Outliers, Leverage, and Influence

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Outliers, Leverage, and Influence

Introduction

- After an initial fit of a model (including possible nonlinear transformations of X and or Y), it may become clear that certain observations are unusual in the sense that they are extremely atypical of X and or Y, either in the univariate or bivariate sense.
- In this module, we develop and discuss some of the common techniques and related terminology connected with the identification and evaluation of unusual observations.

Introduction

- In our examination of diagnostics based on residuals or possibly rescaled residuals, we begin by recalling that our basic model includes a model for the regression errors and, in a sense, the residuals themselves.
- If the fitted model does not give a set of residuals that appear to be reasonably in agreement with the model for those residuals, then we must question the model and/or its assumptions.

Introduction

- A related issue is the importance of each case on estimation and other aspects of the analysis.
- In some data sets, the observed statistics may change in important ways if just *one* case is deleted from the data.
- We will develop methods for detecting and identifying such *influential* cases.
- This will involve two types of diagnostic statistics, *distance measures* and *leverage values*.
- Although our emphasis will still be graphical, we can also develop numerical indices and related statistical tests.

A Graphical Demonstration

- It will help motivate our discussion if, before we dive into technicalities, we get "the big picture" in terms of a simple example.
- So let's start by digressing to a graphical example in RStudio.
- Open RStudio, make sure the `manipulate` package is installed, then, in the console window, type

```
> source("http://www.statpower.net/R2101/Leverage.R")
```

- This will open up a demonstration scatterplot, with sliders.
- If the sliders are not visible, click on the cog icon at the upper left of the graphics window.

A Graphical Demonstration I

- You'll see a scatterplot of 20 points on two variables. One of the points is marked in red, and has a value of $X = 0, Y = 1.6$.
- The regression line for the points is plotted in blue, and at the top of the plot, 3 statistics for this red point are given.
- These statistics are:

A Graphical Demonstration II

1. Leverage. This is a measure of how unusual the X value of a point is, relative to the X observations as a whole. Leverage of a point has an absolute minimum of $1/n$, and we can see that the red point is right in the middle of the points on the X axis, and has a residual of 0.05.
2. Studentized Residual. This is a measure of the size of the residual, standardized by the estimated standard deviation of residuals based on all the data *but* the red point. The red point is a barely detectable smidgen below the regression line, and has a Studentized Residual of $-.025$.

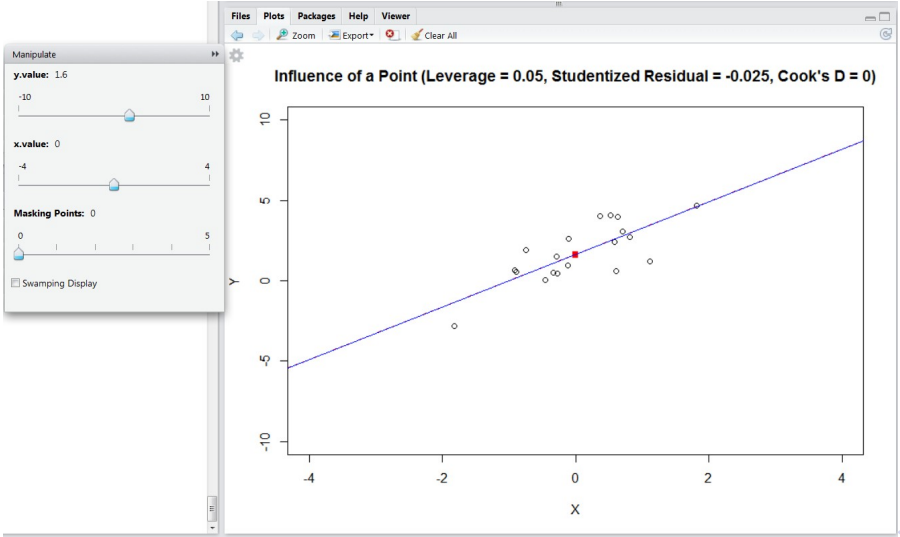
A Graphical Demonstration III

3. Cook's D. This statistic is a measure of the influence of the red point. Influence is the amount that the red point is affecting the regression line, measured by how much the regression line would change if the point were not included in the analysis. Cook's D is within rounding error of zero in this case. This point is almost right on the regression line, and removing it would not change the regression line detectably. In fact, there are actually two regression lines plotted on this graph. One is in red, and represents the linear fit with the red point included, and one is in blue, representing the linear fit for the 19 points excluding the red point. The blue and red regression lines are almost identical, so the red line cannot be seen.

A Graphical Demonstration IV

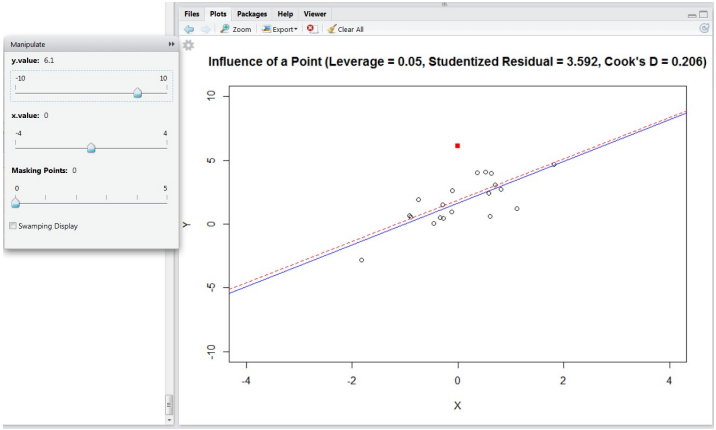
- We can see that, if a point is in the center of the X's and right on the regression line, it has essentially no *influence* on the regression line.

A Graphical Demonstration



A Graphical Demonstration

- Try the following: Click on the Y slider, and manipulate the Y value of the red point until it is 6.1, i.e., exactly 4.5 points higher than it's original Y value, which was 1.6.



A Graphical Demonstration

- In what sense is the red point now "an outlier"?
- Is it an outlier on X?
- Not at all. In fact, the minimal leverage value indicates that it is right in the middle of the X values.
- Is it an outlier on Y?
- Well, it is the highest Y value, but not by much more than the lower left point is removed from the other Y values.
- The red point at $(X = 0, Y = 6.1)$ is one we definitely not consider a univariate outlier on X, and possibly not consider a univariate outlier on Y either.

A Graphical Demonstration

- On the other hand, it is clear that the red point now departs significantly from the regression line. It is "way off" the regression line.
- Has this huge change in Y affected the regression line much? Not really. The Cook's D is only 0.206, and we can see that the regression line has not changed much.
- We've discovered that a point can be a regression outlier and yet *not* have much influence.
- It turns out influence is a function of leverage *and* the amount by which a point deviates from the regression line.

A Graphical Demonstration

- I want you to do the following. First, manipulate X to a value of -1 and investigate how the various indices, *and the regression line*, behave as you move Y close, and far away, from the regression line.
- Next, move X to -3 and repeat the process.
- How are leverage, the Studentized residual, and influence (Cook's D) interrelated?
- Work in groups of 3, spend about 5-10 minutes systematically playing with the plot, and summarize your findings.

Significance Tests for Outliers and Influential Cases

- We've seen that there are several statistical measures for identifying unusual observations.
- Leverage describes how unusual an observations is in predictor(s) data. Leverage becomes more complex when you have more than one regressor, and we'll return to it later.
- The Studentized Residual describes how unusual a point is relative to the regression line computed without that point.
- Cook's distance describes how much a point changes the regression line.
- At what point should we declare the Studentized Residual or Cook's distance to be "significant"?

Significance Tests for Outliers and Influential Cases

An Outlier Test

- Recall that, with the outlier red point positioned at $X = 0, Y = 6.1$, the Studentized Residual was 3.59. This has a t distribution with $n - 2$ degrees of freedom.
- The 2-sided p-value is

```
> 2*(1-pt(3.592,18))
```

```
[1] 0.002083955
```

- This could be considered highly significant — except that we've forgotten one thing. We have looked at the data and have selected a point for analysis, and we need to correct for multiple testing.
- A simple Bonferroni correction would require us to multiply the p-value by 20, obtaining

```
> 20*2*(1-pt(3.592,18))
```

```
[1] 0.04167909
```

- This outlier is significant, and can be rejected.
- It is simply too far from the regression line to be considered chance variation.

Significance Tests for Outliers and Influential Cases

An Outlier Test

- Can you compute a cutoff value for the Studentized Residual beyond which you would declare it significant?

Significance Tests for Outliers and Influential Cases

An Outlier Test

- If an individual test is significant at the $0.05/n$ level, 2-sided, then the outlier is significant.
- In our current situation, the 2-sided p value would be $0.05/20 = 0.0025$, and the 1-sided p value would be half that, or 0.00125. Hence we would need the studentized residual to be

```
> qt(1-0.00125,18)
[1] 3.510104
```

to be declared significant with a Bonferroni correction.

Significance Tests for Outliers and Influential Cases

A Significance Test for Influence

- At what cutoff point should a Cook's distance be declared significant?
- Opinion is divided on this issue. However many authors recommend a value of 1.00, while others such as Chatterjee and Hadi suggest more sophisticated criteria.
- Fox(2008, p. 255), citing Chatterjee and Hadi (1988), cites a cutoff of

$$D_i > \frac{4}{n-k-1} \tag{1}$$

- In general, one should be careful about relying on simplistic cutoffs.
- A simultaneous plot of the Cook's distance and Studentized Residuals for all the data points may suggest observations that need special attention.
- As we shall see in later examples, it is easy to obtain such plots in R.

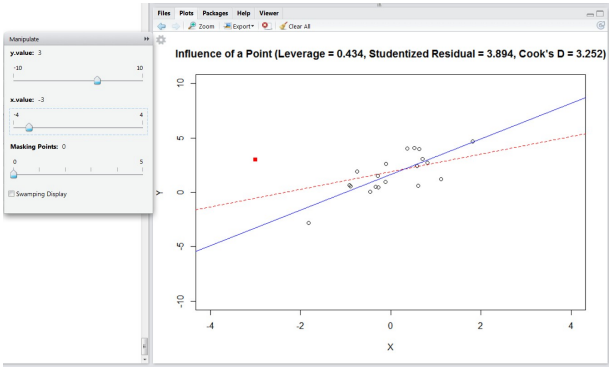
Problems with Multiple Outliers

- The problems relating to outlier identification are exacerbated when there is more than one outlier.
- As we shall see in a later module on multiple regression, these problems become especially acute with more than one predictor.
- For now, we'll use a simple visual demonstration to examine the problems of *masking* and *swamping* which can make it extremely difficult to identify the outliers.

Problems with Multiple Outliers

Masking

- Masking occurs when a group of outliers serves to move the fitted regression line near enough to them that they no longer appear to be outliers.
- To illustrate masking, reopen our outlier demonstration program in RStudio, and move the red point to coordinates of $X = -3, Y = 3$.



Problems with Multiple Outliers

Masking

- As you can see, the red point is now a high leverage, high influence observation with a high Studentized Residual and a high Cook's D.
- Note that there is another point on the lower left of the plot that is vairly close to the blue regression line, but quite a bit more removed from the red regression line calculated with the red outlier point in the data.

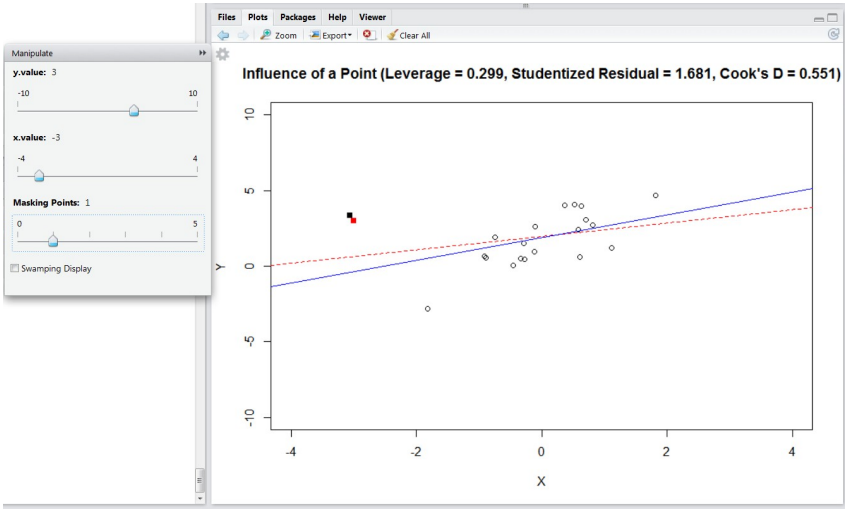
Problems with Multiple Outliers

Masking

- To illustrate masking, suppose that we now randomly generate a point around the red point, to simulate a "disparate process" that generates a second outlier in addition to a first outlier.
- Do do this, open up the slider controls and move the *Masking Points* slider to a value of 1.
- Note that both the blue and the red regression lines moved.
- This is because the blue line is calculated with all the data except the red point, and so it now included the black second outlier.
- Note now that the red outlier no longer seems like an outlier!
- Its leverage, Studentized Residual, and Cook's D have all changed to much more respectable values.

Problems with Multiple Outliers

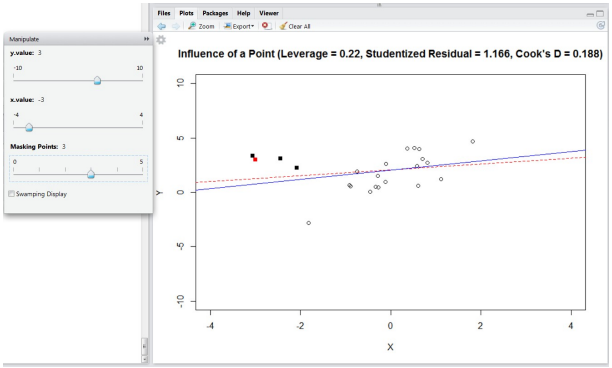
Masking



Problems with Multiple Outliers

Masking

- Now move the slider to add a second masking point. The fact that the red point and the black points are outliers has now been completely obscured.
- Each of the 3 outliers has been masked by the other two.
- Now add a third masking outlier, and things get even worse.



Problems with Multiple Outliers

Masking

- To move on to the next demonstration, remove all the masking outliers by pushing the slider back to zero.

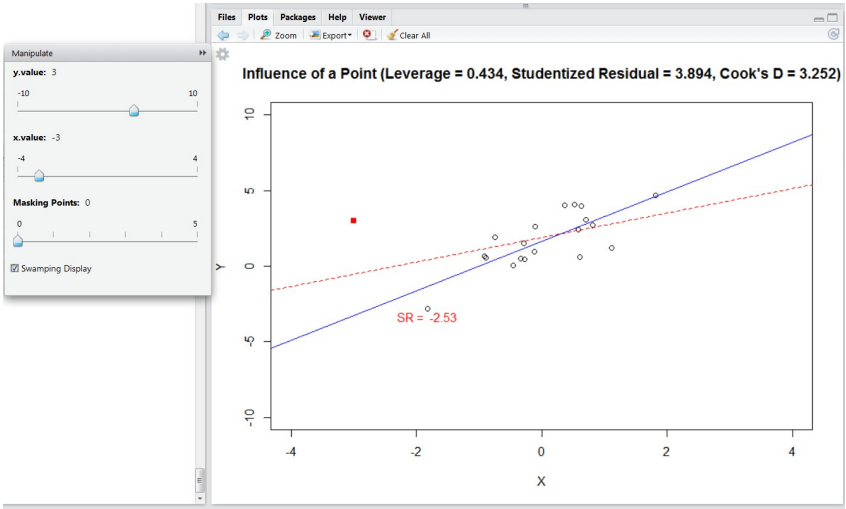
Problems with Multiple Outliers

Swamping

- With all the masking outliers removed and the red outlier point at $X = -3, Y = 3$, we are now going to investigate the problem of swamping.
- Swamping occurs when a group of true outliers mask each other and simultaneously make a point that is not a true outlier appear to be an outlier.
- Look at the point at the lower left of the plot. Now click on the *Display Swamping* checkbox. This will display the Studentizes Residual of the point just below it.
- Note that the Studentized Residual is -2.53 relative to the regression line (in red) based on all the data points, including the outlier, while the point is not particularly far away from the blue regression line based on the points without the outlier.
- Notice also that the red point has a Studentized Residual of 3.894.

Problems with Multiple Outliers

Swamping



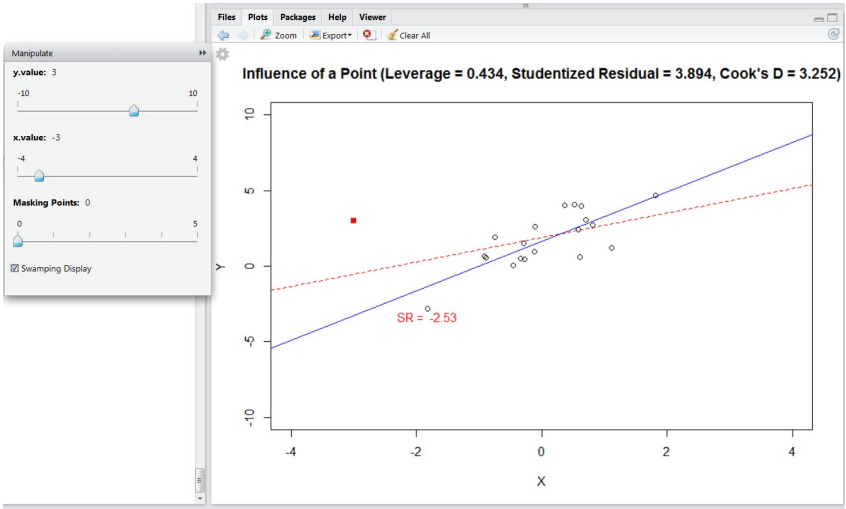
Problems with Multiple Outliers

Swamping

- Now slowly add 1,2,3,4,5 masking outliers to the display.
- Note how the legitimate point begins to look more and more like an outlier, as the Studentized Residual increases.
- We say that the legitimate point has been "swamped" by the outliers.

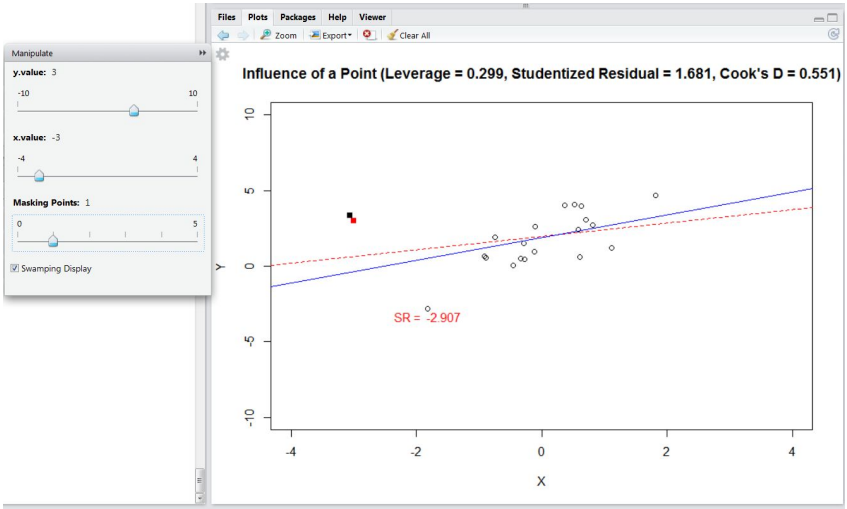
Problems with Multiple Outliers

Swamping



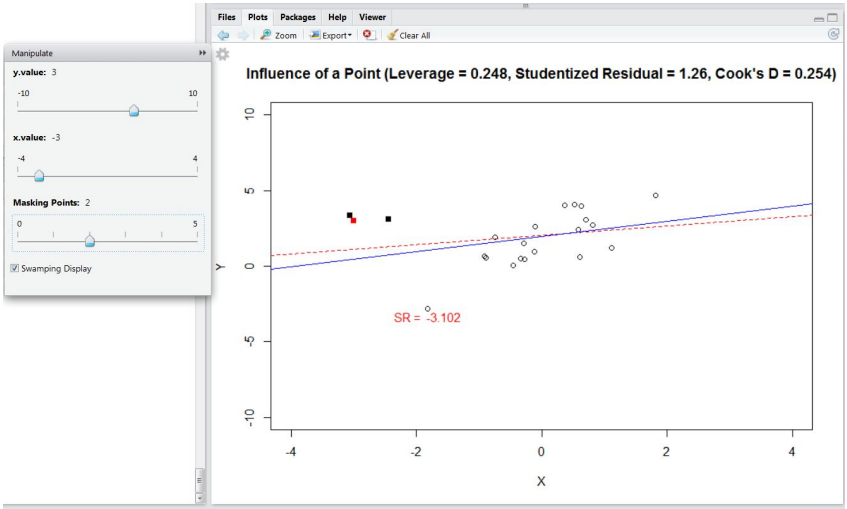
Problems with Multiple Outliers

Swamping



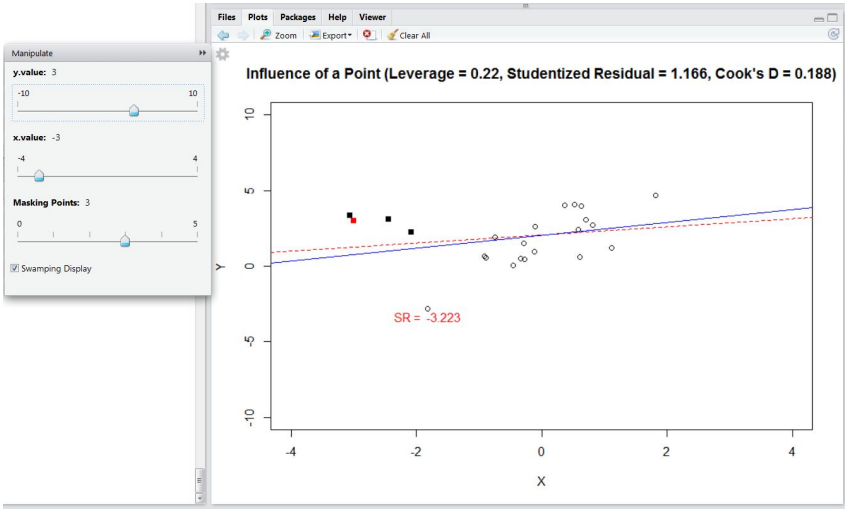
Problems with Multiple Outliers

Swamping



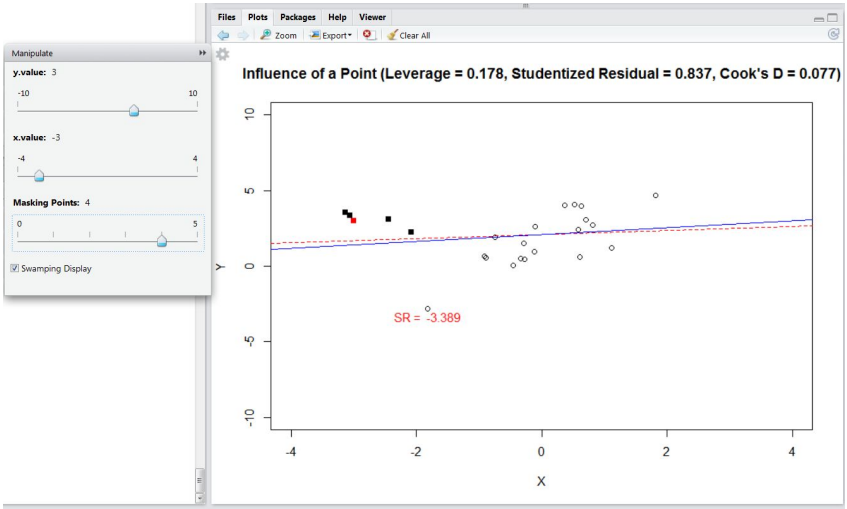
Problems with Multiple Outliers

Swamping



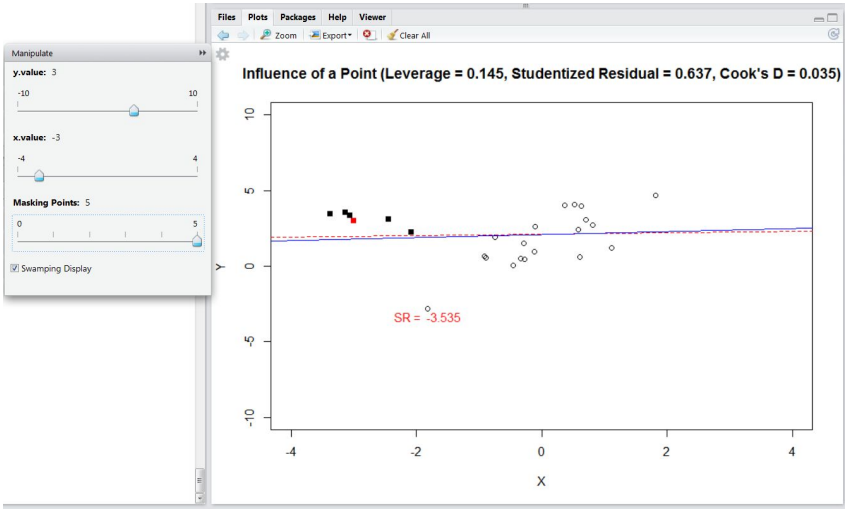
Problems with Multiple Outliers

Swamping



Problems with Multiple Outliers

Swamping



The Problem of Multiple Outliers

- All the well-known statistics for evaluating outliers are based on a *single mean-displaced* outlier model, which posits that a group of n scores has been invaded, as it were, by a single outlying observation.
- We just saw in our masking and swamping demonstrations that multiple outliers severely compromise the ability of this model to detect outlying observations.
- What can we do?

The Problem of Multiple Outliers

- There are two fundamental approaches to detection of multiple outliers.
- The *direct approach* uses one of several sequential detection methods to identify outliers and remove them.
- The *indirect approach* plots the mean function using a *robust regression* method, then uses standard outlier detection methods to identify the outliers in terms of the robust mean function.

What Should We Do with Outliers?

- If we identify an outlier, what should we do?
- Many regression textbooks have a section that deals with this thorny issue.
- Cohen, Cohen, Aiken and West(2003), p 411–419, have a very detailed and perceptive summary of issues and recommendations.
- A key is whether an outlier represents a *contaminated observation* or a *rare case*.

What Should We Do with Outliers?

Contaminated Observations

- A contaminated observation is one that has been damaged in some way. Some examples:
  1. *Error of execution of the research procedure.*
  2. *Inaccurate measurement of the dependent measure.*
  3. *Error in recording or keying in data.*
  4. *Error in calculating a measure.*
  5. *Nonattentive or distracted participants.*

What Should We Do with Outliers?

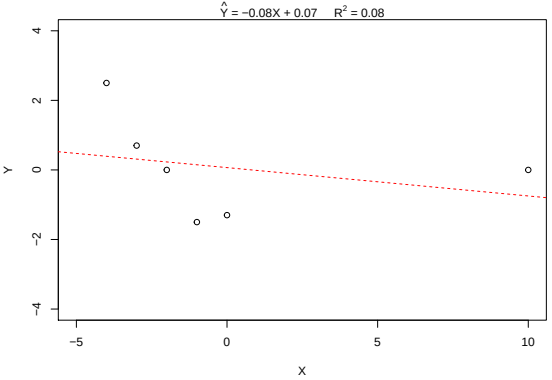
Rare Cases

- The outlier may simply be an extremely rare case.
- For example, a college freshman might be 12 years old and have an 800 SAT in math. Such an individual is extremely rare, but such people exist, and in fact a major research project at Peabody college has been studying such people for more than 30 years.
- Rare cases represent many possible processes. For example, the actual distribution of variables may represent a *mixture distribution*, in which a small percentage of cases come from a population with a mean vector and covariance matrix that are substantially different from the main population.
- The dependent variable may be one that naturally gives rise to extreme cases, such as number of absentee days for employees of a firm in a given month.

What Should We Do with Outliers?

An Incorrect Model

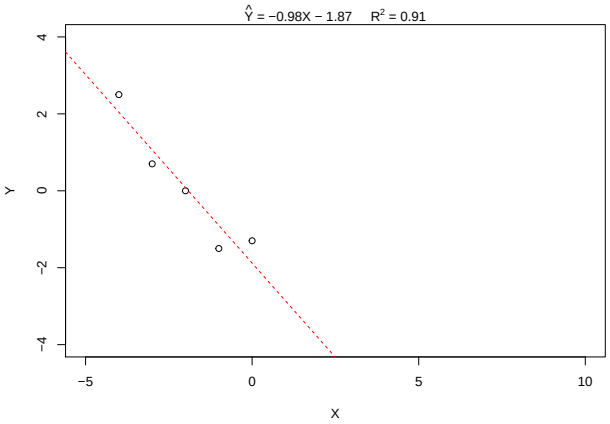
The regression model being studied may simply be incorrect. For example, Huber(1981) gives a case where a single point that might be considered an outlier in a linear fit is not an outlier in a quadratic fit of the data. Here is the linear fit with all points included.



What Should We Do with Outliers?

An Incorrect Model

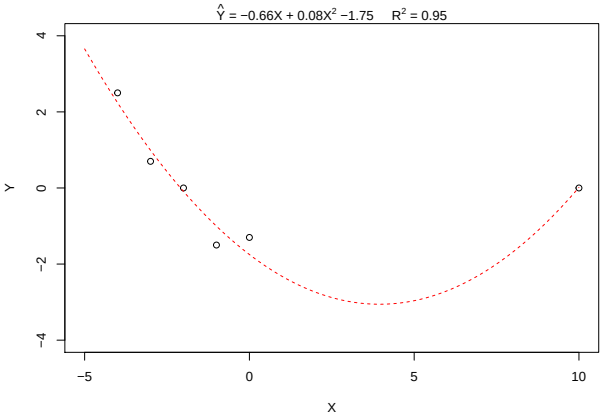
Here is the linear fit with the "outlier" deleted.



What Should We Do with Outliers?

An Incorrect Model

Here is the fit of the (correct) quadratic model with all points included.



Remedial Actions

- Suppose we have identifying one or more observations that may be outliers.
- If n is large, they may have little effect, but if n is small, their effect might be dramatic, as we have seen already.
- One option is outlier deletion and recomputation of the model fit.
- Another option is model respecification via transformation, polynomial fit, or a spline function.
- A third option is the use of "robust techniques."
- Robust techniques redefine the population model to be one that excludes extreme observations, and then proceed to fit that model.
- In one sense, they have "defined away" the problem of outliers, and they should be approached with caution.