Chapter 13

Influential Points

Chanwoo Yoo, Division of Advanced Engineering,
Korea National Open University

Contents

1. Outliers & High Leverage Points

1. Outlier

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

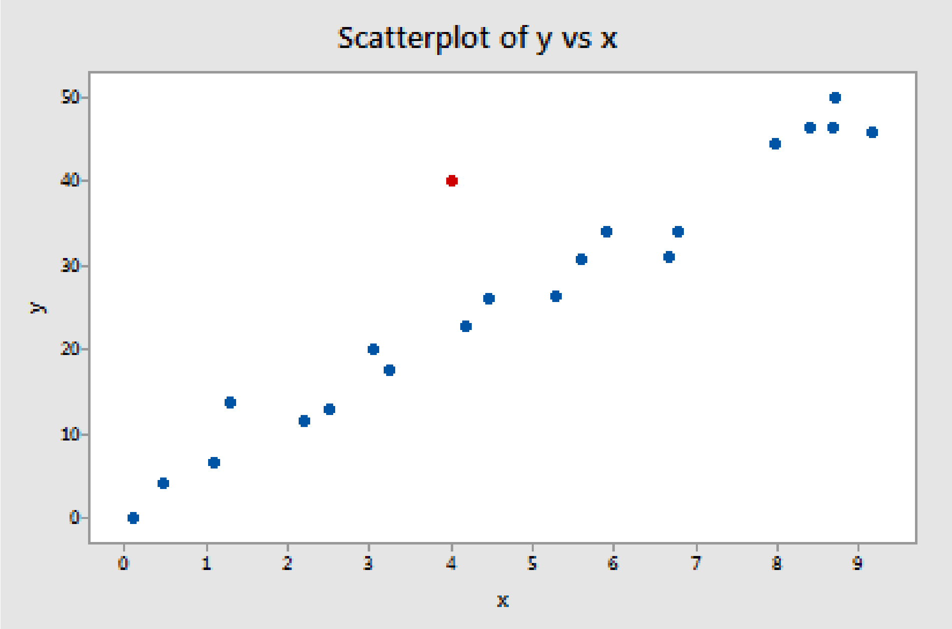
2. High Leverage Point

- A data point has **high leverage** if it has "extreme" predictor x values.
- With a single predictor, an extreme x value is simply one that is particularly high or low.
- With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values.

3. Influential Data Point

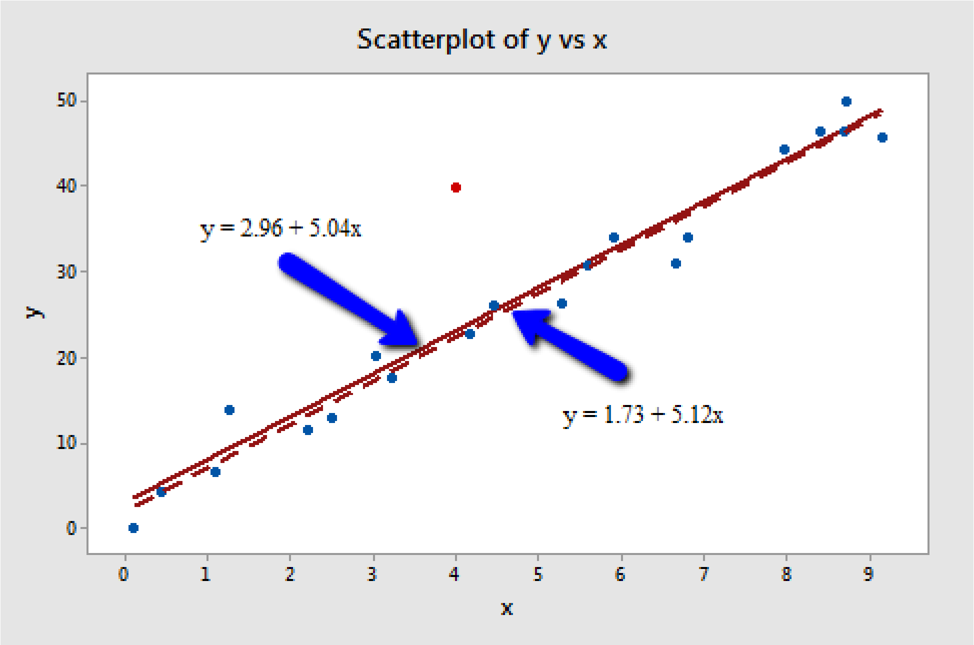
- A data point is influential if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.
- Outliers and high leverage data points have the potential to be influential, but we generally have to investigate further to determine whether or not they are actually influential.

4. Example: Outlier



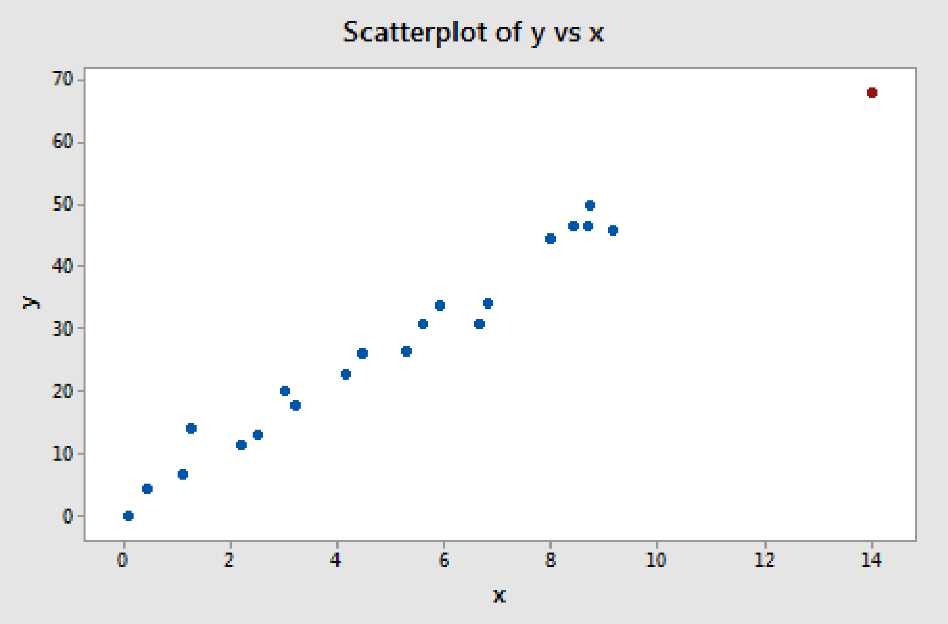
- Influence2 data set
- Because the red data point does not follow the general trend of the rest of the data, it would be considered an outlier.

4. Example: Outlier



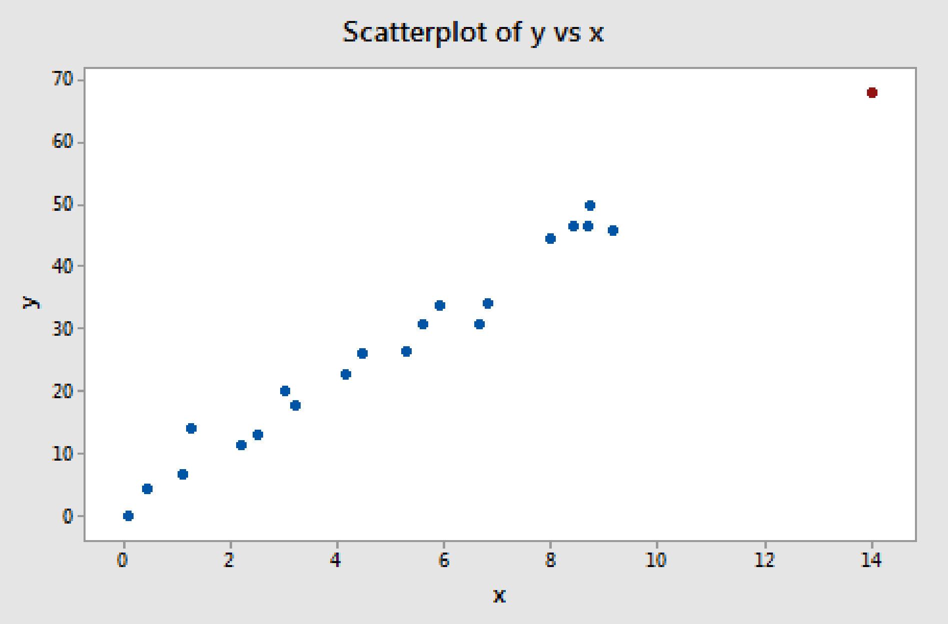
- The plot illustrates two best fitting lines — one obtained when the red data point is included and one obtained when the red data point is excluded.
- The data point is not deemed influential.

5. Example: High Leverage



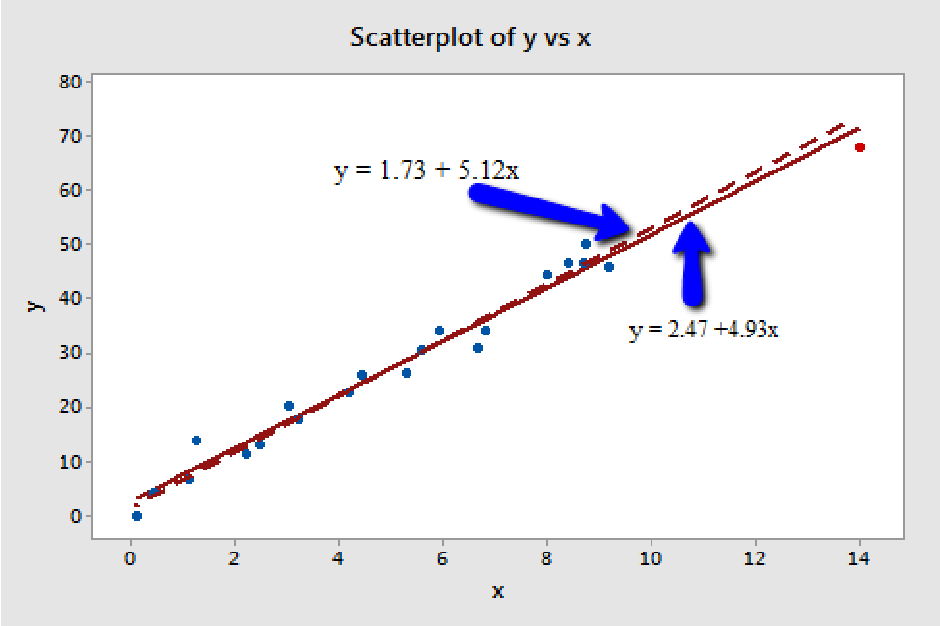
- Influence3 data set
- The red data point does follow the general trend of the rest of the data. Therefore, it is not deemed an outlier here.

5. Example: High Leverage



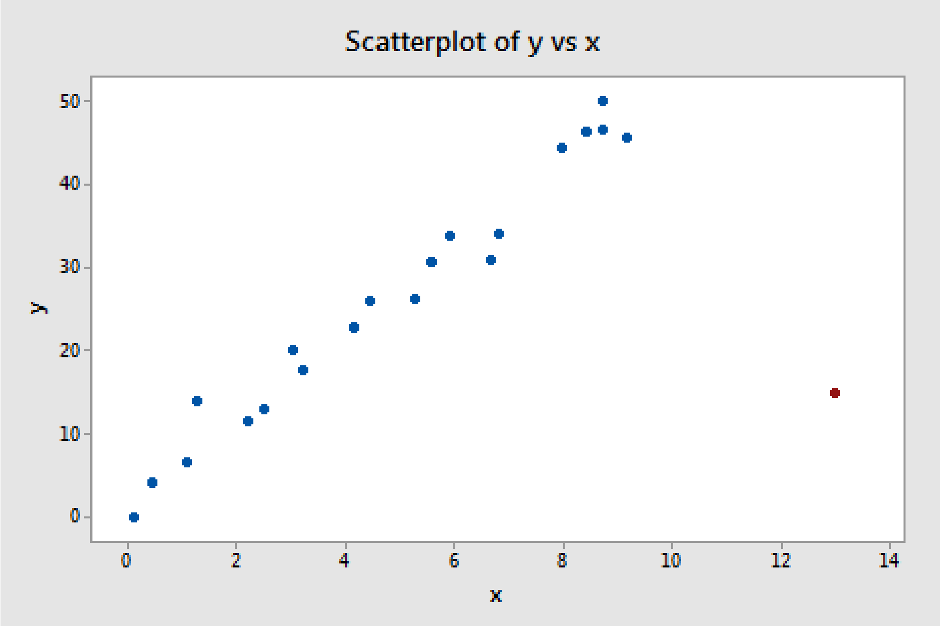
- However, this point does have an extreme x value, so it does have high leverage.

5. Example: High Leverage



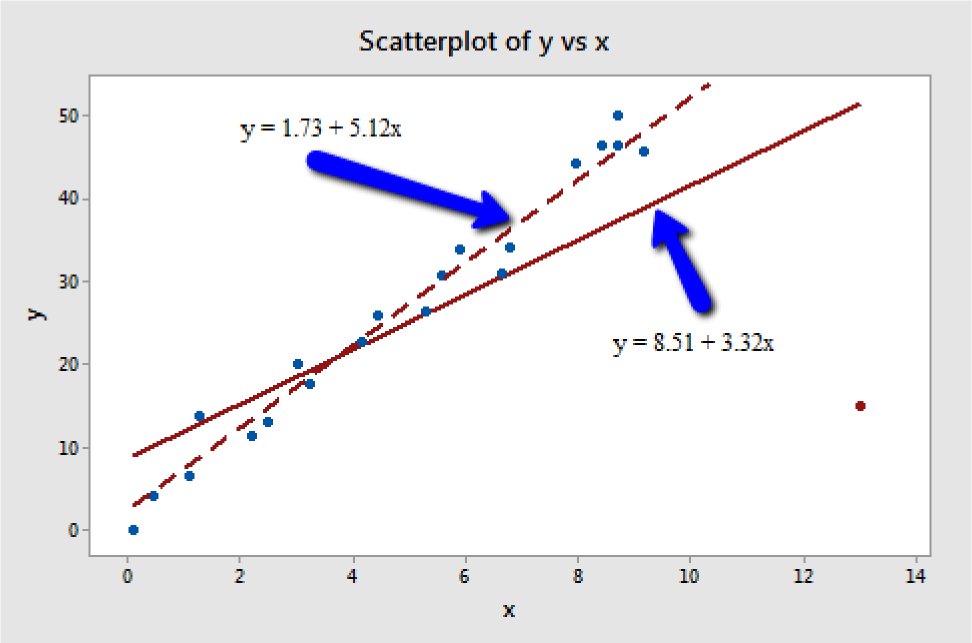
- The plot illustrates two best fitting lines — one obtained when the red data point is included and one obtained when the red data point is excluded.
- The data point is not deemed influential.

6. Example: Influential Data Point



- Influence 4 data set
- The red data point is most certainly an outlier and has high leverage.

6. Example: Influential Data Point



- The two best fitting lines are substantially different.
- The red data point is deemed both high leverage and an outlier, and it turned out to be influential too.

2. Identifying High Leverage Points

1. Hat Matrix

- $\hat{\mathbf{y}} = X(X^T X)^{-1} X^T \mathbf{y}$
  - Hat Matrix H: $X(X^T X)^{-1} X^T$
- $\hat{\mathbf{y}} = H\mathbf{y}$

2. Leverage

- $$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} h_{11}y_1 + h_{12}y_2 + \cdots + h_{1n}y_n \\ h_{21}y_1 + h_{22}y_2 + \cdots + h_{2n}y_n \\ \vdots \\ h_{n1}y_1 + h_{n2}y_2 + \cdots + h_{nn}y_n \end{bmatrix}$$
- The leverage, h_{ii}, quantifies the influence that the observed response y_i has on its predicted value $\hat{y}_i$.

2. Leverage

- If h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response $\hat{y}_i$. On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response $\hat{y}_i$. It's for this reason that the h_{ii} are called the "leverages."

3. Properties of Leverage

- The leverage h_{ii} is a measure of the distance between the x value for the ith data point and the mean of the x values for all n data points.
- The leverage h_{ii} is a number between 0 and 1, inclusive.
- The sum of the h_{ii} equals p, the number of parameters (regression coefficients including the intercept).

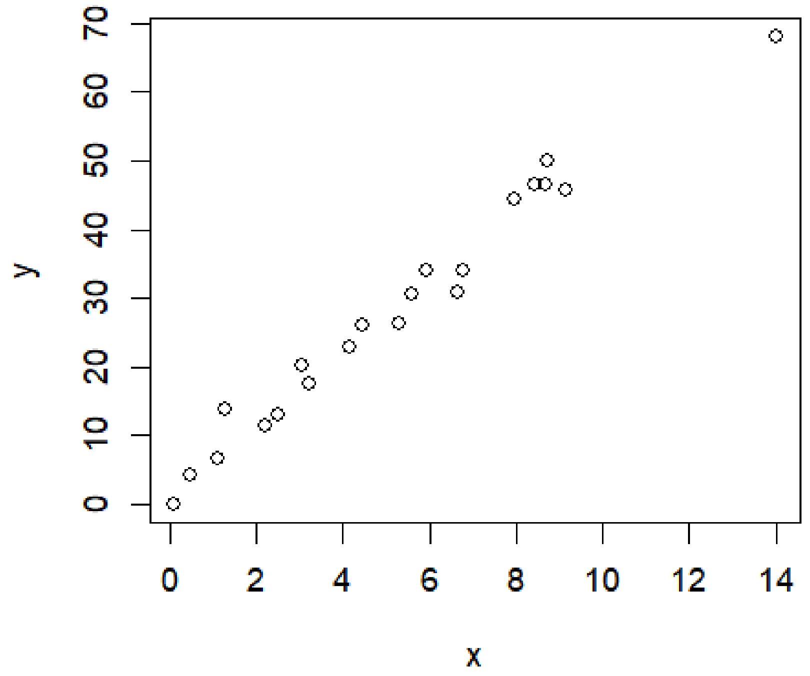
4. Guideline

- Leverages can help us identify x values that are extreme and potentially influential on regression analysis.
- A common rule is to flag any observation whose leverage value, h_{ii}, is more than 3 times larger than the mean leverage value:

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$$

5. Leverage: influence3

```
influence3 <- read.table("influence3.txt", header=T)
attach(influence3)

plot(x, y)

model.1 <- lm(y ~ x)
lev <- hatvalues(model.1)
round(lev, 6)
sum(lev)

detach(influence3)
```



5. Leverage: influence3

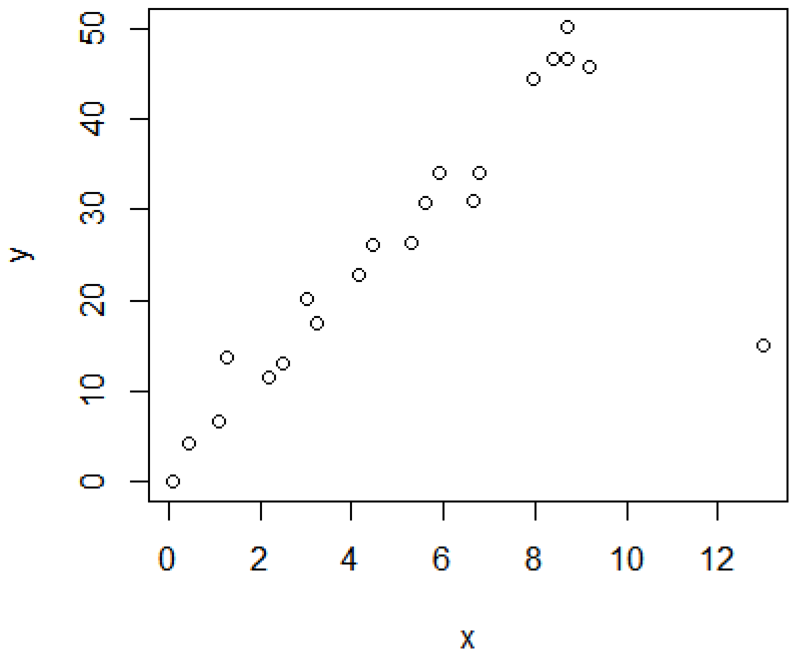
```
> round(lev, 6)
       1        2        3        4        5        6 
0.153481 0.139367 0.116292 0.110382 0.084374 0.077557 
       7        8        9       10       11       12 
0.066879 0.063589 0.050033 0.052121 0.047632 0.048156 
      13       14       15       16       17       18 
0.049557 0.055893 0.057574 0.078121 0.088549 0.096634 
      19       20       21 
0.096227 0.110048 0.357535 
> sum(lev)
[1] 2
```

5. Leverage: influence3

- $n = 21, p = 2$
- $3 \times \frac{p}{n} = 3 \times \frac{2}{21} = 0.286 < 0.357$

6. Leverage: influence4

```
influence4 <- read.table("influence4.txt", header=T)
attach(influence4)

plot(x, y)

model.2 <- lm(y ~ x)
lev <- hatvalues(model.2)
round(lev, 6)

detach(influence3)
```



6. Leverage: influence4

```
> round(lev, 6)
       1        2        3        4        5        6 
0.158964 0.143985 0.119522 0.113263 0.085774 0.078589 
       7        8        9       10       11       12 
0.067369 0.063924 0.049897 0.052019 0.047667 0.048354 
      13       14       15       16       17       18 
0.049990 0.057084 0.058943 0.081446 0.092800 0.101587 
      19       20       21 
0.101146 0.116146 0.311532 
```

6. Leverage: influence4

- $n = 21, p = 2$
- $3 \times \frac{p}{n} = 3 \times \frac{2}{21} = 0.286 < 0.311$

7. Summary

- The leverage merely quantifies the potential for a data point to exert strong influence on the regression analysis.
- The leverage depends only on the predictor values.
- Whether the data point is influential or not also depends on the observed value of the reponse y_i.

3. Identifying Outliers

1. Residuals

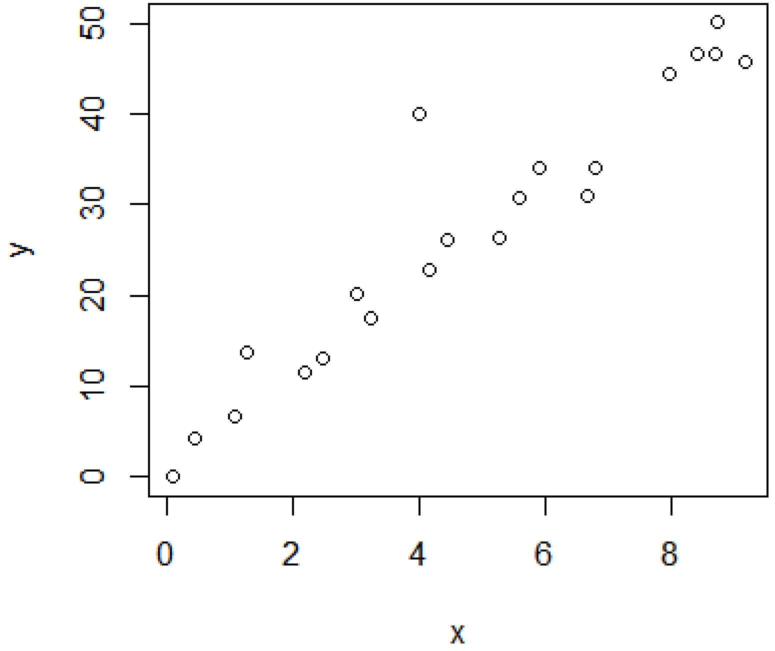
- The problem with ordinary residuals is that their magnitude depends on the units of measurement, thereby making it difficult to use the residuals as a way of detecting unusual y values.
- We can eliminate the units of measurement by dividing the residuals by an estimate of their standard deviation, thereby obtaining what are known as studentized residuals (or internally studentized residuals)

2. Studentized Residuals

- Studentized Residuals (or Internally Studentized Residuals)
  - Ordinary residual divided by an estimate of its standard deviation
  - $r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$
    - $e_i = y_i - \hat{y}_i$

3. Studentized Residuals: influence2

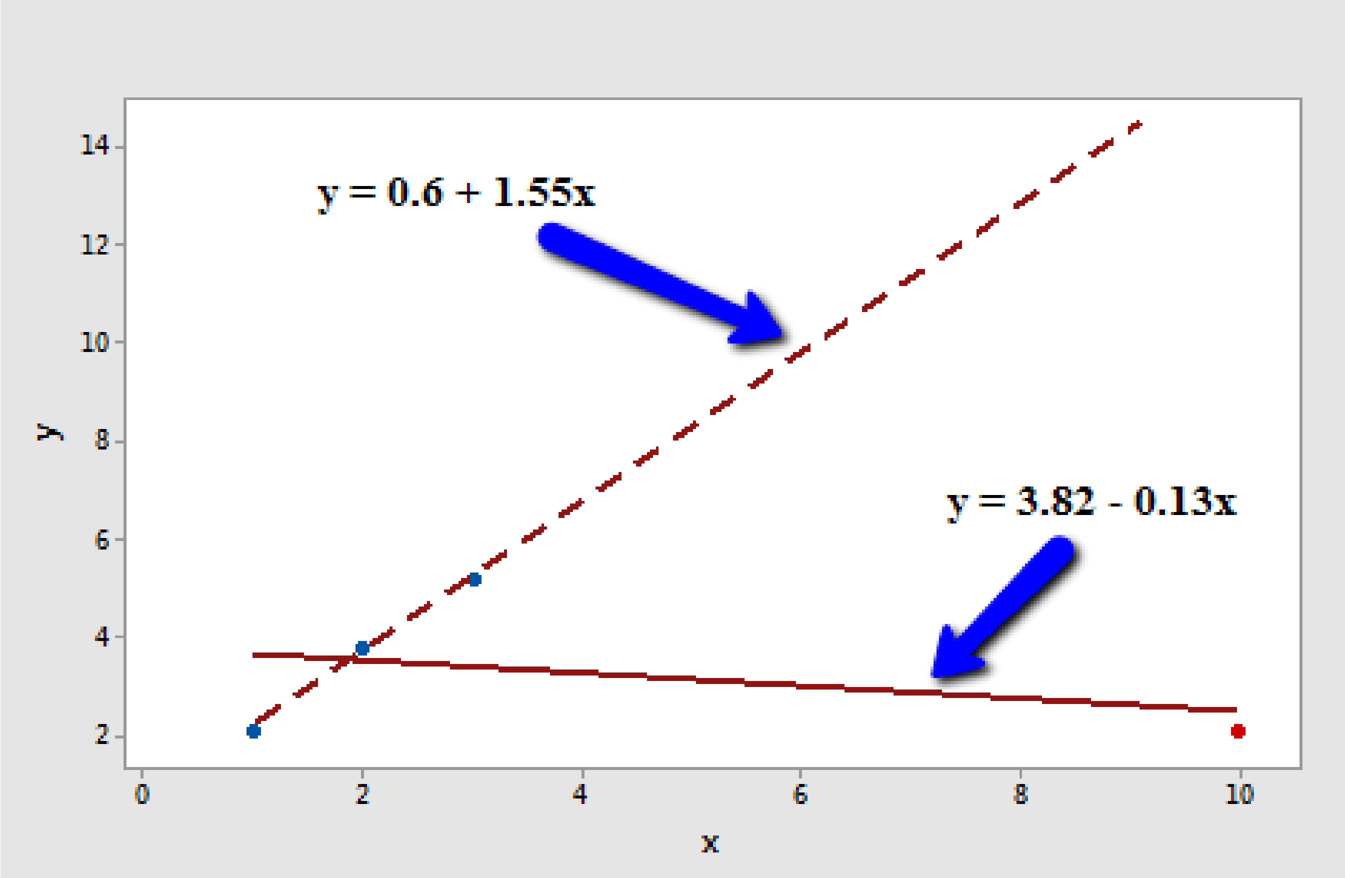
```
influence2 <- read.table("influence2.txt", header=T)
attach(influence2)

plot(x, y)

model.1 <- lm(y ~ x)

sta <- rstandard(model.1)
round(sta, 6)

detach(influence2)
```



3. Studentized Residuals: influence2

```
> round(sta, 6)
        1         2         3         4         5         6 
-0.826351 -0.249154 -0.435445  0.998187 -0.581904 -0.574462 
        7         8         9        10        11        12 
 0.413791 -0.371226  0.139767 -0.262514 -0.713173 -0.095897 
       13        14        15        16        17        18 
 0.252734 -1.229353 -0.683161  0.292644  0.262144  0.731458 
       19        20        21 
-0.055615 -0.776800  3.681098 
```

4. Deleted Residuals



4. Deleted Residuals

- $d_i = y_i - \hat{y}_{(i)}$
  - y_i: observed response for the ith observation
  - $\hat{y}_{(i)}$: predicted response for the ith observation based on the estimated model with the ith observation deleted

5. Externally Studentized Residuals

- Externally Studentized Residuals (or Studentized Deleted Residuals)

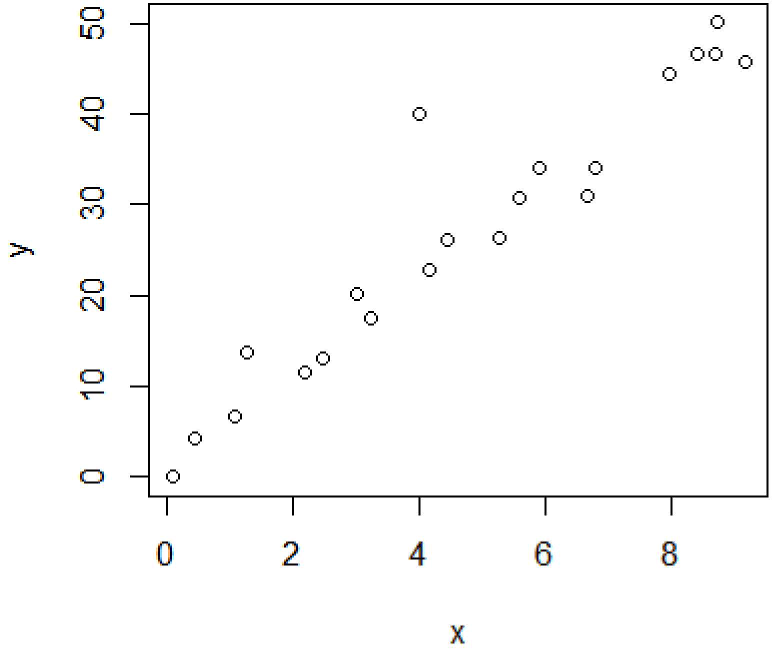
  - $t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(i)}\,(1-h_{ii})}}$
  - Deleted residual divided by its estimated standard deviation
  - Ordinary residual divided by a factor that includes the mean square error based on the estimated model with the ith observation deleted, $MSE_{(i)}$, and the leverage, h_{ii}

5. Externally Studentized Residuals

- If an observation has an externally studentized residual that is larger than 3 (in absolute value) we can call it an outlier.

6. Externally Studentized Residuals: influence2

```
influence2 <- read.table("influence2.txt", header=T)
attach(influence2)

plot(x, y)

model.1 <- lm(y ~ x)

stu <- rstudent(model.1)
round(stu, 6)

detach(influence2)
```



6. Externally Studentized Residuals: influence2

```
> round(stu, 6)
        1         2         3         4         5         6 
-0.819167 -0.242905 -0.425962  0.998087 -0.571499 -0.564060 
        7         8         9        10        11        12 
 0.404582 -0.362643  0.136110 -0.255977 -0.703633 -0.093362 
       13        14        15        16        17        18 
 0.246408 -1.247195 -0.673261  0.285483  0.255615  0.722190 
       19        20        21 
-0.054136 -0.768382  6.690129 
```

4. Identifying Influential Data Points

1. DFFITS (Difference in Fits)

- $DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSE_{(i)}\, h_{ii}}}$
  - The numerator measures the difference in the predicted responses obtained when the ith data point is included and excluded from the analysis.
  - The denominator is the estimated standard deviation of the difference in the predicted responses.

1. DFFITS

- $DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSE_{(i)}\, h_{ii}}}$
  - The difference in fits quantifies the number of standard deviations that the fitted value changes when the ith data point is omitted.

1. DFFITS

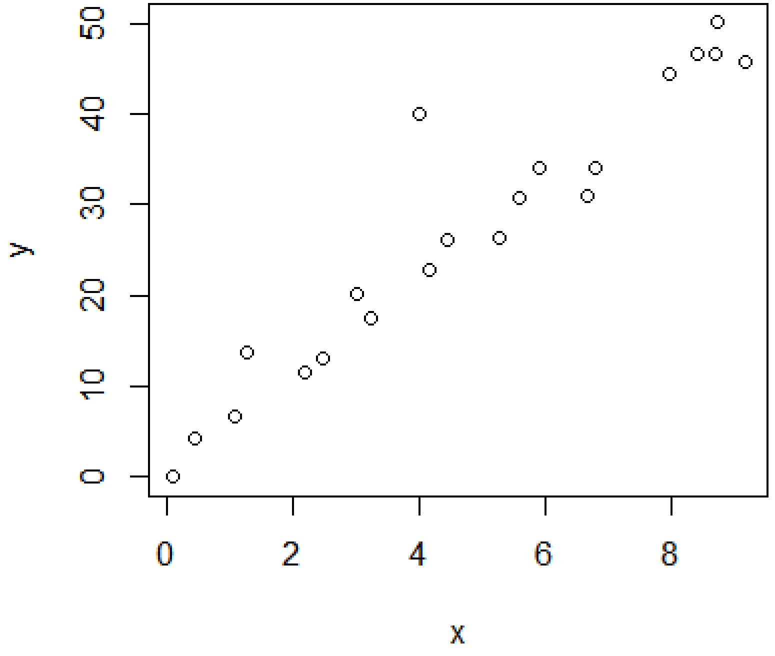
- An observation is deemed influential if the absolute value of its DFFITS value is greater than:

$$2\sqrt{\frac{p+1}{n-p-1}}$$

  - $n: number\ of\ observations, p: number\ of\ parameters$
  - This is not a hard-and-fast rule, but rather a guideline.

2. DFFITS: influence2

```
influence2 <- read.table("influence2.txt", header=T)
attach(influence2)

plot(x, y)

model.1 <- lm(y ~ x)

dffit <- dffits(model.1)
round(dffit, 6)

detach(influence2)
```



2. DFFITS: influence2

```
> round(dffit, 6)
        1         2         3         4         5         6 
-0.378974 -0.105007 -0.162478  0.367368 -0.175466 -0.163769 
        7         8         9        10        11        12 
 0.106698 -0.092652  0.030612 -0.058495 -0.160254 -0.021828 
       13        14        15        16        17        18 
 0.059879 -0.340354 -0.188345  0.100168  0.097710  0.292757 
       19        20        21 
-0.021884 -0.339696  1.550500 
```

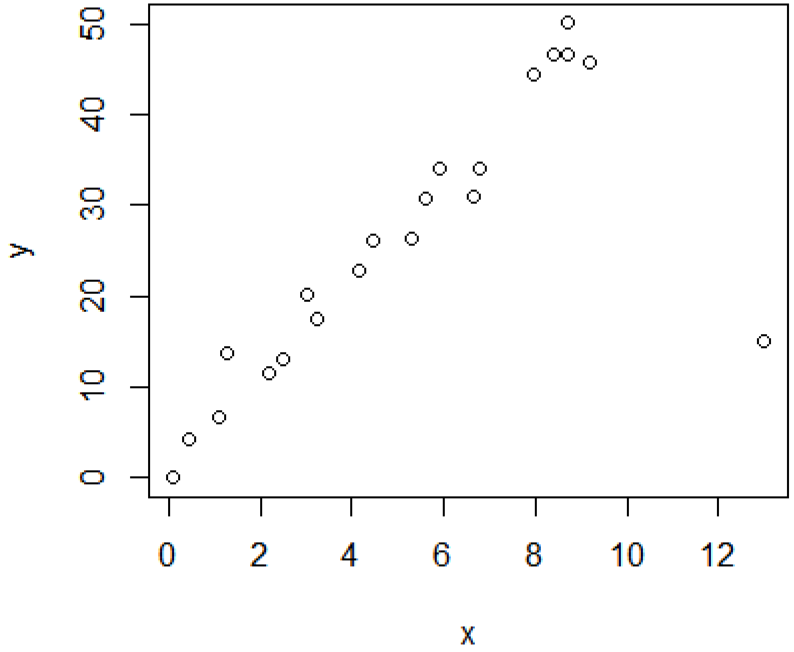
2. DFFITS: influence2

- $n = 21, p = 2$

- $2\sqrt{\frac{p+1}{n-p-1}} = 2\sqrt{\frac{2+1}{21-2-1}} = 0.82 < |1.55|$

3. DFFITS: influence4

```
influence4 <- read.table("influence4.txt", header=T)
attach(influence4)

plot(x, y)

model.2 <- lm(y ~ x)

dffit <- dffits(model.2)
round(dffit, 6)

detach(influence4)
```



3. DFFITS: influence4

```
> round(dffit, 6)
        1         2         3         4         5 
-0.402761 -0.243756 -0.205848  0.037612 -0.131355 
        6         7         8         9        10 
-0.109593  0.040473 -0.042401  0.060224  0.009181 
       11        12        13        14        15 
 0.005430  0.078165  0.127828  0.007230  0.073067 
       16        17        18        19        20 
 0.280501  0.323599  0.436114  0.308869  0.249206 
       21 
-11.467011 
```

3. DFFITS: influence4

- $n = 21, p = 2$

- $2\sqrt{\frac{p+1}{n-p-1}} = 2\sqrt{\frac{2+1}{21-2-1}} = 0.82 < |-11.467|$

4. Cook's Distance

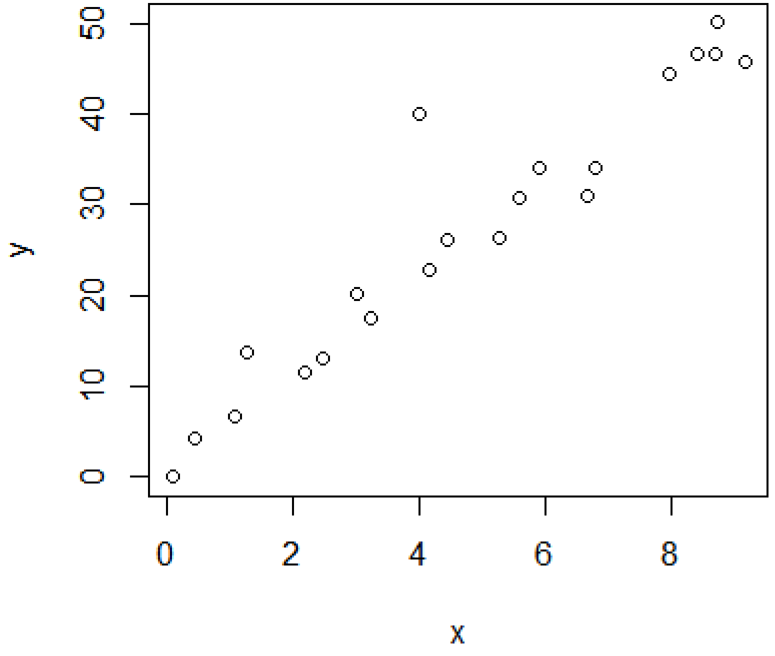
- $D_i = \frac{(y_i - \hat{y}_i)^2}{p \times MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right)$
  - Cook's distance depends on the residual, $y_i - \hat{y}_i$, and the leverage, h_{ii}.

5. Guideline

- If D_i is greater than 0.5, then the ith data point is worthy of further investigation as it may be influential.
- If D_i is greater than 1, then the ith data point is quite likely to be influential.
- Or, if D_i sticks out like a sore thumb from the other D_i values, it is almost certainly influential.

6. Cook's Distance: influence2

```
influence2 <- read.table("influence2.txt", header=T)
attach(influence2)

plot(x, y)

model.1 <- lm(y ~ x)

cook <- cooks.distance(model.1)
round(cook, 6)

detach(influence2)
```

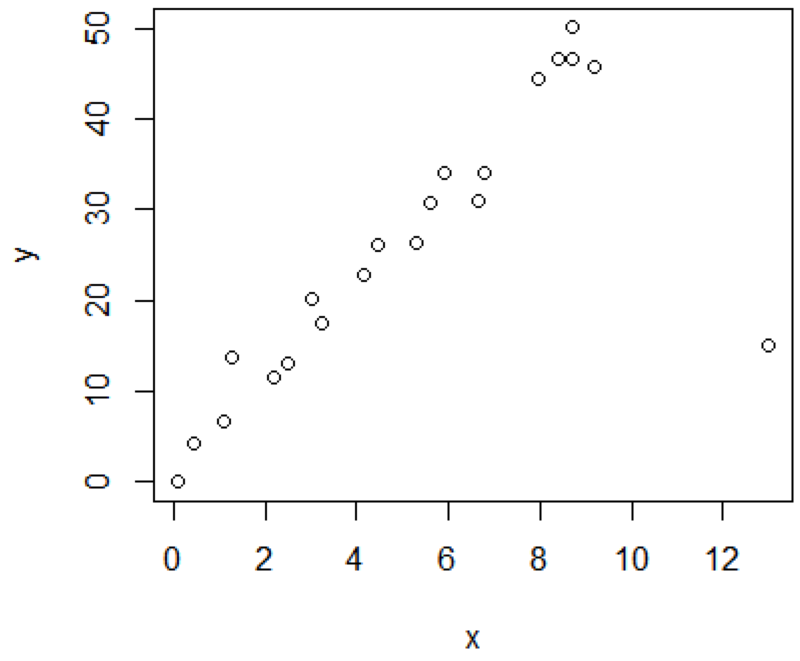


6. Cook's Distance: influence2

```
> round(cook, 6)
       1        2        3        4        5        6        7 
0.073076 0.005800 0.013794 0.067493 0.015960 0.013909 0.005954 
       8        9       10       11       12       13       14 
0.004498 0.000494 0.001799 0.013191 0.000251 0.001886 0.056275 
      15       16       17       18       19       20       21 
0.018262 0.005272 0.005021 0.043960 0.000253 0.058968 0.363914 
```

7. Cook's Distance: influence4

```
influence4 <- read.table("influence4.txt", header=T)
attach(influence4)

plot(x, y)

model.2 <- lm(y ~ x)

cook <- cooks.distance(model.2)
round(cook, 6)

detach(influence4)
```



7. Cook's Distance: influence4

```
> round(cook, 6)
       1        2        3        4        5        6        7 
0.081718 0.030755 0.021983 0.000746 0.009014 0.006290 0.000863 
       8        9       10       11       12       13       14 
0.000947 0.001907 0.000044 0.000016 0.003203 0.008478 0.000028 
      15       16       17       18       19       20       21 
0.002804 0.039575 0.052293 0.091802 0.048085 0.031938 4.048013 
```

5. Dealing with Problematic Data Points

1. Dealing with Problematic Data Points

- Check for obvious data errors:
  - If the error is just a data entry or data collection error, correct it.
  - If the data point is not representative of the intended study population, delete it.
  - If the data point is a procedural error and invalidates the measurement, delete it.

1. Dealing with Problematic Data Points

- Consider the possibility that you might have just misformulated your regression model:
  - Did you leave out any important predictors?
  - Should you consider adding some interaction terms?
  - Is there any nonlinearity that needs to be modeled?

1. Dealing with Problematic Data Points

- Do not delete data points just because they do not fit your preconceived regression model.
- If you delete any data after you've collected it, justify and describe it in your reports.
- If you are not sure what to do about a data point, analyze the data twice — once with and once without the data point — and report the results of both analyses.

Next

Chapter 14
Multicollinearity