Eco311 Optional Reading: Outliers and LAD estimator

(Jing Li, Miami University)

1. Outliers are extreme values. For instance, Bill Gates is an outlier in terms of personal income. It is well known that mean (average) is sensitive to outliers, while median is not. Median is an example of robust estimator.

2. There is a similar situation in the setting of regression. The main takeaway of this note is that OLS is sensitive to outliers, while LAD estimator is not. In short, LAD is robust.

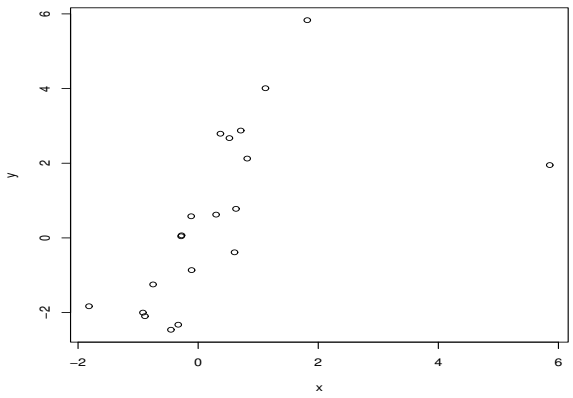
3. We use simulation to illustrate the sensitivity of OLS to outliers.

   (a) We generate 20 random values of y and x that satisfy

   $$y = 2x + e.$$

   Simulation implies that we know the true slope is 2. To produce an outlier, let's pretend the first x value is 0.6, but due to a typo, it is recorded as 6 in the data (i.e., we forget the decimal point, a common recording error).

```
> set.seed(12345)
> n = 20
> x = rnorm(n)
> y = 2*x + rnorm(n)
> x[1] = 10*x[1]
> plot(y~x)
```



The outlier is that rightmost point in the scatter plot.

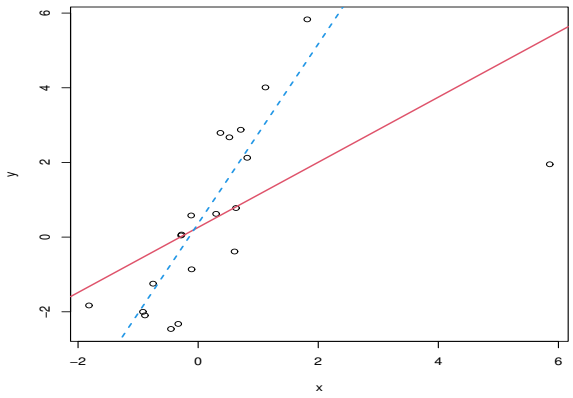
(b) Next we compare the OLS slope estimates when running a regression with whole sample and when the outlier is excluded (called leave-one-out regression)

```
> summary(lm(y~x))$coef
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 0.2603432  0.4462627 0.5833856 0.56687240
x           0.8721763  0.2902704 3.0047025 0.00760721

> summary(lm(y[x<3]~x[x<3]))$coef
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.3639418  0.2814821 1.292948 2.133251e-01
x[x < 3]    2.4049070  0.3404555 7.063793 1.902611e-06

> summary(lm(y~x))$r.squared
[1] 0.3340298
> summary(lm(y[x<3]~x[x<3]))$r.squared
[1] 0.7458786
```

Using the whole sample, the estimated slope is 0.8721763, far away from the true slope 2. When the outlier is excluded from the regression, we get a better estimate 2.4049070. Notice that excluding the outlier leads to better t value (from 3.0047025 to 7.063793) and R-squared (from 0.3340298 to 0.7458786) as well.

(c) We can draw the two regression lines and compare them

```
> plot(y~x)
> abline(lm(y~x),lty=1,lwd=2.0,col=2)
> abline(lm(y[x<3]~x[x<3]),lty=2,lwd=2.0,col=4)
```



We get the red solid line with outlier, and blue dash line without outlier. It is clear that the outlier pulls the OLS regression line toward itself, resulting in a worse fit.

(d) The `studres` function in `MASS` package can be used to detect potential outliers. Google "R, Studentized Residual" to learn more

```
> library(MASS)
> ehat = studres(lm(y~x))
> x[which(abs(ehat)>1.96)]
[1] 5.855288 1.817312
```

In this example, we find two potential outliers with x values of 5.855288 and 1.817312. The first one is indeed that outlier (that rightmost point).

4. Now the question is, how to handle outliers? There are several possibilities: (i) correcting the typo if that is the reason for the outlier; (ii) excluding outliers from the OLS regression; (iii) do not drop outliers, but use an alternative estimator that is robust to outliers, something that works like median

5. The alternative estimator is Least Absolute Deviations (LAD) estimator, which aims to solve the following objective function

$$\hat{\beta}^{LAD} \; minimizes \; \sum_i |residual_i| \tag{1}$$

Because residual or prediction error is not squared in above formula, the penalty for prediction error is smaller compared to the OLS estimator. Put differently, LAD downplays outliers by not paying too much attention to them.

6. One disadvantage of LAD estimator is that formula (1) involves absolute value function, which is not smooth or differentiable. As a result, we can not apply calculus (taking derivative and setting it to zero) to obtain an analytic solution. Numerical methods should be used to find LAD estimate. Another disadvantage is that bootstrap is needed to obtain the standard error of LAD, or measurement of sampling uncertainty.

7. Here we use the simplest numerical method called grid search. Basically we try a range of possible values as the slope estimate, evaluate the objective for each value, and see which one minimizes the sum of absolute residual.

```
> # Least Absolute Deviations Estimate
> blad = seq(0, 5, by=0.01)
> obj = rep(0, length(blad))
> for (i in 1:length(blad)) {
+ obj[i] = sum(abs(y - blad[i]*x))
+ }
> blad[obj==min(obj)]
[1] 1.24
```

   The LAD slope estimate is 1.24, which is better than the OLS estimate 0.8721763. Notice that we obtain this better estimate without excluding the outlier.

8. The `L1pack` package can be used to obtain an almost the same estimate

```
> library(L1pack)
> lad(y~x-1, method = "EM")
...
Coefficients:
x
1.2356
```

   The -1 in the `lad` function means "do not include intercept".

9. To summarize, it is a good idea to compare OLS results with and without outliers, or report the robust LAD estimate.