

Outliers

WAIT....WHAT EXACTLY IS AN OUTLIER?

Unusual data are problematic in linear models fit by least squares because they can influence the results of the analysis. But what is an outlier? What are the criteria used to determine if a case is an outlier or not?

A univariate outlier is a value of Y or X that is unconditionally unusual; such value may or may not be a regression outlier. In a regression context, a regression outlier is an observation whose dependent variable value is conditionally unusual given the value of the independent variable(s).

Conceptually, we can think of two distinct concepts related to outliers--- 'influence' and 'leverage'. Leverage assesses how far away a value of the independent variable is from the mean value: the further away the observation, the more leverage it has. Observations whose inclusion or exclusion result in substantial changes in the fitted model (i.e. coefficients, fitted values) are said to be influential.

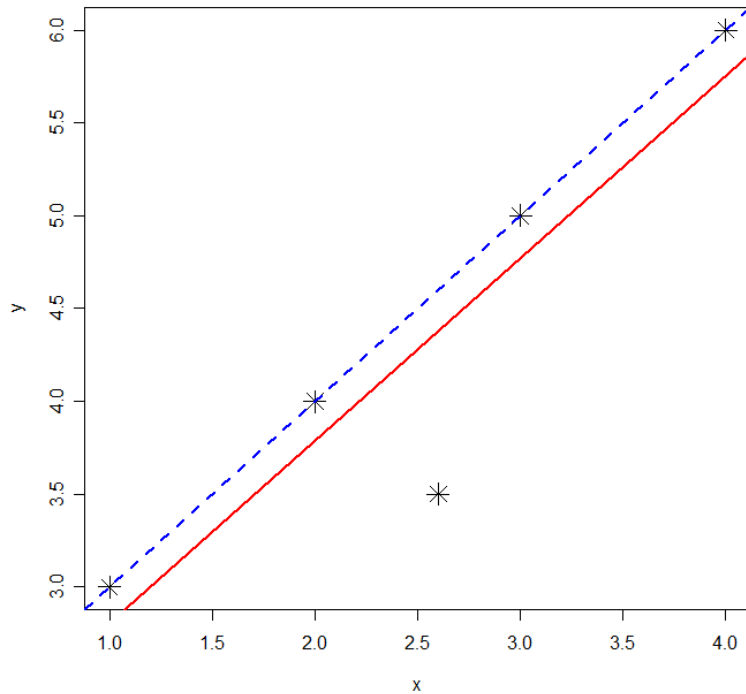
Outliers with respect to the predictors (a.k.a. independent variables) are called 'leverage points'. They can affect the regression model, though their corresponding response (a.k.a. dependent) variables need NOT be outliers. These leverage point may or may not substantively change the regression model.

To elaborate, let's consider the following three cases:

Case 1:

An outlier can have x-value near the center of the x distribution. As a result, deleting this outlier has minimal impact on the least squares fit, leaving the slope (b) largely unchanged and only affecting the intercept slightly. This case only exerts a low leverage and low influence on the regression fit.

The red line shows the regression fit using ALL data points. The blue dotted line shows the regression fit after deleting the extreme case. Note the two lines are rather similar.



Regression fit using all data, intercept=1.82, slope=.9824.

Call:

```
lm(formula = y ~ x, data = data1)
```

Residuals:

```
1 2 3 4 5
0.1933 -0.8786 0.2109 0.2284 0.2460
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.8243	0.6877	2.653	0.0768 .
x	0.9824	0.2536	3.873	0.0305 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5676 on 3 degrees of freedom

Multiple R-Squared: 0.8334, Adjusted R-squared: 0.7778

F-statistic: 15 on 1 and 3 DF, p-value: 0.03046

Regression fit after deleting the extreme case—the new slope = 1 and intercept=2-- the coefficients barely changed!

Call:

```
lm(formula = y ~ x, data = data2)
```

Residuals:

1	3	4	5
-2.533e-17	4.222e-17	-8.445e-18	-8.445e-18

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.000e+00	4.388e-17	4.558e+16	<2e-16 ***
x	1.000e+00	1.602e-17	6.241e+16	<2e-16 ***

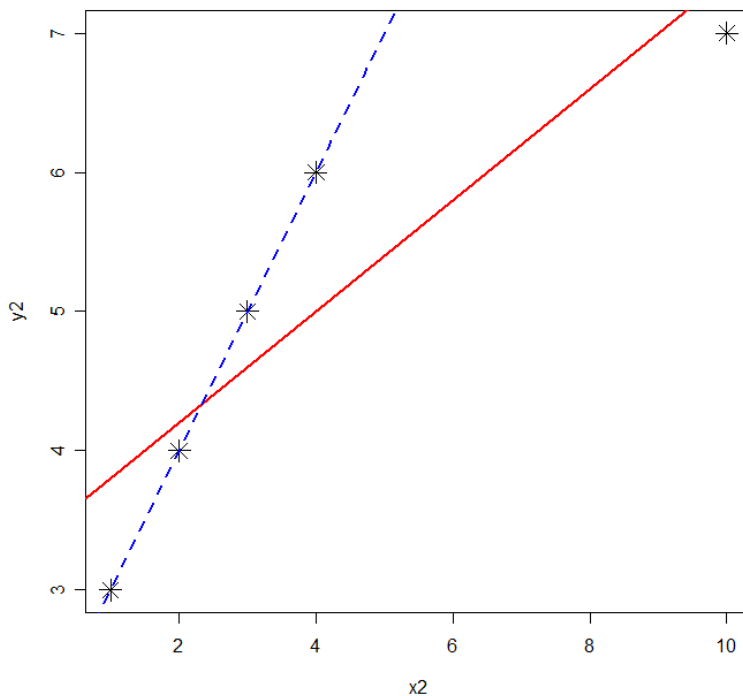
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.583e-17 on 2 degrees of freedom
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 3.895e+33 on 1 and 2 DF, p-value: < 2.2e-16

CASE 2:

In contrast, another outlier can have an unusual x value (i.e. outside the range of X distribution--high leverage). Deleting this case would markedly change both slope and intercept. In other words, this case exerts a strong leverage and influence on the regression coefficients.

The red line depicts the regression line with all observations. The blue dotted line shows the new fit with the extreme case dropped. Both slope and intercept changed markedly.



Note that the regression coefficients changed significantly.
Using all cases:

Call:

```
lm(formula = y2 ~ x2, data = data1)
```

Residuals:

```
 1  2  3  4  5
-0.8 -0.2 0.4 1.0 -0.4
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.4000    0.5888  5.775 0.0103 *
x2           0.4000    0.1155  3.464 0.0405 *
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8165 on 3 degrees of freedom

Multiple R-Squared: 0.8, Adjusted R-squared: 0.7333

F-statistic: 12 on 1 and 3 DF, p-value: 0.04052

Regression---excluding the extreme observation.

```
> summary(lm(y2~x2, data=data2)) #intercept and slope change significantly!
```

Call:

```
lm(formula = y2 ~ x2, data = data2)
```

Residuals:

```
 1  2  3  4
-2.533e-17 4.222e-17 -8.445e-18 -8.445e-18
```

Coefficients:

```
      Estimate Std. Error  t value Pr(>|t|)
(Intercept) 2.000e+00 4.388e-17 4.558e+16 <2e-16 ***
x2          1.000e+00 1.602e-17 6.241e+16 <2e-16 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

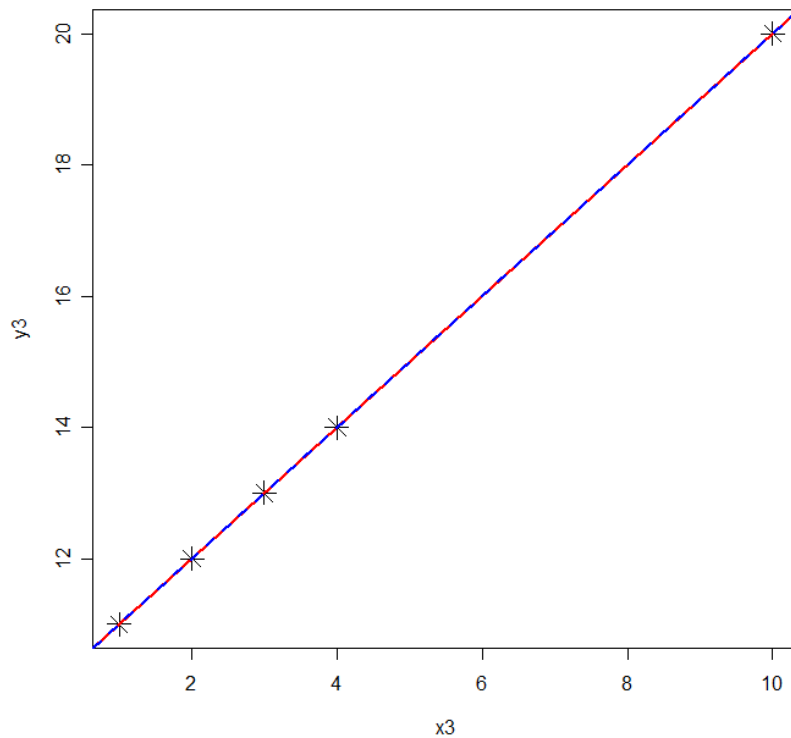
Residual standard error: 3.583e-17 on 2 degrees of freedom

Multiple R-Squared: 1, Adjusted R-squared: 1

F-statistic: 3.895e+33 on 1 and 2 DF, p-value: < 2.2e-16

Case 3:

The outlier (the last observation with high X and Y value) is far from the range of values for X and Y. Although this point has high leverage, excluding this point does NOT change the regression coefficients. In other words, this outlier has high leverage but no influence. This point is in line with the rest of the data and hence it is NOT a regression outlier. The red line represents regression line using all data points. The blue dotted shows the regression fit after dropping the outlier---the two lines overlap entirely.



Including or excluding this high leverage point does NOT change the regression result at all.

Call:

```
lm(formula = y3 ~ x3, data = data1)
```

Residuals:

```
      1      2      3      4      5  
-1.116e-15 2.899e-16 6.290e-16 5.536e-16 -3.565e-16
```

Coefficients:

```
      Estimate Std. Error  t value Pr(>|t|)  
(Intercept) 1.000e+01 6.117e-16 1.635e+16 <2e-16 ***  
x3          1.000e+00 1.200e-16 8.336e+15 <2e-16 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.483e-16 on 3 degrees of freedom
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 6.949e+31 on 1 and 3 DF, p-value: < 2.2e-16

```
> summary(lm(y3~x3, data=data2)) #intercept and slope change significantly!
```

Call:

```
lm(formula = y3 ~ x3, data = data2)
```

Residuals:

```
    1    2    3    4  
-2.533e-17 4.222e-17 -8.445e-18 -8.445e-18
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)  
(Intercept) 1.000e+01 4.388e-17 2.279e+17 <2e-16 ***  
x3          1.000e+00 1.602e-17 6.241e+16 <2e-16 ***
```

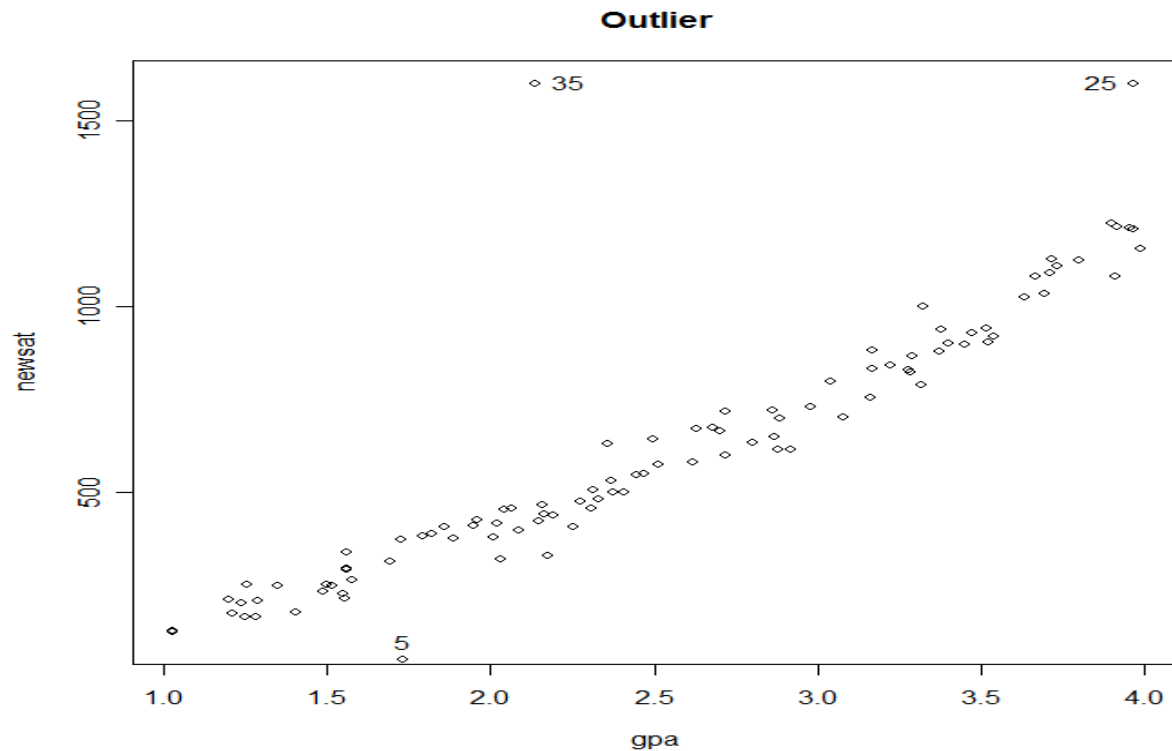
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.583e-17 on 2 degrees of freedom
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 3.895e+33 on 1 and 2 DF, p-value: < 2.2e-16

In sum, when we try to look for outliers, we have to be careful---are we looking for observations that have a) high leverage; b) high influence; c) both high leverage and influence?

EXAMPLE

I created a dataset with 2 variables: SAT score and GPA. I added in three cases--5th, 25th, 35th --- that seem to stand out from the rest of the data.



I fitted a regression line using ALL the data points. Here's the output. Notice the slope for GPA is 362.36.

Call:

```
lm(formula = newsat ~ gpa)
```

Residuals:

```
   Min     1Q  Median     3Q    Max
-279.74 -52.89 -12.59  31.05 1122.91
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -297.59    43.79  -6.796 8.45e-10 ***
gpa          362.36    16.43  22.058 <2e-16 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

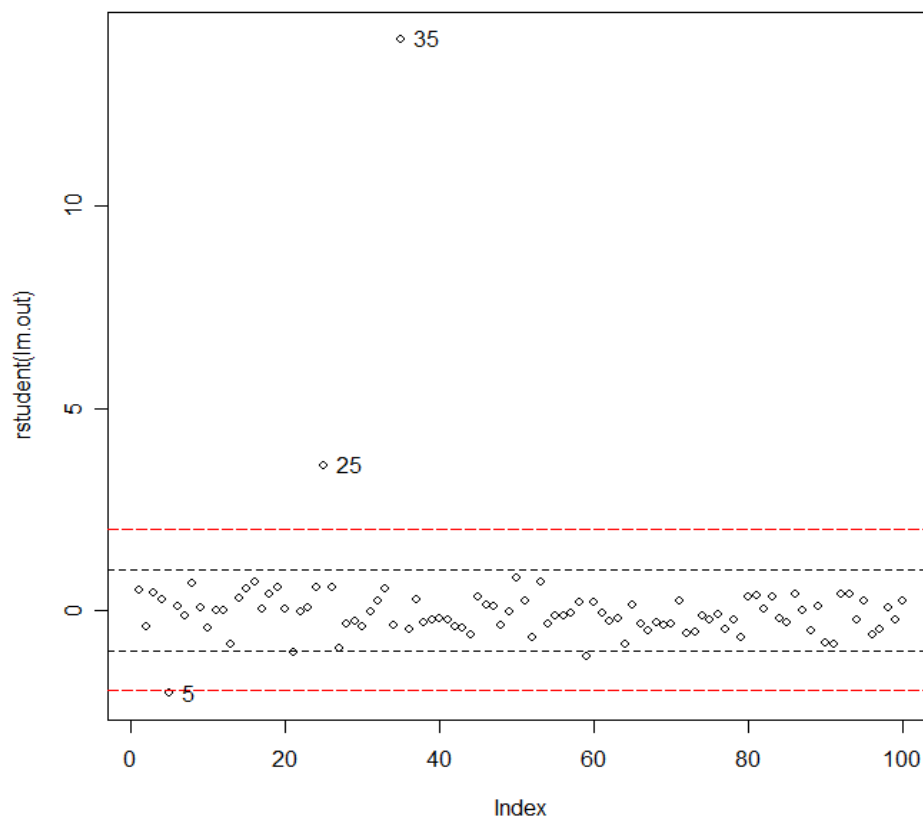
Residual standard error: 139.1 on 98 degrees of freedom

Multiple R-Squared: 0.8324, Adjusted R-squared: 0.8306
F-statistic: 486.6 on 1 and 98 DF, p-value: < 2.2e-16

MEASURING LEVERAGE

As we know for a normal distribution, 95% of data would fall within 2 standard deviations from the mean of the distribution. We can plot the standardized residuals and observe which cases fall beyond the two standard deviations.

The bulk of the observations fall within the two black dotted lines (which delineate the range of values one standard deviation from the mean). The red dotted lines show the bound for two standard deviations from the mean. Observation no. 5 falls about two standard deviations from the mean while the 25th and 35th are markedly further from the mean.



MEASURING INFLUENCE

Dfbetas gives the CHANGE in the estimated parameters if an observation is excluded, relative to its standard error.

$$D_{ij}^* = D_{ij} / (SE_{-i}(B_j))$$

the denominator contains the $SE(B_j)$ calculated without observation i

where:

$$D_{ij} = B_j - B_{\{j(-i)\}}$$

B_j is the regression coefficient obtained from using all data

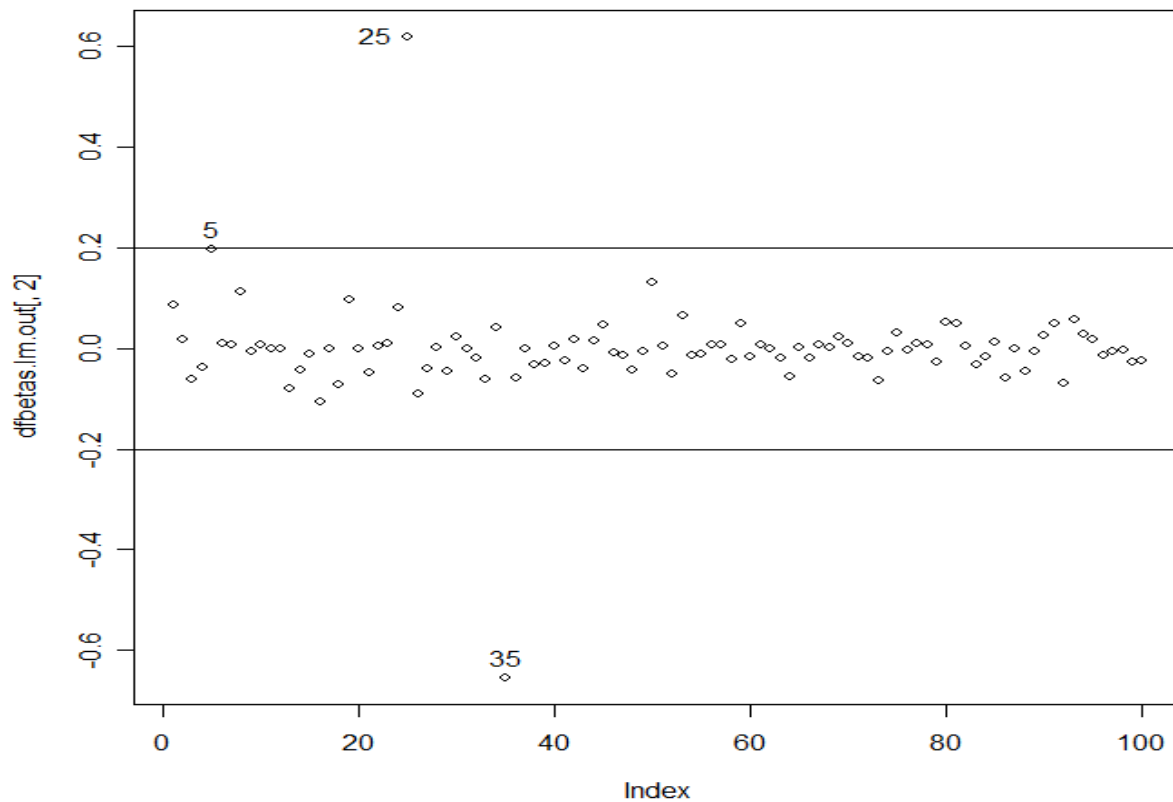
B_{j-1} is the regression coefficient obtained from using all data EXCEPT the excluded observation

Values can be positive or negative:

-Positive values indicate that deleting the observation yields a smaller coefficient

-Negative values indicate that deleting the observation yields a larger coefficient

$abs(D_{ij}^*) > 2/(sqrt(n))$ are typically considered large



The 25th and 35th observation fall sufficiently far from the bounds. The 5th observation seems to be a borderline case.

If we print out the dfbeta table, it shows how the coefficients would change if we were to drop that observation from the regression.

(Intercept) gpa

```

25 -0.4716727792 0.619518876
26 0.1027550295 -0.089333076
27 0.0074283060 -0.039610557
28 -0.0128458209 0.002244282
29 0.0347022996 -0.046013025
30 -0.0348672389 0.022383029
31 0.0008975843 -0.002119721
32 0.0257992524 -0.019645814
33 0.0750893037 -0.061286521
34 -0.0518092466 0.042302613
35 1.0734344772 -0.655392165

```

Since the GPA coefficient is negative, deleting 35th observation would make the regression coefficient bigger. Without 35th observation, the regression line will be pulled up to get closer to the 25th observation, and we would obtain 368.5 for the slope (compared to 362.36).

Call:

```
lm(formula = newsat ~ gpa, data = datatemp2)
```

Residuals:

```

      Min       1Q   Median       3Q      Max
-263.447 -41.559  -3.721   46.443  462.876

```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept) -324.589    25.222  -12.87 <2e-16 ***
gpa          368.541     9.445   39.02 <2e-16 ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79.87 on 97 degrees of freedom

Multiple R-Squared: 0.9401, Adjusted R-squared: 0.9395

F-statistic: 1523 on 1 and 97 DF, p-value: < 2.2e-16 #References