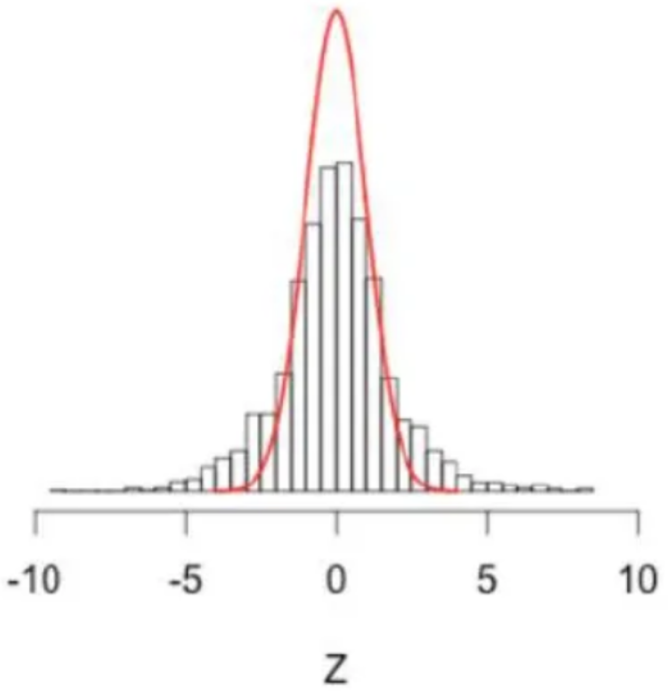
Lecture 14: Regression Diagnostics III

Ailin Zhang

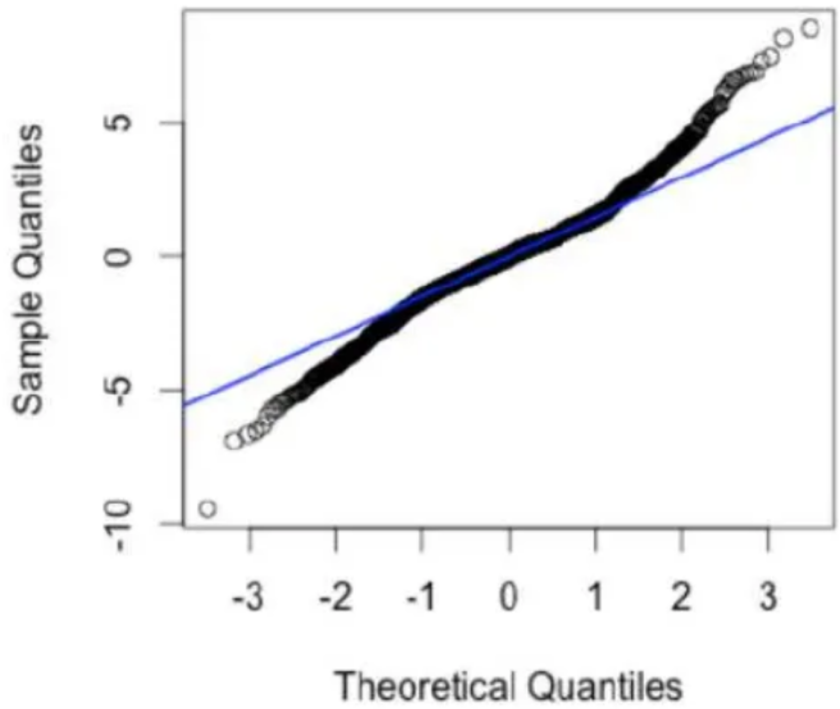
2023-06-13

Normal Probability Plot

Heavy Tailed

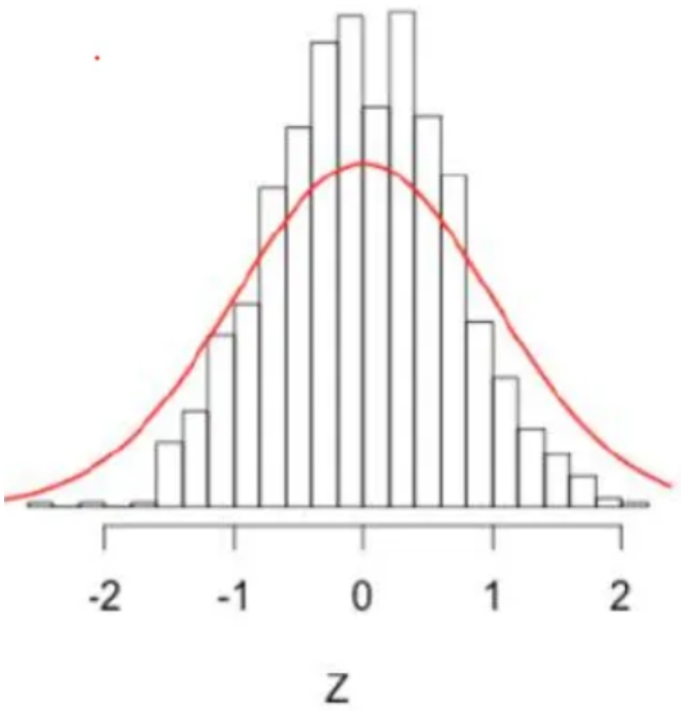


Normal QQ Plot

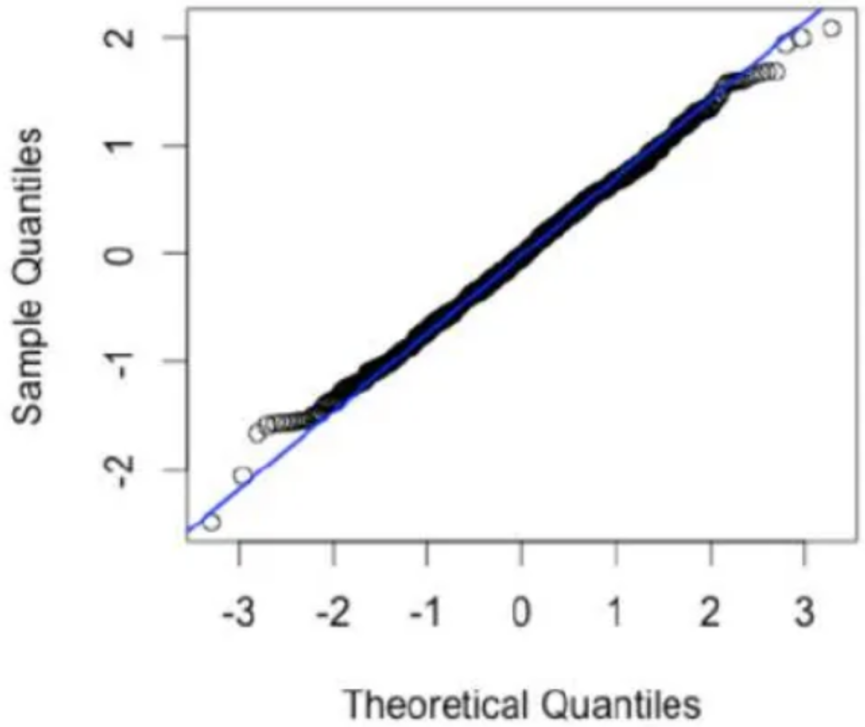


Normal Probability Plot

Light Tailed



Normal QQ Plot



Agenda

- Revisit Regression Assumptions
- Leverage
- Types of Residuals
- Residual Plots
- Graphical Methods
- Influential Points and Outliers (Today)
- Measure of Influence (Today)

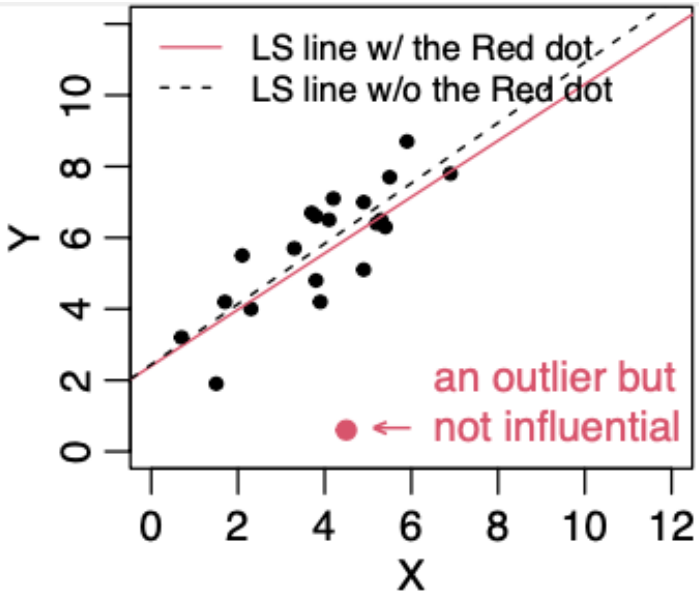
Outliers

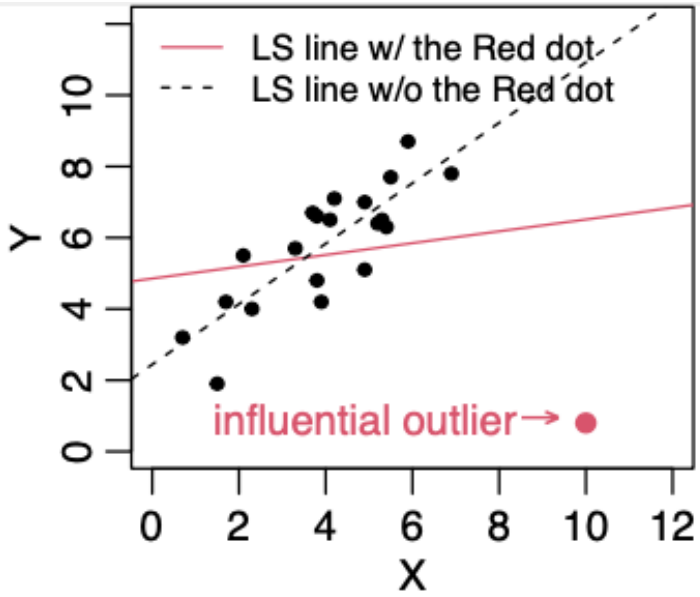
- An **outlier** is a point that the model fails to explain.
- Outliers are observations that do not follow the same model as the majority.
- The Outlier has a large residual. (Note: not all large residual are outliers)
- We use the studentized residual r_i^* to detect outliers.

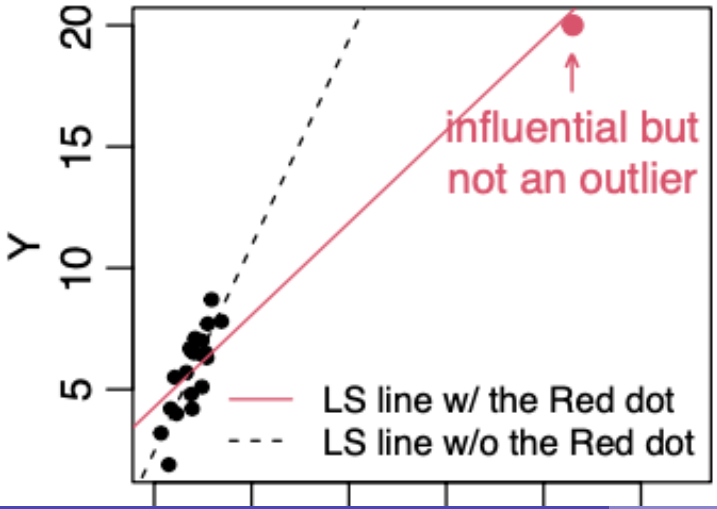
Influential Points

- An influential point has an unduly large effect on the model. The fitted model changes drastically when it is included.
- Observations whose removal will cause major changes in the regression analysis are called influential points.
- Changes such as the predicted responses, the estimated slope coefficients, or the hypothesis test results can be considered
- Influential points are not necessarily outliers
- A point can be influential, an outlier, or both

Outliers vs Influential Points







Example – New York Rivers

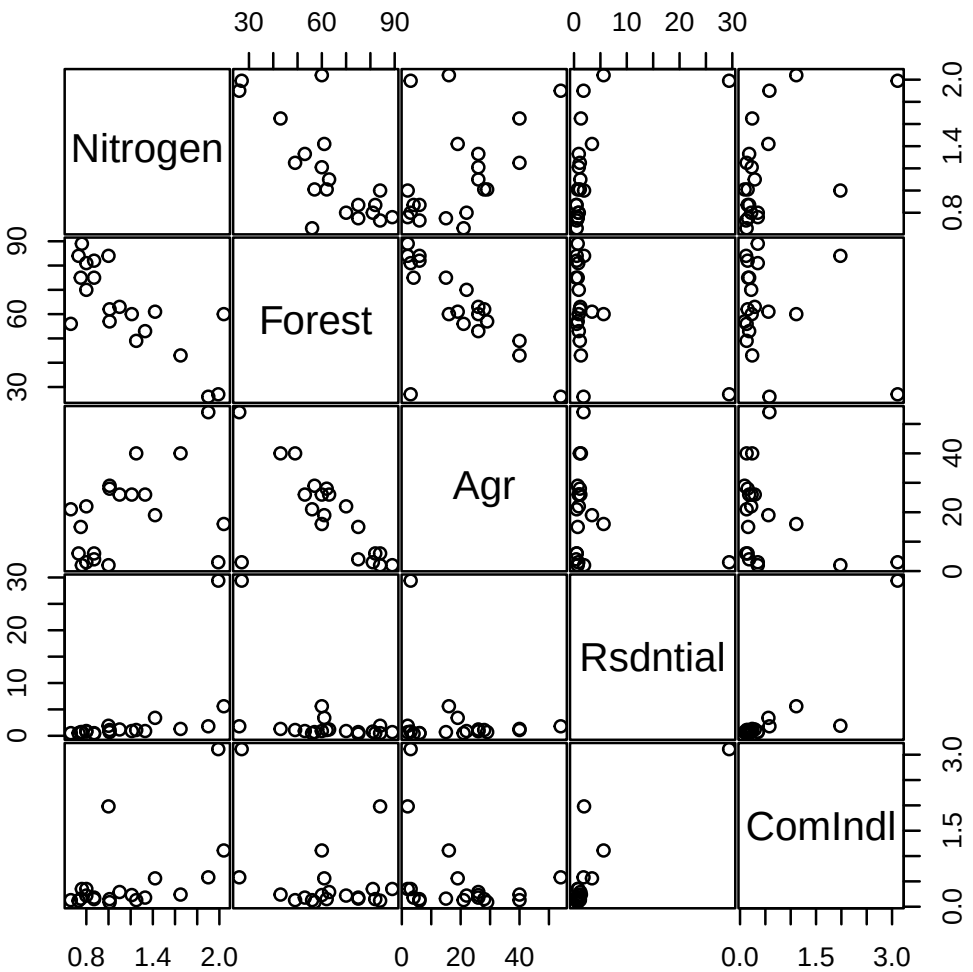
Data on Water Pollution in New York Rivers

- Nitrogen = Mean nitrogren concentration (mg/liter) measured at regular intervals (Response)
- Agr = % of land currently in Agricultural use
- Forest = % of Forest land
- Rsdntial = % of land in Residential use
- ComIndl = % of lane in Commercial or Industrial use

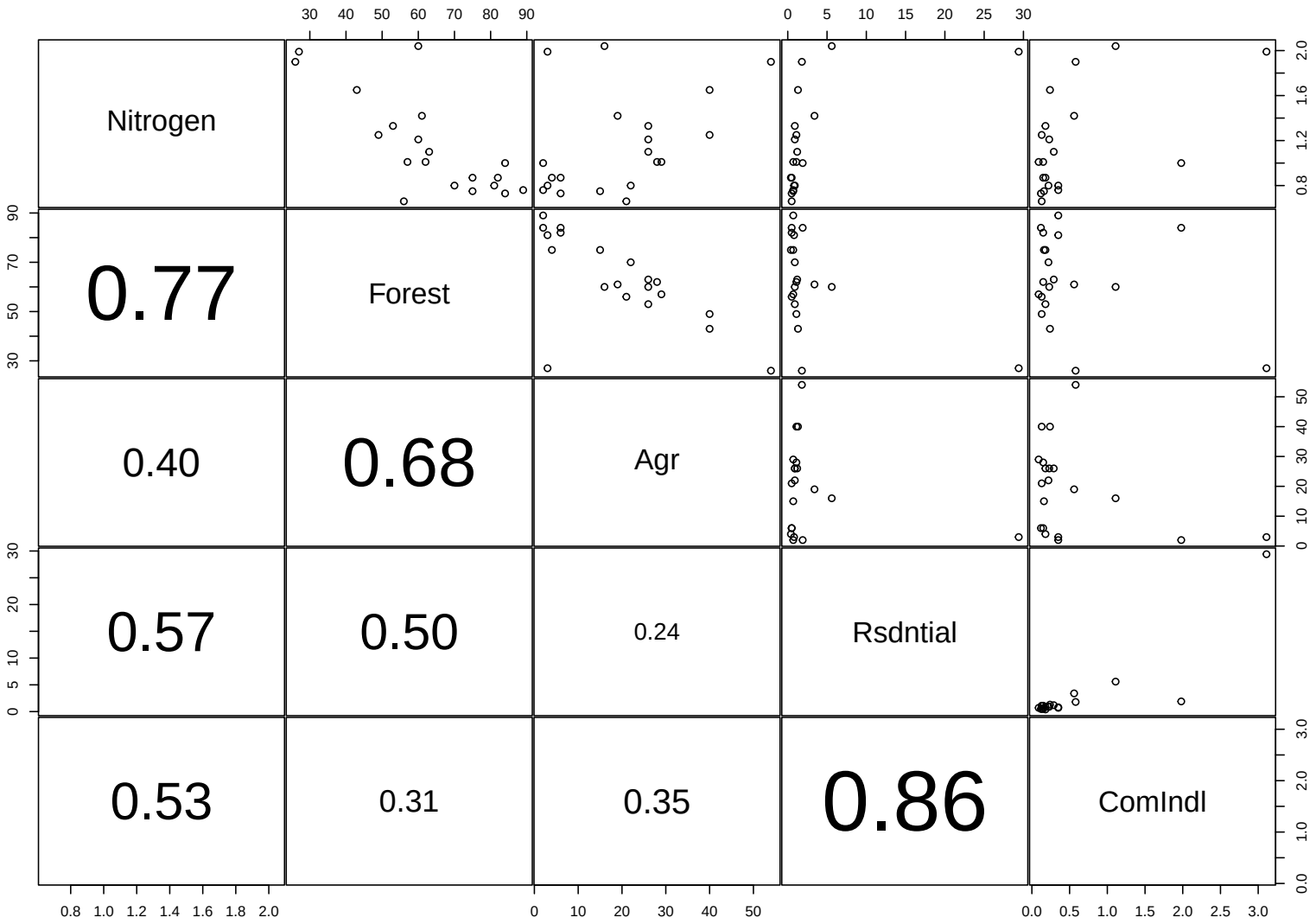
```
NYrivers = read.table("river.txt", h = T, sep="\t")
```

Pairwise Scatterplots

```
pairs(Nitrogen ~ Forest + Agr + Rsdntial + ComIndl,
      data=NYrivers, gap=0.1, oma=c(2,2,2,2))
```



A Fancier Scatterplot Matrix



R codes for the Fancier Scatter Plot Matrix

```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor)
    {
     usr <- par("usr"); on.exit(par(usr))
     par(usr = c(0, 1, 0, 1))
     r <- abs(cor(x, y))
     txt <- format(c(r, 0.123456789), digits = digits)[1]
     txt <- paste0(prefix, txt)
if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
     text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs( ~ Nitrogen + Forest + Agr + Rsdntial + ComIndl ,
       data=NYrivers, gap=0.1,oma=c(2,2,2,2),
       lower.panel = panel.cor)
```

Linear models

```
lm1 = lm(Nitrogen ~ Forest + Agr + Rsdntial + ComIndl,
         data=NYrivers)
lm1noH = lm(Nitrogen ~ Forest + Agr + Rsdntial + ComIndl,
            data=subset(NYrivers, River!="Hackensack"))
lm1noN = lm(Nitrogen ~ Forest + Agr + Rsdntial + ComIndl,
            data=subset(NYrivers, River!="Neversink"))
```

We will focus on the coefficient of `Rsdntial`

```r
summary(lm1)$coef # all data
```

```
##                 Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)  1.722213529 1.23408166  1.3955426 0.18316946
## Forest      -0.012967887 0.01393148 -0.9308336 0.36667966
## Agr          0.005809126 0.01503396  0.3864001 0.70462624
## Rsdntial    -0.007226768 0.03383005 -0.2136198 0.83372002
## ComIndl      0.305027765 0.16381667  1.8620069 0.08230952
```

```r
summary(lm1noH)$coef # w/o Hackensack
```

```
##                 Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept)  1.626014115 0.781091103  2.0817215 0.05619948
## Forest      -0.012760349 0.008814947 -1.4475809 0.16975563
## Agr          0.002352222 0.009539146  0.2465863 0.80880737
## Rsdntial     0.181160986 0.044390049  4.0811171 0.00112280
## ComIndl      0.075617570 0.113957223  0.6635610 0.51774981
```

```r
summary(lm1noN)$coef # w/o Neversink
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  1.099471134 0.91163575  1.2060421 0.2477883387
## Forest      -0.007589231 0.01022207 -0.7424360 0.4700975391
## Agr          0.010136685 0.01098383  0.9228732 0.3717054741
## Rsdntial    -0.123792917 0.03933701 -3.1469831 0.0071342930
## ComIndl      1.528956204 0.34371909  4.4482725 0.0005512227
```

Model Summary

For the coefficient of `Rsdntial`

- NOT significant using all data
- significantly positive if Hackensack is removed
- significantly negative if Neversink is removed

How to systematically detect influential points?

- Cook's Distance

Measure of Influence

- Suppose we suspect that observation i is influential.
- To test this, re-fit the model without ith observation.
  - $\hat{\beta}_{j(i)}$: fitted regression coefficient for jth predictor
  - $\hat{y}_{j(i)}$: jth fitted value
  - $\hat{\sigma}_{(i)}$: residual standard error
- Measurements of influence look at quantities like $\hat{\beta}_j - \hat{\beta}_{j(i)}$ or $\hat{y}_j - \hat{y}_{j(i)}$

Cook's Distance

Cook's distance measures the difference between the fitted model values between the full data set and the -(i) data set.

$$C_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)}$$

for $i = 1, 2, \ldots, n$

- A more computationally efficient way to C_i:

$$C_i = \frac{1}{p+1} \cdot r_i^2 \cdot \frac{h_{ii}}{1 - h_{ii}}$$

- The second term $h_{ii}/(1 - h_{ii})$ is called the potential.
- Influential points have high a C_i compared to the other points

Indentifying Influential Points Using Cook's Distance

- Simple Rule: Influential if $C_i > 1$
- A more sophisticated rule: Influential if C_i exceeds the 50th percentile of the F-distribution with $df = (p+1, n-p-1)$ degrees of freedom

```
qf(0.5, p+1, n-p-1)
```

For the NY Rivers data $n = 20, p = 4$, the threshold is

```
qf(0.5, 4+1, 20-4-1)
```

```
## [1] 0.9107243
```

- A graph of C_i vs. i will help us to detect influential points.

Summary R commands for Diagnostics

```
model$fit ## fitted value
model$res ## raw residuals
rstandard(model) #internally standardized residuals
rstudent(model) #externally studentized residuals
hatvalues(model) #leverage
cooks.distance(model) # Cook's distance
```

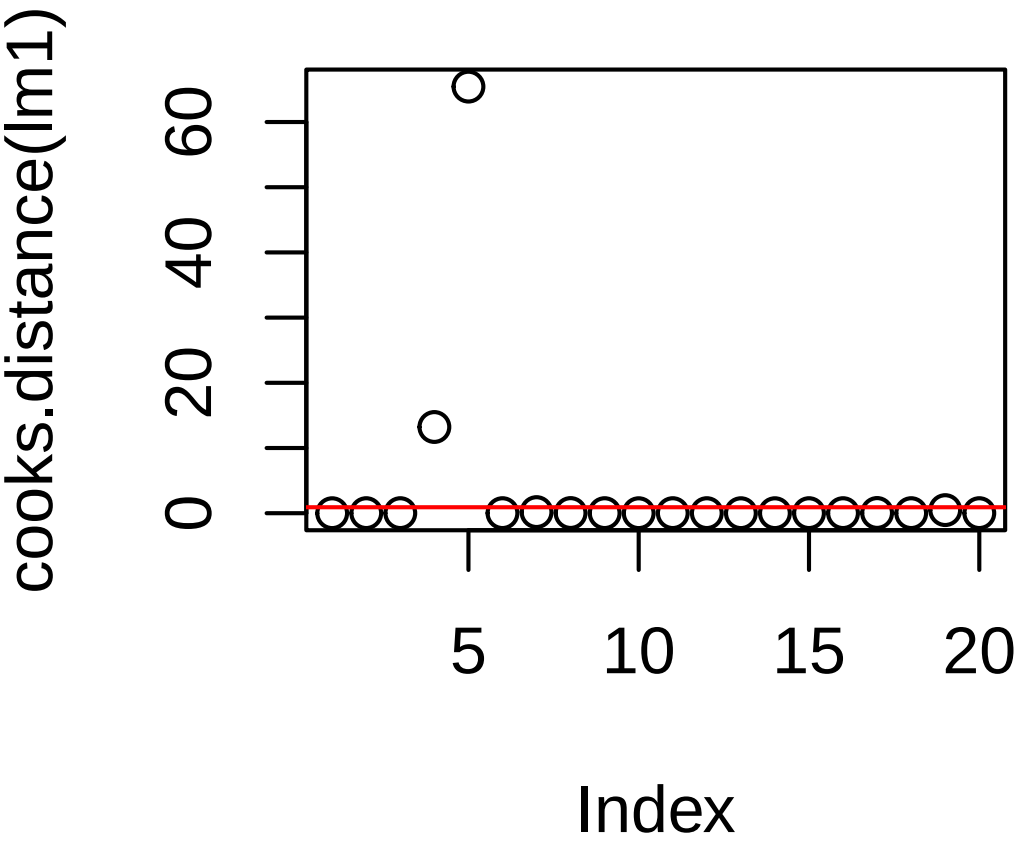
Leverage and Cook's Distance for NY River Data

```
data.frame(NYrivers$River,
           cooksD = round(cooks.distance(lm1),2),
           lev = round(hatvalues(lm1),2),
           rstu= round(rstudent(lm1),2))
```

```
##    NYrivers.River cooksD  lev  rstu
## 1           Olean   0.00 0.11 -0.14
## 2       Cassadaga   0.01 0.08 -0.63
## 3           Oatka   0.01 0.44  0.18
## 4       Neversink  13.22 0.90 -3.80
## 5      Hackensack  65.43 0.97 -4.84
## 6       Wappinger   0.01 0.05  0.89
## 7        Fishkill   0.16 0.10  3.91
## 8        Honeoye   0.02 0.16  0.77
## 9     Susquehanna   0.01 0.14 -0.43
## 10       Chenango   0.00 0.06  0.19
## 11    Tioughnioga   0.01 0.19  0.39
## 12    West Canada   0.01 0.12 -0.51
## 13    East Canada   0.00 0.17  0.12
## 14        Saranac   0.00 0.16  0.04
## 15        Ausable   0.01 0.20  0.32
## 16          Black   0.01 0.14  0.53
## 17      Schoharie   0.03 0.17 -0.83
## 18       Raquette   0.00 0.34  0.20
## 19    Oswegatchie   0.47 0.32 -2.68
## 20       Cohocton   0.01 0.21 -0.42
```

Indentifying Influential Points Using Cook's Distance

```
plot(cooks.distance(lm1))
abline(h=qf(0.5, 4+1, 20-4-1), col="red")
```



The 4th (Neversink) and 5th observation (Hackensack) is influential.

Identifying outlier using residuals

- If you want to test whether the ith data point is an outlier:

  we compute r_i^* where $r_i^* \sim t_{df=n-p-2}$: Outlier if p-value $< \alpha$, not outlier otherwise

- If you want to test whether there is any outlier in our data:

  1. compute t-statistics $t_1, \ldots, t_n$ (derived from r_i^*)
  2. From t-distribution with $df = n - p - 2$, we obtain p-values: $p_1, \ldots, p_n$
  3. If all p-values: $\min p_i \geq \frac{\alpha}{n} \Rightarrow$ no outliers

Added-Variable Plot

For MLR, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$ the LS estimate $\hat{\beta}_j$ would be identical to the slope for the SLR model computed as follows.

1. Regress Y on all other X_k's except X_j
2. Regress X_j on all other X_k's except X_j
3. Fit a SLR model using the residuals from Step 1 as the response and the residuals from Step 2 as the predictor.

We have explored this in an exercise

An added-variable plot is a plot with

- the residuals from Step 1 in the vertical axis
- the residuals from Step 2 in the horizontal axis

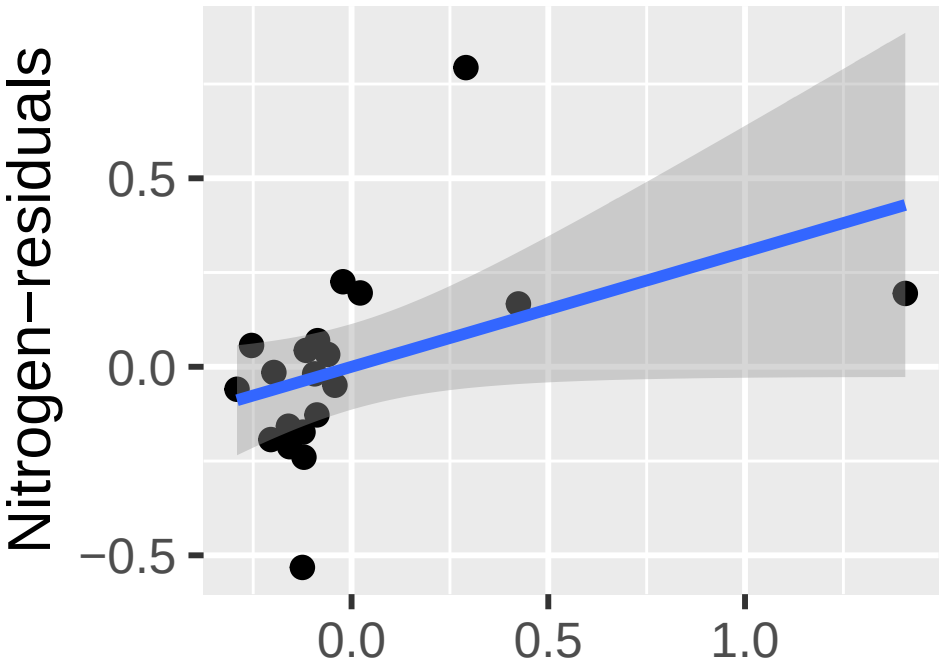
This plot helps to identify points that are highly influential in determining $\hat{\beta}_j$

Added-Variable Plot

The added-variable plot for ComIndl can be generated as following:

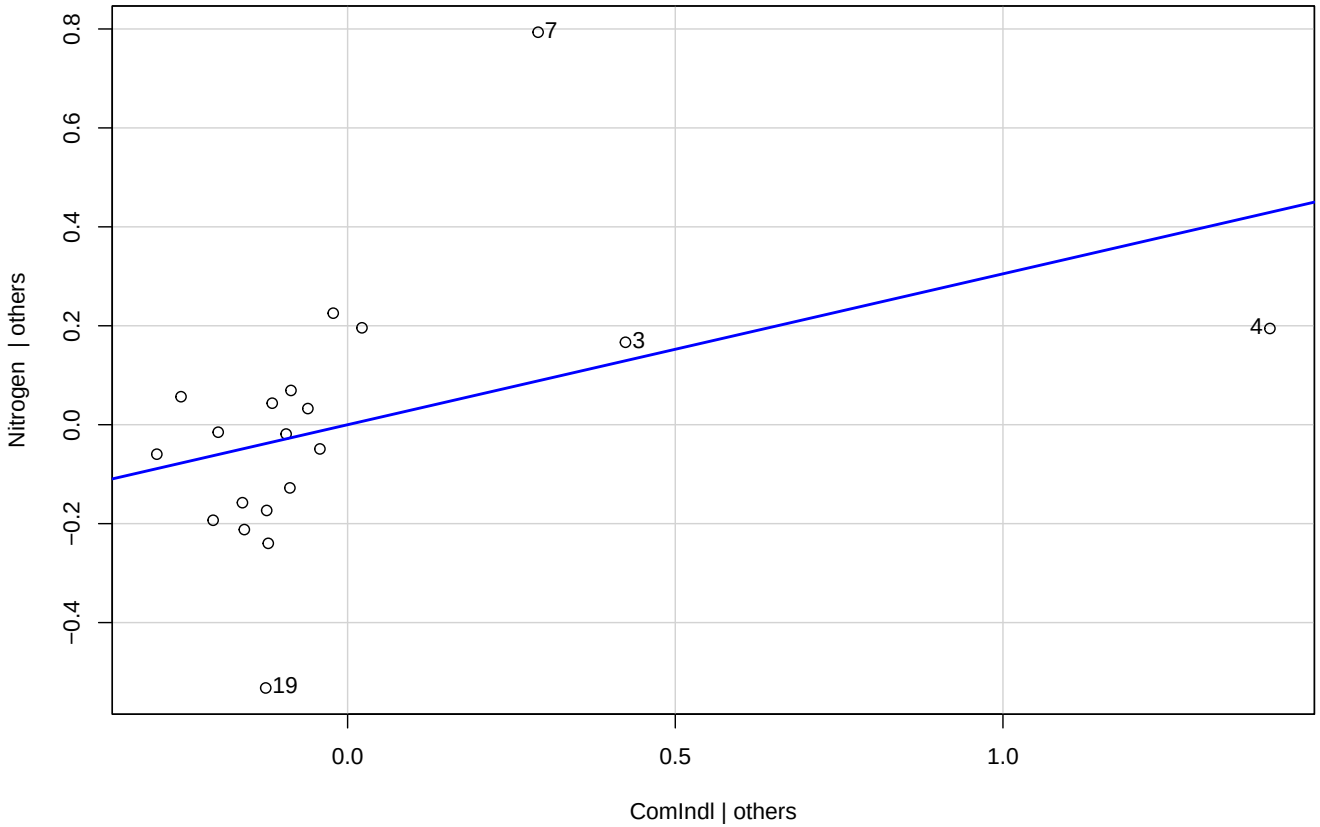
```
RN = lm(Nitrogen ~ Forest + Agr + Rsdntial, data=NYrivers)$res
RC = lm(ComIndl ~ Forest + Agr + Rsdntial, data=NYrivers)$res

ggplot(data.frame(RN, RC), aes(x=RC, y=RN)) + geom_point() +
  geom_smooth(method='lm') +
  labs(x="ComIndl-residuals", y="Nitrogen-residuals")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



avPlots()

```
library(car)
avPlots(lm1, "ComIndl")
```



Added-Variable Plot for All Variables

```
avPlots(lm1, layout=c(2,2))
```