MTH499/599 Lecture Notes 08

Donghui Yan

Department of Math, Umass Dartmouth

Apr 08, 2015

Outline

- Leverage of an observation

Leverage of an observation

- Amount $\hat{y}_i$ would change if y_i is shifted by one unit
- Leverage of the i^{th} observation equals h_{ii}

Proof.

Consider the i^{th} observation. Since $\hat{Y} = HY$, we have

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + ... + h_{in}y_n = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j.$$

Assume y_i is increased by one, i.e., $\tilde{y}_i = y_i + 1$. Then $\hat{y}_i$ becomes

$$\tilde{\hat{y}}_i = h_{ii}(y_i + 1) + \sum_{j \neq i} h_{ij}y_j,$$

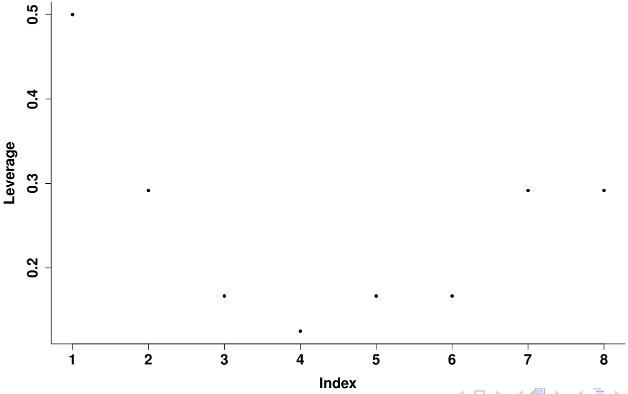
and the result follows. $\square$

Leverage of an observation (continued)

- The leverage of a point is considered large if it exceeds $2p/n$
  - The total leverage of all observations is $trace(H) = p$
  - The average leverage is p/n
- Observation: The further x_i is from $\overline{x}$, the larger leverage and more sensitive is to changes in y_i.

Leverage of an observation (continued)

```
> leverage<-hat(model.matrix(mylm));
> plot(leverage,xlab="Index", ylab="Leverage", pch=19);
x: 6  5  4  3  2  2  1  1
y: 6  9  8 10 11 12 11 13
```



Influential point and Cook's distance

- An *influential* point is one if removed would significantly change the *estimate*
  - Note difference between influential and high leverage
- An influential point may either be an outlier or have large leverage, or both
  - Typically true for at least one
- *Cook's distance* is a commonly used influence measure.

Cook's distance

- Cook's distance is defined by

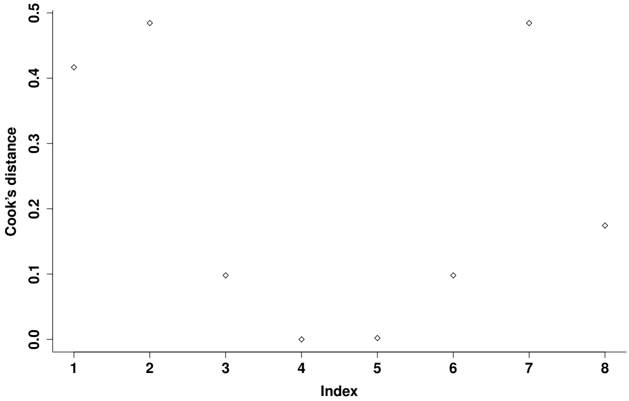
$$D_i = \frac{\sum_j (\hat{y}_j - \hat{y}_j(i))^2}{ps^2}$$

  - Where $\hat{y}_j(i)$ is the fit of j^{th} point with point i removed
  - Cook's distance uses sum of squared differences
    - As an surrogate for changes in estimate for simplicity
- A rule of thumb for potential outliers if

$$D_i \geq 4/(n-p).$$

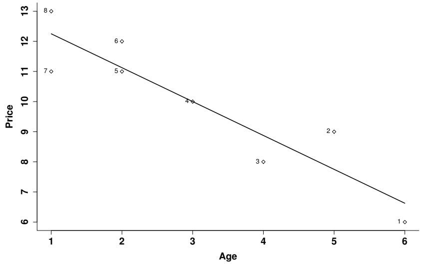
Cook's distance

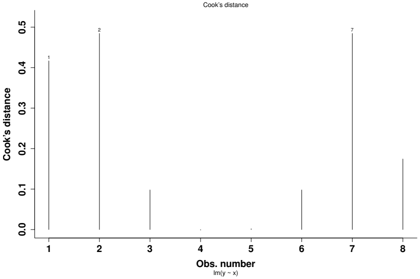
```
> cook<-cooks.distance(mylm);
```



Cook's distance

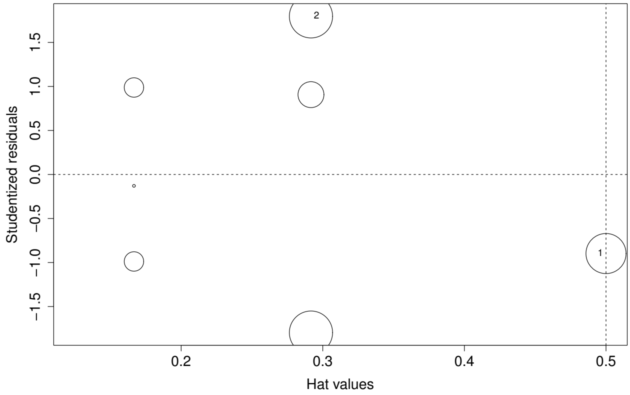
```
>cutoff<-4/(length(x)-length(mylm$coefficients)-2);
>plot(mylm,which=4, cook.levels=cutoff, main="", cex.lab=1.5,
cex.axis=1.5,bty="l",pch=20, font.axis=2, font.lab=2);
```





Influence plot of the toy example

```
>influencePlot(mylm,xlab="Hat values",
                        ylab="Studentized residuals",
                        cex.lab=1.5,cex.axis=1.5);
```

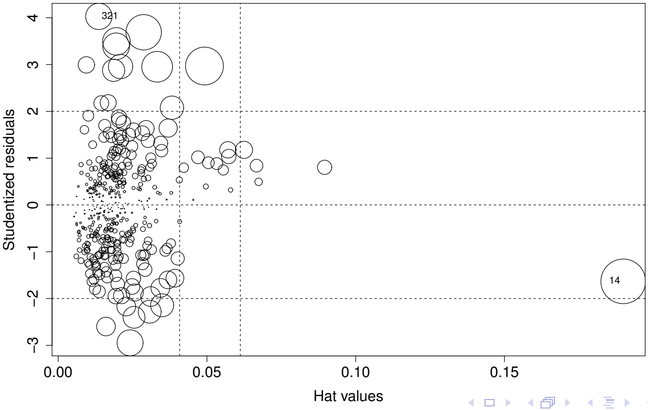


The auto MPG example

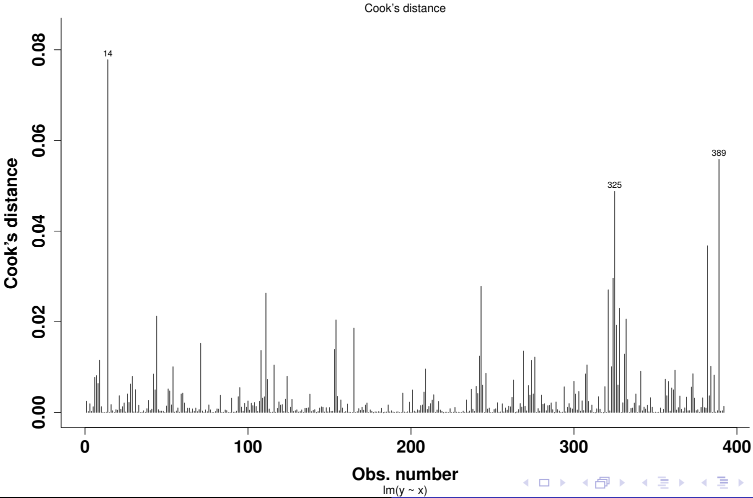
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.218435   4.644294  -3.707  0.00024 ***
x1           -0.493376   0.323282  -1.526  0.12780
x2            0.019896   0.007515   2.647  0.00844 **
x3           -0.016951   0.013787  -1.230  0.21963
x4           -0.006474   0.000652  -9.929  < 2e-16 ***
x5            0.080576   0.098845   0.815  0.41548
x6            0.750773   0.050973  14.729  < 2e-16 ***
x7            1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

The auto MPG example

	mpg	cyl.	disp.	hp	wt	acc	yr	origin	carname
14	14.0	8	455	225	3086	10.0	70	1	buick estate wagon(sw)
321	46.6	4	86	65	2110	17.9	80	3	mazda glc



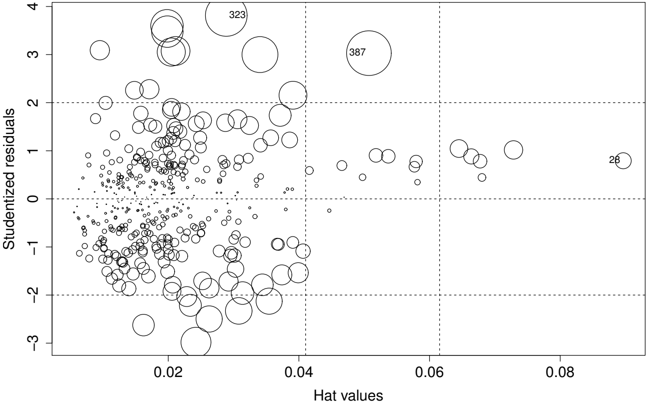
The auto MPG example (Cook's distance)



The auto MPG example (removing potential outliers

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.831032   4.559113  -3.692 0.000255 ***
x21          -0.564607   0.317985  -1.776 0.076599 .
x22           0.022678   0.007596   2.986 0.003011 **
x23          -0.010906   0.013948  -0.782 0.434764
x24          -0.006874   0.000691  -9.948  < 2e-16 ***
x25           0.107458   0.099091   1.084 0.278852
x26           0.747046   0.049982  14.946  < 2e-16 ***
x27           1.342664   0.272952   4.919 1.29e-06 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
Residual standard error: 3.256 on 382 degrees of freedom
Multiple R-squared:  0.8254, Adjusted R-squared:  0.8222
F-statistic: 257.9 on 7 and 382 DF,  p-value: < 2.2e-16
```

The auto MPG example (outliers removed)



The auto MPG example (outliers removed)