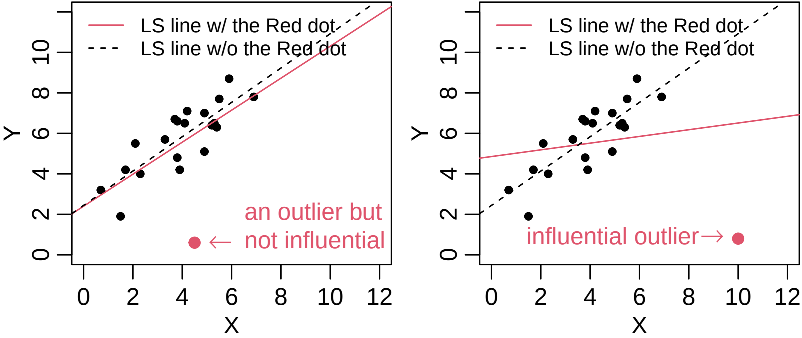
# STAT 224 Lecture 12
# Chapter 4 Model Diagnostics, Part 3
# Leverage, Influence, and Outliers

Yibi Huang

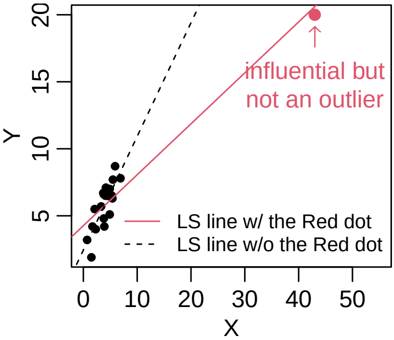
# Influential Points and Outliers

# Outliers vs. Influential Points

- An *outlier* is a point that the model fails to explain. It has a large residual.
- An *influential point* has an unduly large effect on the model. The fitted model changes drastically when it is included.
- A point can be influential, an outlier, or both.
  See the examples on the next page
- Influential points are not necessarily outliers




- For SLR, influential points and outliers can be identified by inspecting scatterplots
- For MLR, identification of influential points is more difficult



# Example – New York Rivers

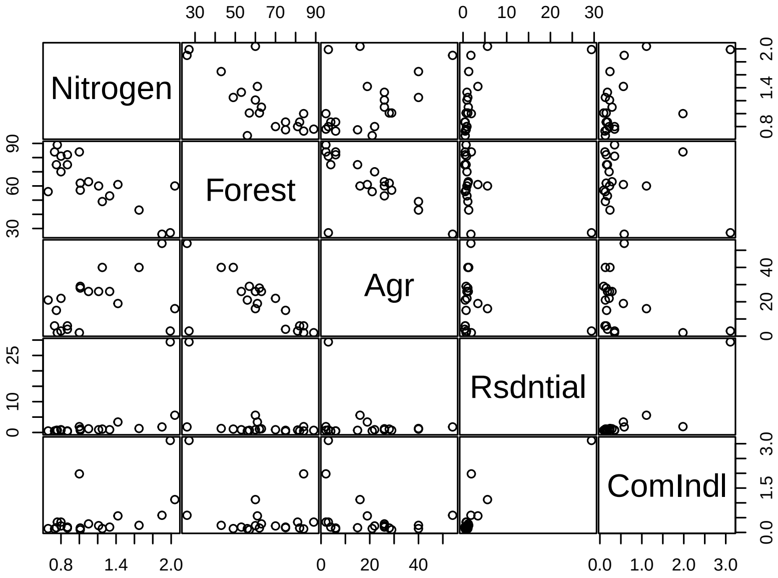
Data on Water Pollution in New York Rivers (Table 1.8, 1.9 on p.10 of textbook), which can be download at

http://www.stat.uchicago.edu/~yibi/s224/data/P010.txt

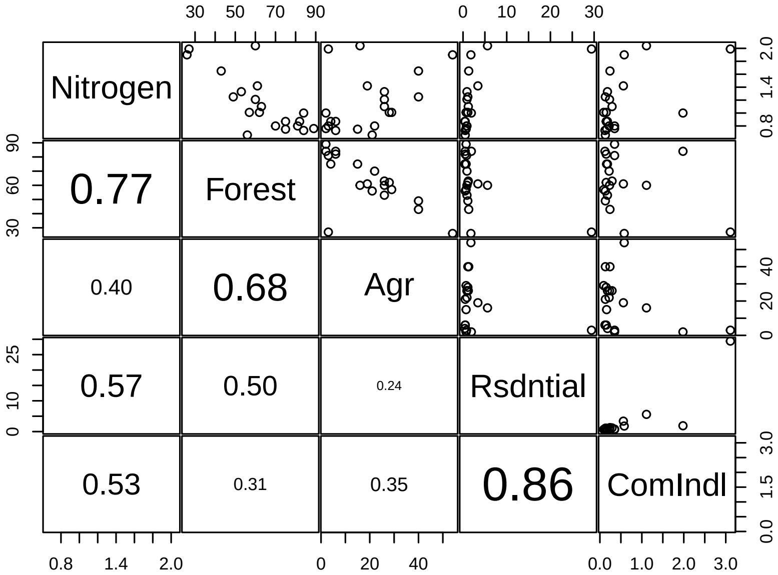
- `Nitrogen` = Mean nitrogren concentration (mg/liter) measured at regular intervals (Response)
- `Agr` = % of land currently in Agricultural use
- `Forest` = % of Forest land
- `Rsdntial` = % of land in Residential use
- `ComIndl` = % of lane in Commercial or Industrial use

```
NYrivers = read.table("P010.txt", h = T, sep="\t")
```

```
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

```r
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)
```

```r
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"))
```

On the next page, observe that the coefficient of `Rsdntial` is

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

```
summary(lm1)$coef       # all data
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.722214    1.23408  1.3955  0.18317
Forest      -0.012968    0.01393 -0.9308  0.36668
Agr          0.005809    0.01503  0.3864  0.70463
Rsdntial    -0.007227    0.03383 -0.2136  0.83372
ComIndl      0.305028    0.16382  1.8620  0.08231
summary(lm1noH)$coef    # w/o Hackensack
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.626014   0.781091  2.0817 0.056199
Forest      -0.012760   0.008815 -1.4476 0.169756
Agr          0.002352   0.009539  0.2466 0.808807
Rsdntial     0.181161   0.044390  4.0811 0.001123
ComIndl      0.075618   0.113957  0.6636 0.517750
summary(lm1noN)$coef    # w/o Neversink
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.099471    0.91164  1.2060 0.2477883
Forest      -0.007589    0.01022 -0.7424 0.4700975
Agr          0.010137    0.01098  0.9229 0.3717055
Rsdntial    -0.123793    0.03934 -3.1470 0.0071343
ComIndl      1.528956    0.34372  4.4483 0.0005512
```

# Hat Matrix, Leverages, High Leverage Points

# Hat Matrix (Review)

Recall in `L10.pdf`, for the MLR model,

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + \varepsilon_j.$$

we define the *hat matrix* $\mathbf{H}$ to be $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ where $\mathbf{X}$ is the *model matrix*

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

# Leverages $h_{ii}$

Recall in L10.pdf, we showed that the predicted Value $\widehat{\mathbf{Y}}$ of $\mathbf{Y}$ is

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}Y$$

in other words,

$$\overbrace{\begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{pmatrix}}^{\widehat{\mathbf{Y}}} = \overbrace{\begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix}}^{\mathbf{H}} \overbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}^{\mathbf{Y}}$$

$\widehat{\mathbf{Y}} = \mathbf{HY}$ means every predicted value $\widehat{y}_i$ is a linear combination of $y_1, \ldots, y_n$

$$\widehat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n,$$

and $h_{ij}$ is the $(i, j)$th element of the matrix $\mathbf{H}$.

# Leverages $h_{ii}$ (2)

$$\widehat{y_i} = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n,$$

- $h_{ij}$ = the contribution of $y_j$ in predicting $\hat{y}_i$.
- $h_{ii}$ = the contribution of $y_i$ in predicting itself $\hat{y}_i$, is called the *leverage* of $i$th observation, $i = 1, 2, \ldots, n$.
- Hence, an **influential point must have a high leverage $h_{ii}$**

# Leverages $h_{ii}$ (2)

$$\widehat{y_i} = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n,$$

- $h_{ij}$ = the contribution of $y_j$ in predicting $\hat{y}_i$.
- $h_{ii}$ = the contribution of $y_i$ in predicting itself $\hat{y}_i$, is called the *leverage* of $i$th observation, $i = 1, 2, \ldots, n$.
- Hence, an **influential point must have a high leverage $h_{ii}$**
- For SLR

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$

  So, a high-leverage point in SLR is an outlier of the $X$-variable. The further $x_i$ is away from $\bar{x}$, the higher leverage it has

# Leverages $h_{ii}$ (2)

$$\widehat{y_i} = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n,$$

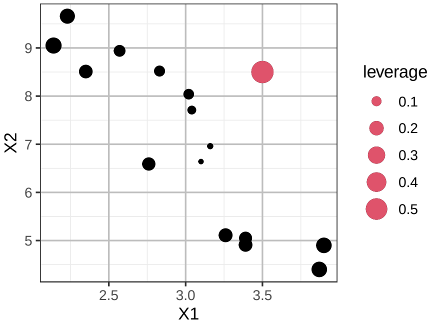
- $h_{ij}$ = the contribution of $y_j$ in predicting $\hat{y}_i$.
- $h_{ii}$ = the contribution of $y_i$ in predicting itself $\hat{y}_i$, is called the *leverage* of $i$th observation, $i = 1, 2, \ldots, n$.
- Hence, an **influential point must have a high leverage $h_{ii}$**
- For SLR

  $$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$

  So, a high-leverage point in SLR is an outlier of the $X$-variable. The further $x_i$ is away from $\bar{x}$, the higher leverage it has
- $h_{ij}$ and $h_{ii}$ are completely determined by the predictors $\mathbf{X}$ since $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

# High Leverage Points Are Outliers in $X$-Space

```
hamilton = read.table("P103.txt", h = T)
hamilton = rbind(c(11,3.5,8.5), hamilton)  # adding a new obs
lmHamilton = lm(Y ~ X1 + X2, data=hamilton)
leverage = hatvalues(lmHamilton)
library(ggplot2)
ggplot(hamilton, aes(x = X1, y = X2, size=leverage)) +
  geom_point() + geom_point(aes(x=X1[1],y=X2[1]), col=2)
```



# High Leverage Points

- Recall in `L10.pdf`, we have mentioned that
  - leverages lie between $0$ and $1$, and
  - $\sum h_{ii} = p + 1$,
    hence $h_{ii}$'s have an average value of $(p + 1)/n$.
- Points with $h_{ii} > 2(p + 1)/n$ are considered to have high leverage.
  These points should be flagged and checked to see if they are unduly influential.
  - For the NY Rivers data, $n = 20$, $p = 4$, points w/ $h_{ii} > \frac{2(p+1)}{n} = \frac{2(4+1)}{20} = 0.5$ are high leverage points
- Finding leverage In R: hatvalues(model)

```r
data.frame(NYrivers$River, lev = round(hatvalues(lm1),2),
res = round(lm1$res,2), rstu= round(rstudent(lm1),2))
   NYrivers.River  lev   res  rstu
1           Olean 0.09 -0.12 -0.62
2       Cassadaga 0.18 -0.03 -0.15
3           Oatka 0.63  0.05  0.41
4       Neversink 0.56 -0.19 -1.46
5      Hackensack 0.89 -0.13 -2.28
6       Wappinger 0.20 -0.04 -0.21
7        Fishkill 0.27  0.42  3.14
8        Honeoye  0.16  0.19  1.05
9     Susquehanna 0.17 -0.15 -0.79
10       Chenango 0.07  0.06  0.30
11    Tioughnioga 0.11  0.17  0.90
12    West Canada 0.10 -0.12 -0.63
13    East Canada 0.19  0.10  0.52
14        Saranac 0.14 -0.04 -0.19
15        Ausable 0.18  0.00  0.02
16          Black 0.14  0.20  1.10
17      Schoharie 0.09 -0.25 -1.38
18       Raquette 0.33  0.21  1.35
```

# Relationship between Residual and Leverage

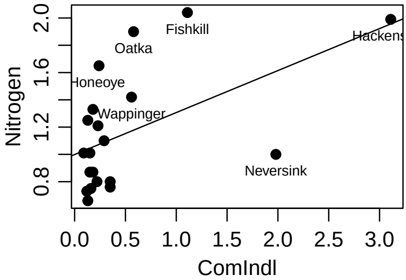
- The raw residuals, $e_i$, and the leverage, $h_{ii}$, satisfy

$$h_{ii} + \frac{e_i^2}{\text{SSE}} \leq 1.$$

- Therefore points with high leverage tend to have small residuals.
- We must examine both residuals and leverages to identify possible model violations.

# Masking and Swamping

- Masking occurs when we miss outliers (false negative).
    - This can occur when an outlier is hidden by other outliers,
- Swamping occurs when we incorrectly label a point as an outlier (false positive).
    - This can occur since large outliers tend to pull the fitted line toward them, possibly away from other points.
- We need other methods of measuring influence to get around these problems.



# Measures of Influence

# Measures of Influence

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

# 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 $i = 1, 2, \ldots, n$.

- Equivalently, Cook's distance can be shown to be

$$C_i = \frac{r_i^2}{p+1} \times \underbrace{\frac{h_{ii}}{1-h_{ii}}}_{\text{potential}}$$

where $r_i$ = $i$th internally studentized residual

- 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 $p + 1$ and $n - p - 1$ degrees of freedom, i.e.,

```
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] 0.9107
```

- A graph of $C_i$ vs. $i$ helps us to see influential points.

# R Commands for Diagnostics

- fitted values

```
model$fit
```

- raw residuals

```
model$res
```

- internally studentized residuals

```
rstandard(model)
```

- externally studentized residuals

```
rstudent(model)
```

# R Commands for Diagnostics

- leverage

```
hatvalues(model)
```

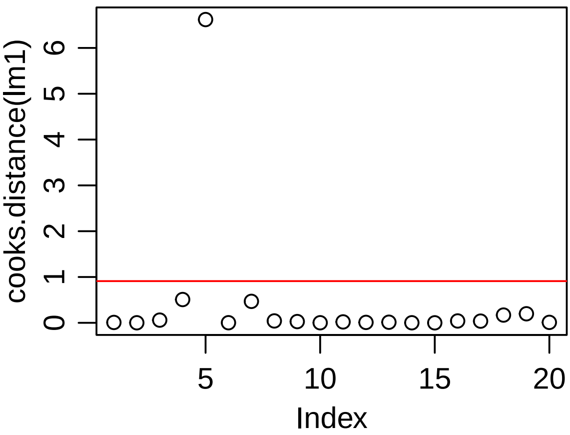
- Cook's distance

```
cooks.distance(model)
```

# Leverage & 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   0.01 0.09 -0.62
2       Cassadaga   0.00 0.18 -0.15
3           Oatka   0.06 0.63  0.41
4       Neversink   0.51 0.56 -1.46
5      Hackensack   6.62 0.89 -2.28
6       Wappinger   0.00 0.20 -0.21
7        Fishkill   0.47 0.27  3.14
8        Honeoye    0.04 0.16  1.05
9     Susquehanna   0.03 0.17 -0.79
10       Chenango   0.00 0.07  0.30
11    Tioughnioga   0.02 0.11  0.90
12    West Canada   0.01 0.10 -0.63
13    East Canada   0.01 0.19  0.52
14        Saranac   0.00 0.14 -0.19
```

```r
par(mai=c(.55,.55,.02,.02),mgp=c(1.8,.7,0))
plot(cooks.distance(lm1))
abline(h=qf(0.5, 4+1, 20-4-1), col="red")
```




The 5th observation (Hackensack) is influential.

# Added-Variable Plot

# Added-Variable Plot

In slides `L02.pdf`, we said the LS estimate $\widehat{\beta}_j$ for $\beta_j$ in the MLR model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

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.

# Added-Variable Plot

In slides `L02.pdf`, we said the LS estimate $\widehat{\beta}_j$ for $\beta_j$ in the MLR model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

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.

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 $\hat{\beta}_j$.

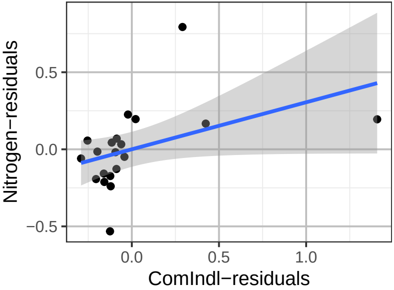
# Added-Variable Plot

If we fit the model below,

```
lm1 = lm(Nitrogen ~ Forest + Agr + Rsdntial + ComIndl, data=NYrivers)
```

the added-variable plot for `ComIndl` is

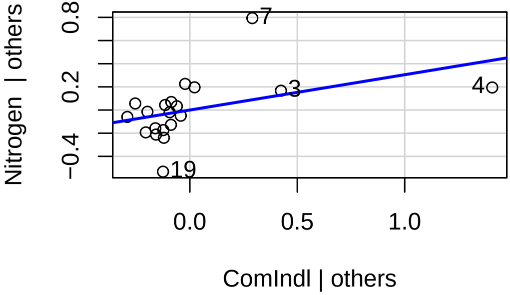
```
RN = lm(Nitrogen ~ Forest + Agr + Rsdntial, data=NYrivers)$res
RC = lm(ComIndl ~ Forest + Agr + Rsdntial, data=NYrivers)$res
ggplot(data.frame(RN, RC), aes(x=RC, y=RN)) + geom_point() +
  geom_smooth(method='lm') + labs(x="ComIndl-residuals", y="Nitrogen-re
```



# Making Added-Variable Plot Using `avPlots()` in the `car` Library

The avPlots() function in the car library can produce added-variable plots automatically

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



# Added-Variable Plot for All Variables

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