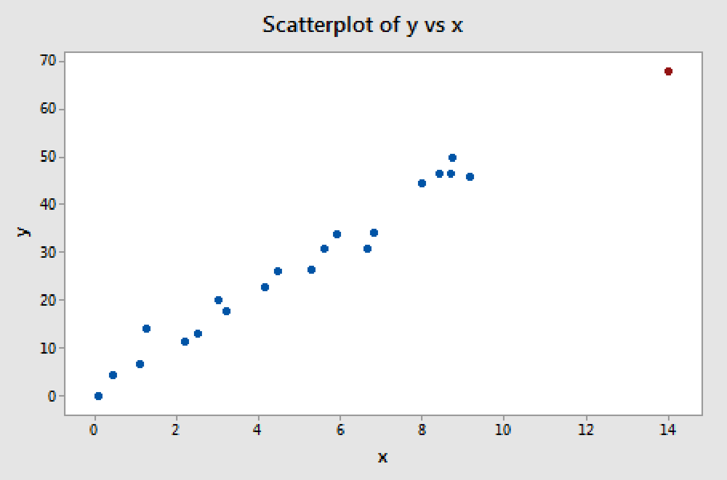
- why bother? in certain situations they may highly influence the estimated regression function, so need to identify
- leverage v_{ii}:
  - $\hat{y}_i = v_{i1}y_1 + \ldots + \mathbf{v_{ii}y_i} + \ldots + v_{in}y_n$
  - it is i^{th} row element of **VY** matrix
  - quantifies the influence that the observed response y_i has on its predicted value $\hat{y}_i$
- why they are called leverages?
  - v_{ii} quantifies how far away the i^{th} x value is from the rest of the x values
  - $0 \leq v_{ii} \leq 1$
  - $\sum_{i=1}^{n} v_{ii} = p$, reason?
- we select points with large leverage values as potential influential points

Detecting high leverage points II

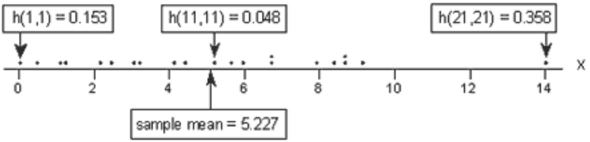
- what value should be considered large?
  - though the cut-off value depends on the situation and the analyst
  - common rule: more than 3 times larger than the mean leverage value

$$v_{ii} > 3\frac{\sum_{i=1}^{n} v_{ii}}{n} = 3\frac{p}{n}$$

High leverage point example I



High leverage point example II



- $n = 21$ and $p = 2$ (SLR); flag value $3\frac{p}{n} = 0.286$
- red point($x = 14, y = 68$) has leverage value $= 0.358$
- the data point **should** be flagged as high leverage point
- leverages only take into account the extremeness of the x values, but a high leverage observation may or may not actually be influential

Outlier detection I

- residuals can help in detecting outliers as measures the difference between the observed and predicted responses
- why need studentized residuals?
    - the major problem with ordinary residuals is that their magnitude depends on the units of measurement, thereby making it difficult to use the residuals as a way of detecting unusual y values

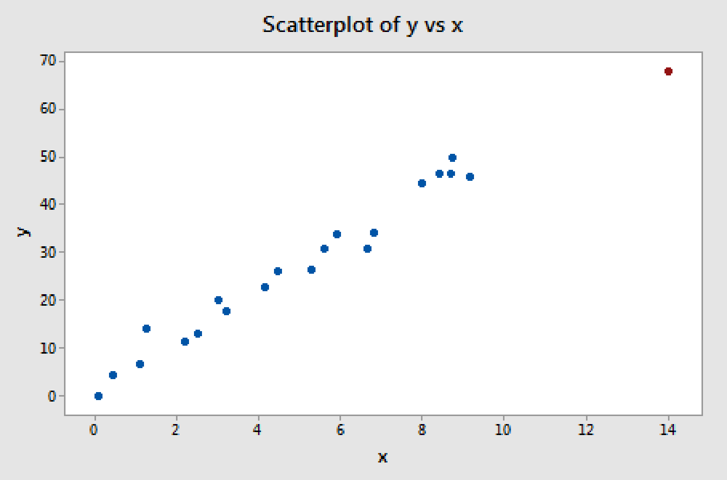
$$t_i = \frac{r_i}{sd(r_i)} = \frac{r_i}{\sqrt{MSE(1 - v_{ii})}}$$

    - depends on the leverage h_{ii}
- how to use it for outlier detection?
    - they quantify how large the residuals are in standard deviation units

Outlier detection II

- studentized residual that is larger than 3 (in absolute value) is generally decided as outlier
- again depends on the situation and the analyst

Outlier detection III



- std residual for red point $= 3.68 > 3$

Outlier detection IV

- deemed as an outlier, further investigation to decide whether it is an influential point or not

Intermission

- When trying to identify outliers, one problem that can arise is when there is a potential outlier that influences the regression model to such an extent that the estimated regression function is "pulled" towards the potential outlier, so that it isn't flagged as an outlier using the standardized residual criterion.
- To address this issue, deleted residuals offer an alternative criterion for identifying outliers.
- The basic idea is to delete the observations one at a time, each time refitting the regression model on the remaining $n-1$ observations. Then, we compare the observed response values to their fitted values based on the models with the ith observation deleted.

Cooks distance measure

- the influence of the i_{th} data point be judged by using the distance measure

$$D_i = \frac{(\hat{\beta} - \hat{\beta_{(i)}})'(X'X)(\hat{\beta} - \hat{\beta_{(i)}})}{ps^2} \tag{1}$$

  here, $s^2 = MSE = \frac{R'R}{(n-p)}$

- large value of D_i indicates that the associated i^{th} point has a strong influence on the estimate of regression parameters β
- magnitude of the distance between $\hat{\beta}$ and $\hat{\beta_{(i)}}$
  - compare D_i value to probability points of $F_{p,n-p,ncp=0}$
  - equivalent to finding level of the confidence ellipsoid

Alternative form

- using $\hat{\beta} - \hat{\beta_{(i)}} = \frac{(X'X)^{-1}x_i r_i}{1-v_{ii}}$ result, we get

$$D_i = \frac{t_i^2 w_i}{p}$$

- D_i can be large if either t_i^2 or w_i is large
- $t_i = \frac{r_i}{s\sqrt{(1-v_{ii})}}$, is i_{th} deleted studentized residual
- it depends on the residual, measures outlier properties of the observation
- $w_i = \frac{v_{ii}}{(1-v_{ii})}$, ratio of the variance of the i_{th} predicted value and the variance of the i_{th} residual
- it also depends on the leverage of the observation, measures the location of the i_{th} observation
- Cook's distance incorporates both outlier(y value) and high leverage(x value) properties of an observation

Using Cook's distance measures

- for MLR models we need to rely on guidelines/rules for deciding when a Cook's distance measure is large enough to deem a data point as influential observation
- common rule:
  - $D_i > 0.5$ then may be influential, but needs further investigation
  - $D_i > 1$, quite likely to be influential
  - $D_i >> 1$, almost certainly influential

Examples I

R scripts

Residual Correlation I

Goal:

to find out $t_{j(i)}$ that is $\mathbf{j_{th}}$(not i_{th}) deleted studentized residual based on the data set with i_{th} point removed,

- consider

$$v_{kl} = x_k'(X_{(i)}'X_{(i)})^{-1}x_l' \tag{2}$$

here, $X_{(i)}'X_{(i)} = X'X - x_i x_i'$ and x_i is i_{th} row

Residual Correlation II

- using general identity

$$\left[B + uz^{'}\right]^{-1} = B^{-}1 - \frac{B^{-}1uz^{'}B^{-}1}{1 + u^{'}B^{-1}z}$$

after using this on eq(1) expressions and a bit algebra,

$$v_{kl} = v_{kl(i)} - \frac{v_{ki(i)}v_{li(i)}}{1 + v_{ii}} \tag{3}$$

$$v_{kl(i)} = v_{kl} + \frac{v_{ki}v_{li}}{1 - v_{ii}} \tag{4}$$

Residual Correlation III

- other two results that are used to get $t_{j(i)}$

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{(X'X)^{-1} x_i r_i}{1 - v_{ii}} \tag{5}$$

$$(n-p)s^2 = (n-p-1)s^2_{(i)} + \frac{r_i^2}{1 - v_{ii}} \tag{6}$$

- ρ_{ij}: residual correlation in **full** dataset
- ratio $w_{j(i)}$, when $j \neq i$

$$\frac{v_{jj(i)}}{1 - v_{jj(i)}} = \frac{v_{jj}(1 - \rho_{ij}^2) + \rho_{ij}^2}{(1 - v_{jj}(1 - \rho_{ij}^2))} \tag{7}$$

  - ratio will be large if either v_{jj} or ρ_{ij}^2 is large
  - high v_{jj} would have been detected in the full data analysis

Residual Correlation IV

- now if the ratio is high, it must be due to large correlation between i_{th} and j_{th} residuals in full dataset
- and if ρ_{ij} is negligible then $w_{j(i)} = w_j$

- using results from eq(2)-(5), and with lot more algebra we get

$$t_{j(i)}^2 = \frac{(n-p-1)(t_j - \rho_{ij} t_i)^2}{(n-p-t_i^2)(1-\rho_{ij}^2)} \tag{8}$$

- if residual correlation ρ_{ij} is negligible, and $t_i > 1$, then for all remaining deleted studentized residuals will increase
- if ρ_{ij} is large, then j_{th} deleted studentized residuals increases substantially then i_{th} point is deleted

- note that both expressions of $w_{j(i)}$ and $t_{j(i)}^2$ are in terms of full dataset
- their product gives Cook's distance measure

Paper Example

- available in R as **stack**, multipler outlier example
- Number of Observations: 21
- Number of Variables: 4
- Variables:
  - y: STACKLOSS, x: AIRFLOW, WATERTEMP, ACIDCONC
- model setup: from Daniel and Wood (1971)
  - found 4 outliers: (1,3,4,21)
  - model containes a linear and quadratic term of AIRFLOW
  - a linear term for WATERTEMP
  - ACIDCONC is not needed for prediction
- R script

References

R. Dennis Cook (1979)
Influential Observations in Linear Regression
Journal of the American Statistical Association 74(365), 169 – 174.

The End

Questions? Comments?