A note on the relationship between leverage and residuals

Billy Fang

November 2, 2017

TL;DR: This is a slightly technical clarification of a claim made on page 7 of "Regression diagnostics," to dispel the misconception that "larger leverage means smaller residual." I also provide some code for a toy example that hopefully visualizes this concept.

Page 7 of "Regression Diagnostics" claims that if h_i is close to 1, then the fitted value $\widehat{y}_i$ is close to y_i. Specifically, it is shown that when $H_{ii} = h_i$ is very close to 1, then H_{ij} for all $j \neq i$ are each very close to zero, and thus

$$\widehat{y}_i = h_i y_i + \sum_{i \neq j} H_{ij} y_j \approx y_i.$$

This statement is a little vague because "h_i close to 1" and "$\widehat{y}_i$ close to y_i" are not quantified. How close h_i needs to be to 1 will depend on the y_i values (see the "$\approx$" in the above equation). By playing with the y_i value, it is easy to construct an example where a point has arbitrarily large leverage but still a "large" residual.

Perhaps a more concrete/mathematical statement of the above is the following.

> If the y_i values are *fixed*, and you change a particular point x_i so that its leverage h_i increases to 1, then the residual $y_i - \widehat{y}_i$ will [eventually] decrease to zero.

This can be verified by the equations above, since $h_i \to 1$ forces $H_{ij} \to 0$ for $j \neq i$, and then $\widehat{y}_i \to y_i$. Note that this statement is about what happens *in the limit* as $h_i \to 1$, and is not guaranteeing anything quantitative about *how* small the residual is when $h_i = 0.9$ for example. So, for example, at the end of "Regression Diagnostics II," the claim "if the leverage h_i is very large, then the residual $\widehat{e}_i$ will be small" is not entirely precise. Admittedly the statement is rather technical.

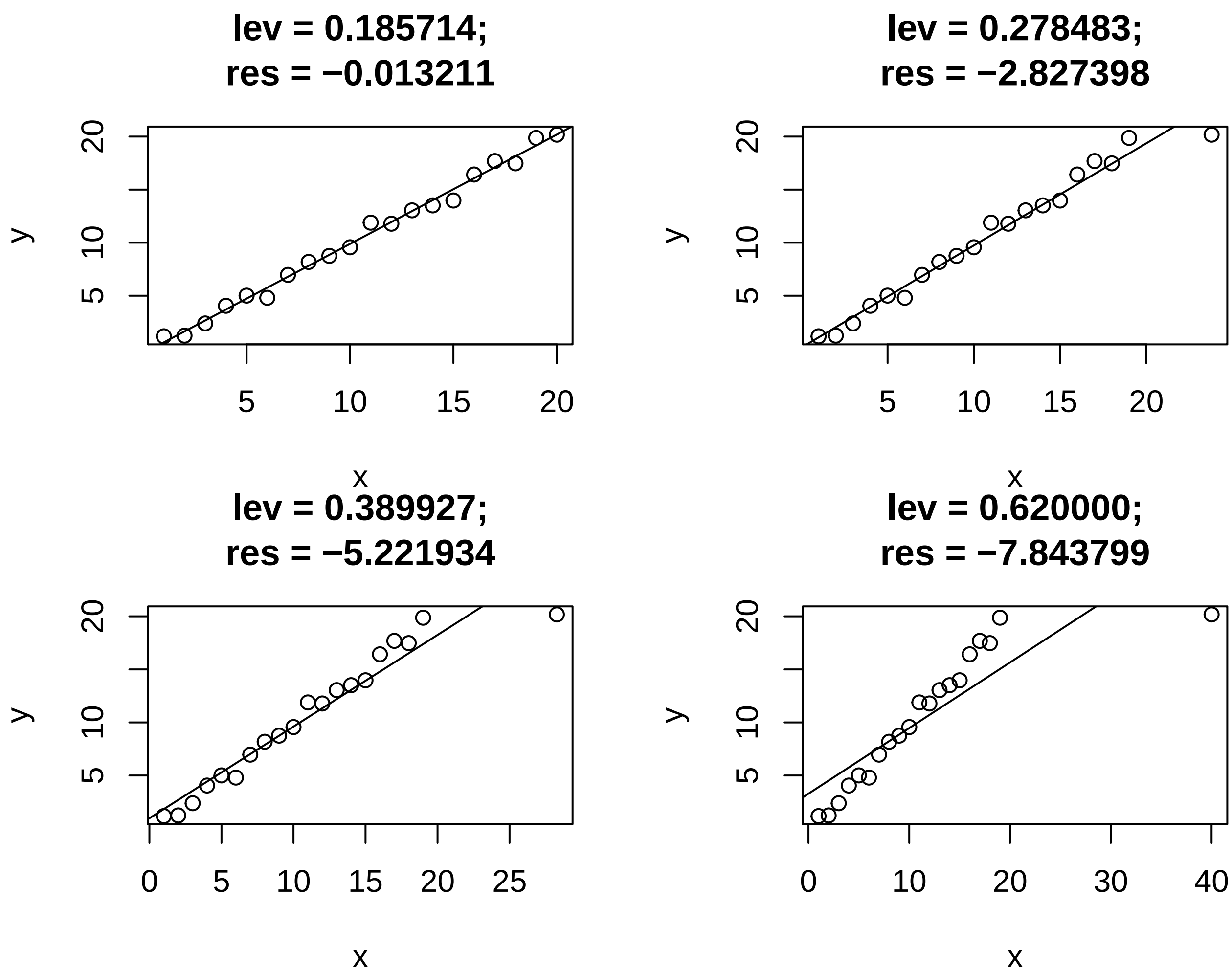
Another misinterpretation of the above result is "increasing a point's leverage a little bit will decrease the residual $y_i - \widehat{y}_i$" which is also not true. Again, the above result is only talking about what happens *in the limit* when $h_i \to 1$. It is possible for h_i to increase from 0.4 to 0.5 while the residual $y_i - \widehat{y}_i$ decreases, as in the example below.

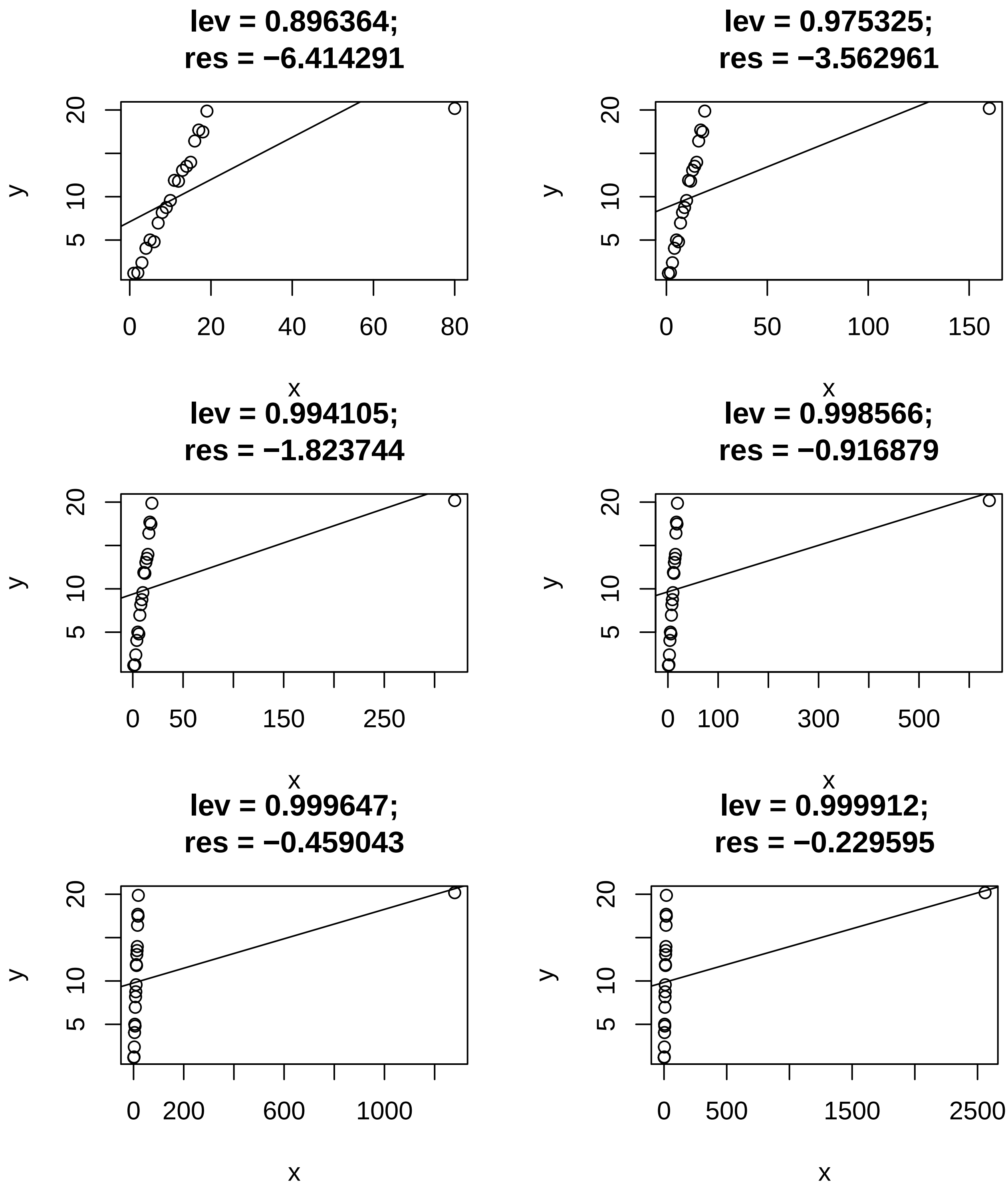
Below is a simple illustration of how, even when the y_i values are fixed, increasing leverage may cause the residual to increase. However, it also supports the claim that the residuals *eventually* decrease to zero as the leverage approaches 1. It may also be interesting to see how the line of fit changes as the leverage changes.

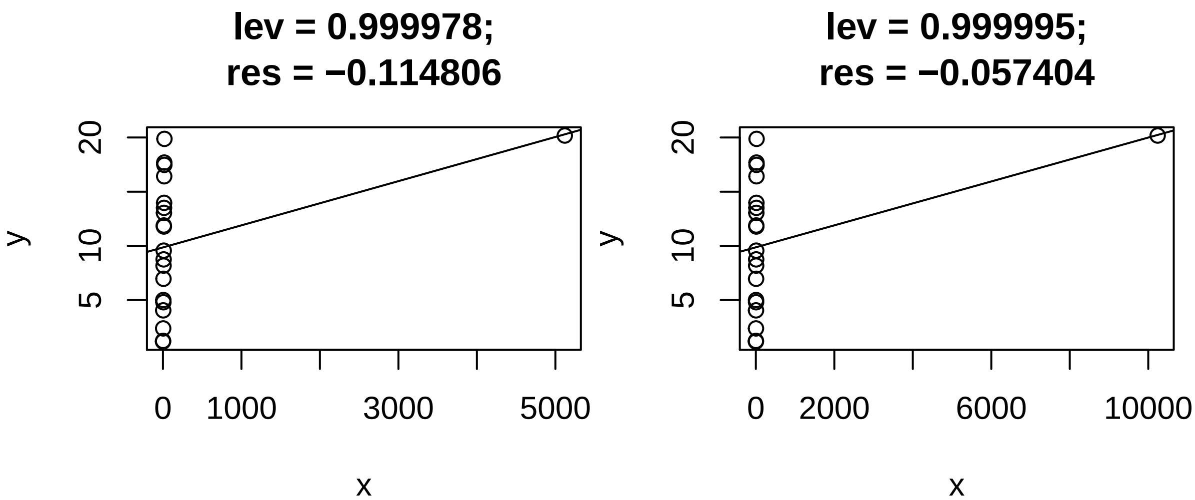
```r
n <- 20
x1 <- 1:n
y <- x1 + rnorm(n, sd=0.5)
# y <- 10 * x1 + rnorm(n, sd=0.5)
# try using this line instead, and see what happens to the scale of the residuals

mults <- 2^(c(0, 1/4, 1/2, 1:9))

x.list <- list(x1)

for (i in 1:length(mults)) {
  x.list <- c(x.list, list(c(1:(n-1), mults[i] * n)))
}
```

```r
lev <- rep(0, length(x.list))
res <- rep(0, length(x.list))
```

```r
for (i in 1:length(x.list)) {
  x <- x.list[[i]]
  mod <- lm(y ~ x)
  lev[i] <- influence(mod)$hat[n]
  res[i] <- resid(mod)[n]
  a <- coef(mod)[1]
  b <- coef(mod)[2]
  plot(x, y, main=sprintf("lev = %f;\nres = %f", lev[i], res[i]))
  abline(a=a, b=b)
}
```

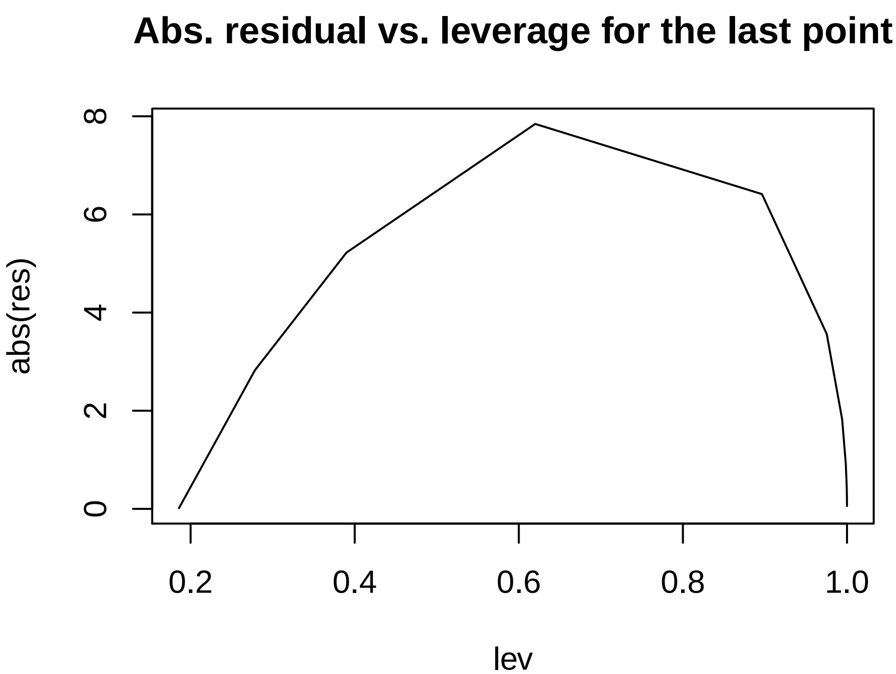



lev = 0.896364;
res = −6.414291
y
x
0 20 40 60 80
5 10 20
lev = 0.975325;
res = −3.562961
y
x
0 50 100 150
5 10 20
lev = 0.994105;
res = −1.823744
y
x
0 50 150 250
5 10 20
lev = 0.998566;
res = −0.916879
y
x
0 100 300 500
5 10 20
lev = 0.999647;
res = −0.459043
y
x
0 200 600 1000
5 10 20
lev = 0.999912;
res = −0.229595
y
x
0 500 1500 2500
5 10 20



Below is the plot of how the residual of the last point changes as the leverage increases. *Note that the residuals approach zero as leverage approaches* 1*, but for moderate values of leverage the residuals increase as well.*

```
plot(lev, abs(res), type='l', main="Abs. residual vs. leverage for the last point")
```



In the definition of `y` at the beginning of the code, try using the commented line `y <- 10 * x1 + rnorm(n, sd=0.5)` instead. The shape of all the plots will look exactly the same, but the scales on the vertical axis will change. In particular, in the last plot, compare what happens to the residual corresponding to leverage $h_i = 0.9$ before and after the change of the model. Thus, you cannot really say that the residual is “small” for large values of leverage like $h_i = 0.9$ or $h_i = 0.99$, since this depends on the y_i values.