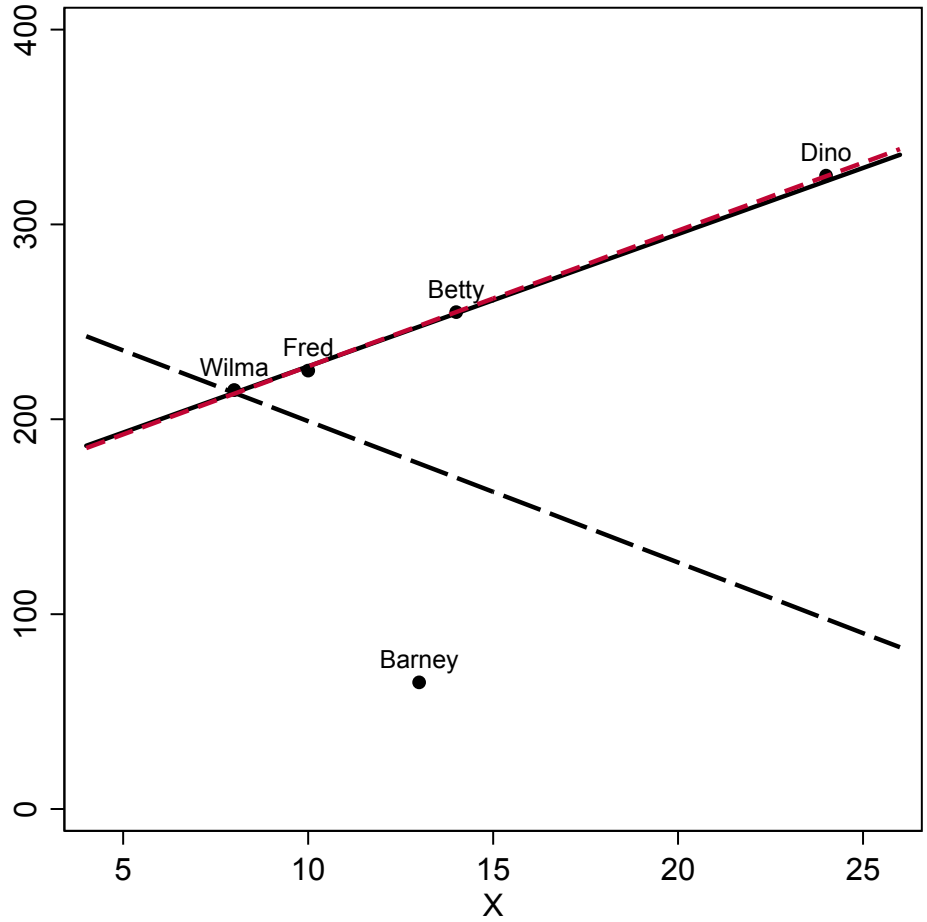
GSERM 2017

Regression III

Outliers / Influence and Robust Inference

June 21, 2017 (morning session)

Discrepancy, Leverage, and Influence



Note: Solid line is the regression fit for Wilma, Fred, and Betty only.
Long-dashed line is the regression for Wilma, Fred, Betty, and Barney.
Short-dashed (red) line is the regression for Wilma, Fred, Betty and Dino.

Discrepancy, Leverage, and Influence

$$\text{Influence} = \text{Leverage} \times \text{Discrepancy}$$

Leverage

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= \mathbf{H}\mathbf{Y} \end{aligned}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

$$h_i = \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i'$$

Residuals

Variation:

$$\widehat{\text{Var}(\hat{u}_i)} = \hat{\sigma}^2[1 - \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i'] \tag{1}$$

$$\begin{aligned} \widehat{\text{s.e.}(\hat{u}_i)} &= \hat{\sigma}\sqrt{[1 - \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i']} \\ &= \hat{\sigma}\sqrt{1 - h_i} \end{aligned} \tag{2}$$

"Standardized":

$$\tilde{u}_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_i}} \tag{3}$$

"Studentized": define

$$\begin{aligned} \hat{\sigma}^2_{-i} &= \text{Variance for the } N-1 \text{ observations} \neq i \\ &= \frac{\hat{\sigma}^2(N-K)}{N-K-1} - \frac{\hat{u}_i^2}{(N-K-1)(1-h_i)}. \end{aligned} \tag{4}$$

Then:

$$\hat{u}'_i = \frac{\hat{u}_i}{\hat{\sigma}_{-i}\sqrt{1-h_i}} \tag{5}$$

Influence

"DFBETA":

$$D_{ki} = \hat{\beta}_k - \hat{\beta}_{k(-i)} \tag{6}$$

"DFBETAS" (the "S" is for "standardized):

$$D^*_{ki} = \frac{D_{ki}}{\widehat{\text{s.e.}(\hat{\beta}_{k(-i)})}} \tag{7}$$

Cook's D:

$$\begin{aligned} D_i &= \frac{\tilde{u}_i^2}{K} \times \frac{h_i}{1-h_i} \\ &= \frac{h_i \hat{u}_i^2}{K\hat{\sigma}^2(1-h_i)^2} \end{aligned} \tag{8}$$

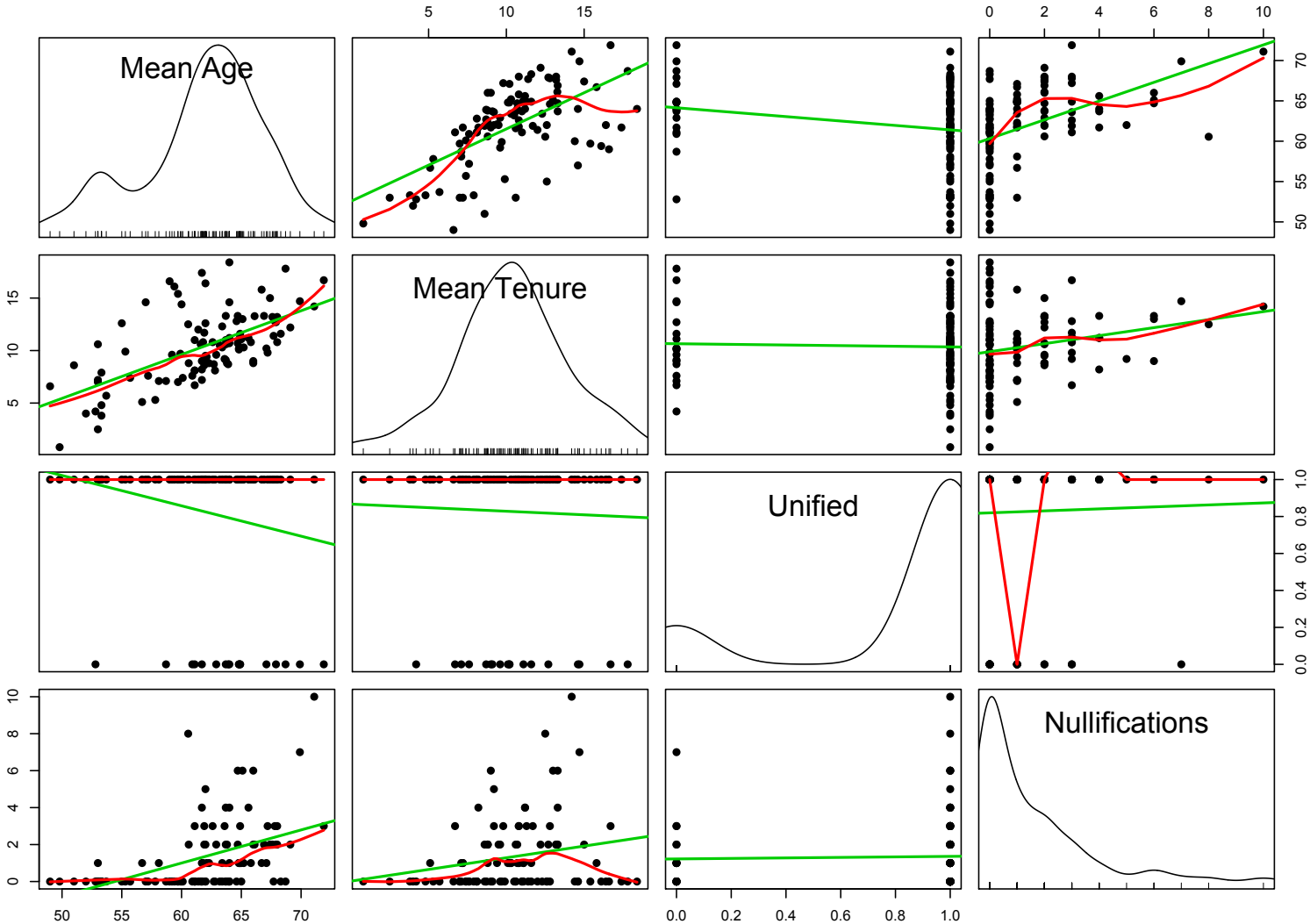
```
> # No Barney OR Dino...
> summary(lm(Y~X,data=subset(flintstones,name!="Dino" & name!="Barney")))

Residuals:
     2      4      5
 0.714 -2.143  1.429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  159.286      6.776    23.5    0.027 *
X              6.786      0.619    11.0    0.058 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.67 on 1 degrees of freedom
Multiple R-squared: 0.992,Adjusted R-squared: 0.984
F-statistic:  120 on 1 and 1 DF,  p-value: 0.0579
```

```
> # No Barney (Dino included...)
> summary(lm(Y~X,data=subset(flintstones,name!="Barney")))

Residuals:
        2         3         4         5
-8.88e-16  2.63e-01 -2.11e+00  1.84e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  157.368      2.465    63.8  0.00025 ***
X              6.974      0.161    43.3  0.00053 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.99 on 2 degrees of freedom
Multiple R-squared: 0.999,Adjusted R-squared: 0.998
F-statistic: 1.87e+03 on 1 and 2 DF,  p-value: 0.000534
```

"COVRATIO":

$$\text{COVRATIO}_i = \left[(1 - h_i) \left(\frac{N - K - 1 + \hat{u}_i'^2}{N - K} \right)^K \right]^{-1} \tag{9}$$

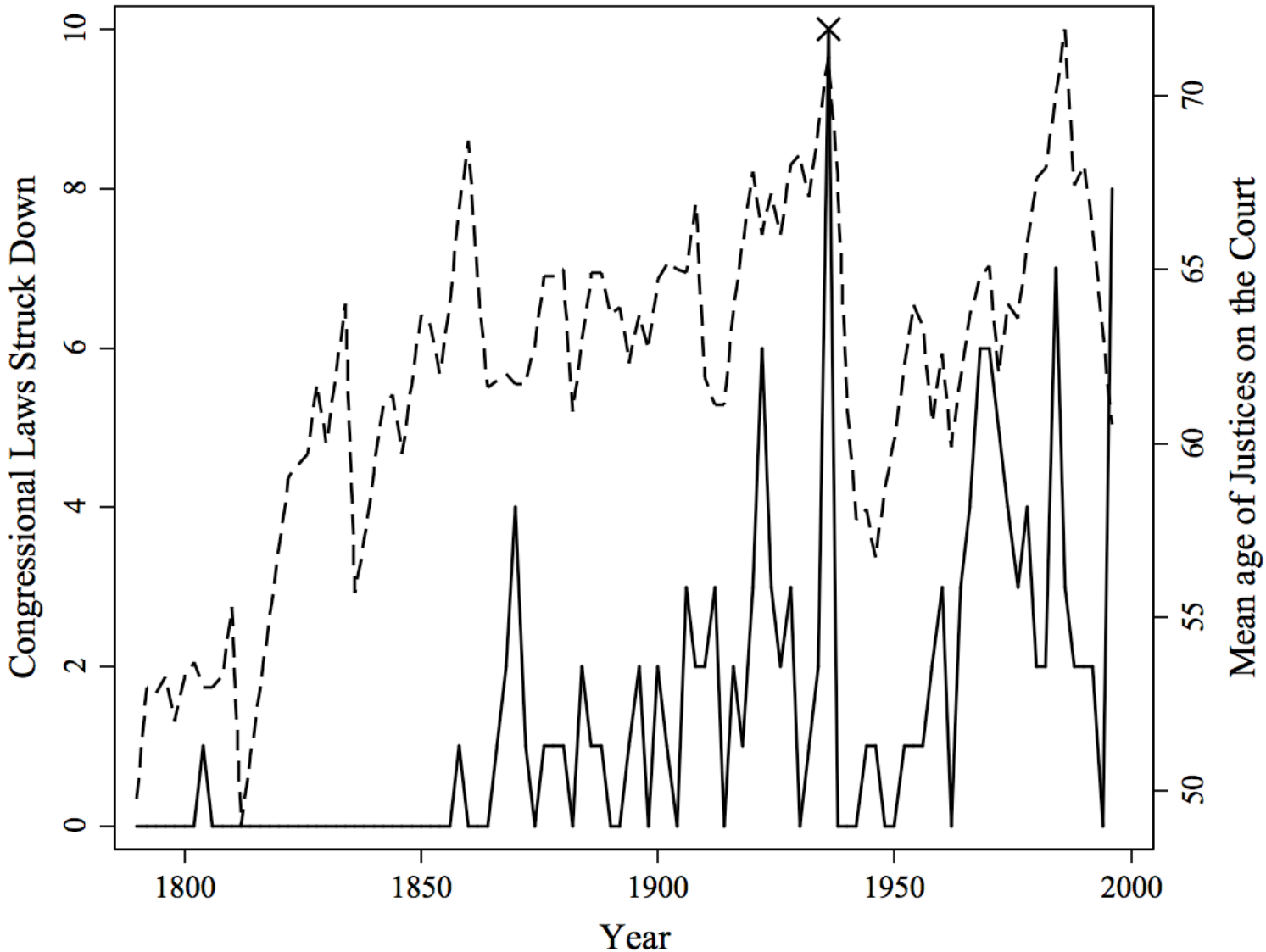
Example: Federal Judicial Review, 1789-1996



Variance

```
> Fit<-lm(nulls~age+tenure+unified)
> summary(Fit)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7857 -1.0773 -0.3634  0.4238  6.9694

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.10340    2.54324  -4.759 6.57e-06 ***
age           0.21886    0.04484   4.881 4.01e-06 ***
tenure       -0.06692    0.06427  -1.041    0.300
unified       0.71760    0.45844   1.565    0.121
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.715 on 100 degrees of freedom
Multiple R-squared: 0.2324,Adjusted R-squared: 0.2093
F-statistic: 10.09 on 3 and 100 DF,  p-value: 7.241e-06
```

Federal Judicial Review and Mean SCOTUS Age



Residuals, etc.

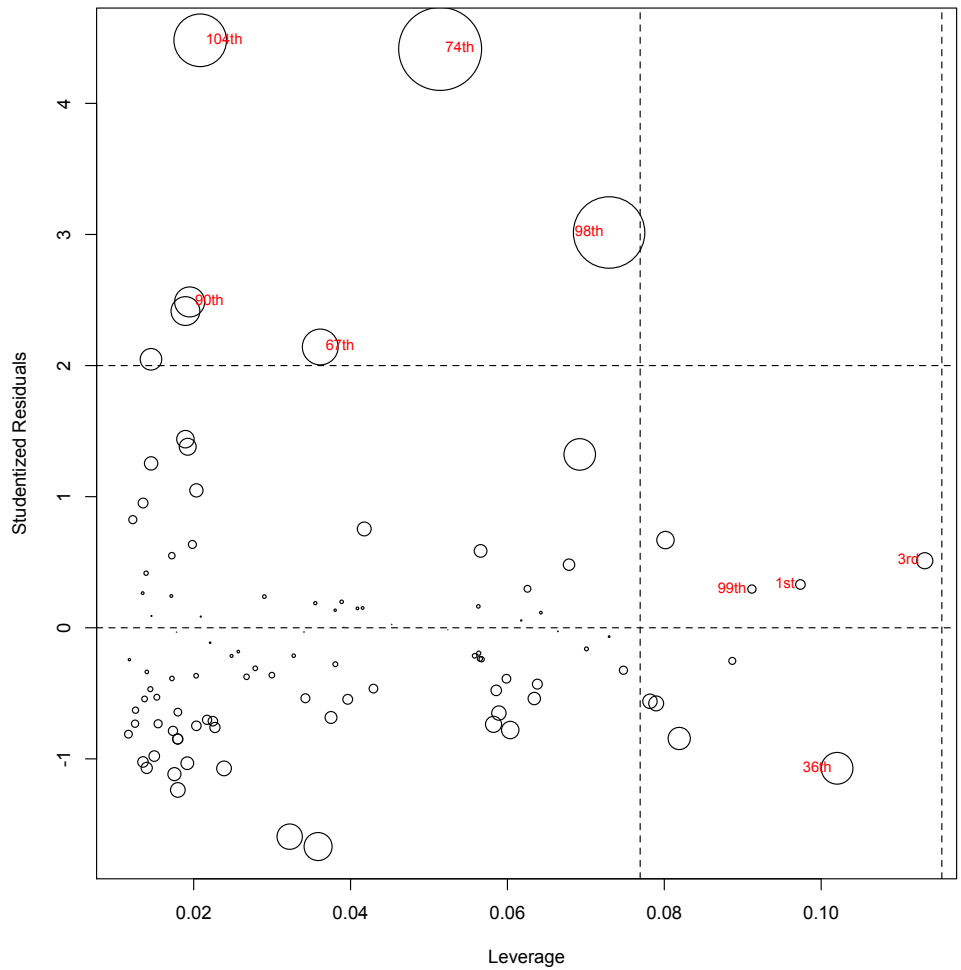
```
> FitResid<-(nulls - predict(Fit)) # residuals
> FitStandard<-rstandard(Fit) # standardized residuals
> FitStudent<-rstudent(Fit) # studentized residuals
> FitCooksD<-cooks.distance(Fit) # Cook's D
> FitDFBeta<-dfbeta(Fit) # DFBeta
> FitDFBetaS<-dfbetas(Fit) # DFBetaS
> FitCOVRATIO<-covratio(Fit) # COVRATIOs
```

Studentized Residuals

```
> FitStudent[74]
      74
4.415151

> Congress74<-rep(0,length=104)
> Congress74[74]<-1

> summary(lm(nulls~age+tenure+unified+Congress74))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.17290    2.37692  -4.280 4.33e-05 ***
age           0.18820    0.04177   4.505 1.82e-05 ***
tenure       -0.06356    0.05905  -1.076    0.284
unified       0.55159    0.42282   1.305    0.195
Congress74    7.14278    1.61779   4.415 2.58e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.576 on 99 degrees of freedom
Multiple R-squared: 0.3586,Adjusted R-squared: 0.3327
F-statistic: 13.84 on 4 and 99 DF,  p-value: 5.304e-09
```

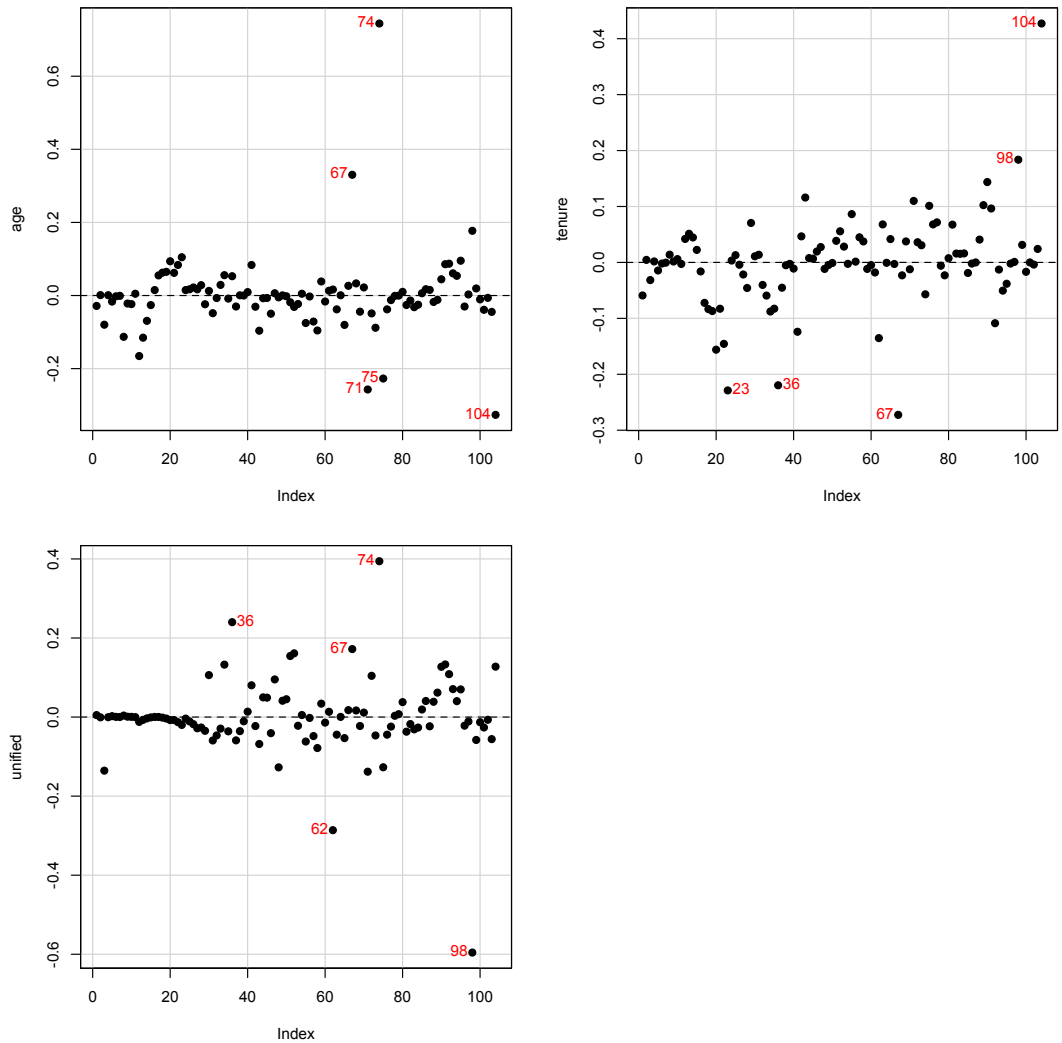
"Bubble Plot"

```
> influencePlot(Fit,id.n=4,labels=Congress,id.cex=0.8,
  id.col="red",xlab="Leverage")
```



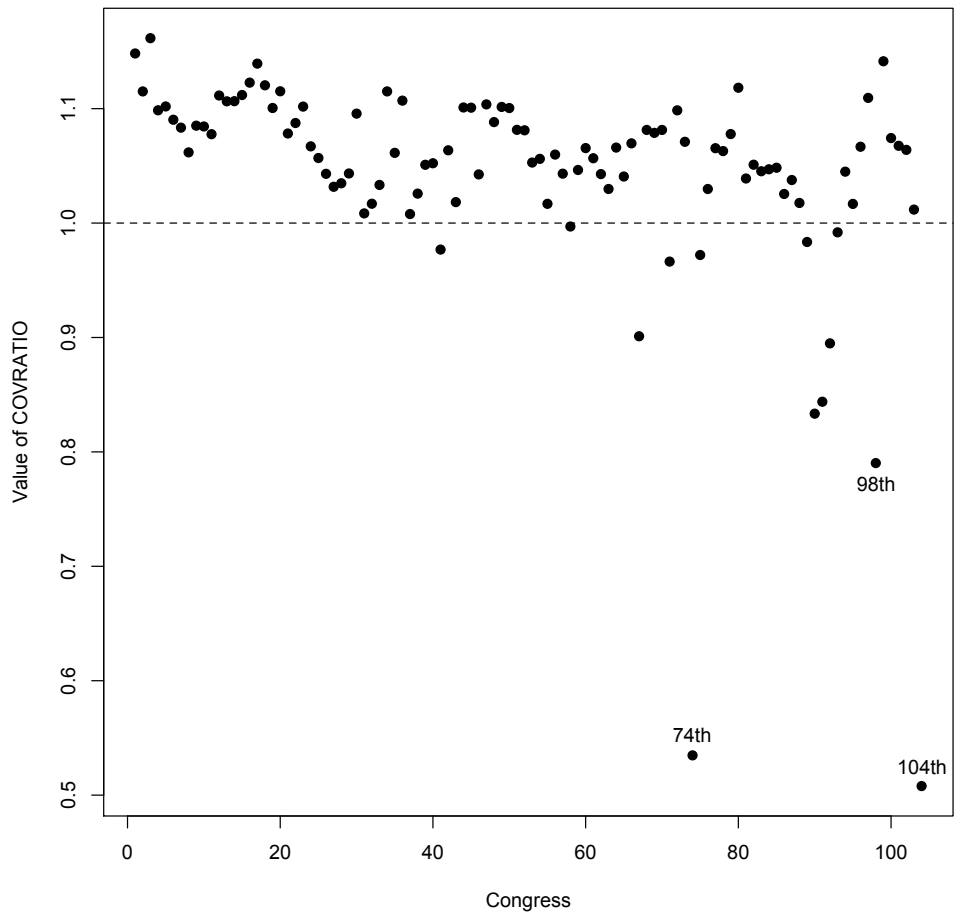
DFBETAS

```
> dfbetasPlots(Fit,id.n=5,id.col="red",main="",pch=19)
```



COVRATIO Plot

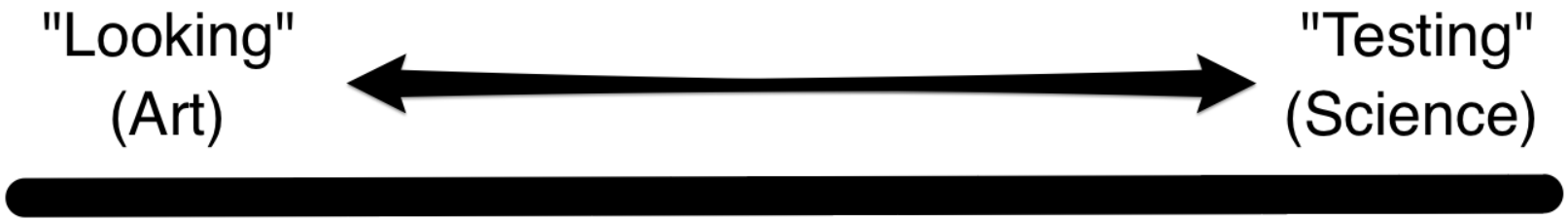
```
> plot(FitCOVRATIO~congress,pch=19,xlab="Congress",ylab="Value of COVRATIO")
> abline(h=1,lty=2)
```



Sensitivity Analyses: Omitting Outliers

```
> Outlier<-rep(0,104)
> Outlier[74]<-1
> Outlier[98]<-1
> Outlier[104]<-1
> DahlSmall<-Dahl[which (Outlier==0),]

> summary(lm(nulls~age+tenure+unified,data=DahlSmall))

Call:
lm(formula = nulls ~ age + tenure + unified, data = DahlSmall)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.38536    1.99470  -5.206 1.08e-06 ***
age           0.19302    0.03512   5.496 3.13e-07 ***
tenure       -0.10069    0.04974  -2.024   0.0457 *
unified       0.76645    0.36069   2.125   0.0361 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.319 on 97 degrees of freedom
Multiple R-squared: 0.2578,Adjusted R-squared: 0.2349
F-statistic: 11.23 on 3 and 97 DF,  p-value: 2.167e-06
```

Thinking About Diagnostics



Observational Data
Complex Data Structure
Informative Missingness
Complex / Uncertain Causality

Experimental Data
Simple Data Structure
No / Uninformative Missingness
Simple / Clear Causality

One Approach

Pena, E.A. and E.H. Slate. 2006. "Global Validation of Linear Model Assumptions." *J. American Statistical Association* 101(473):341-354.

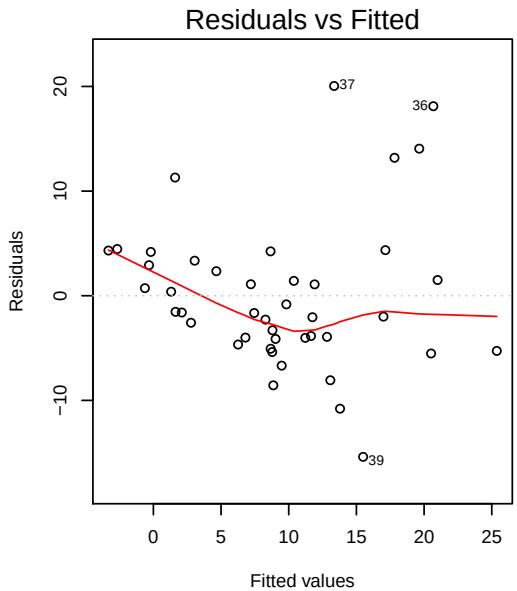
Tests for:

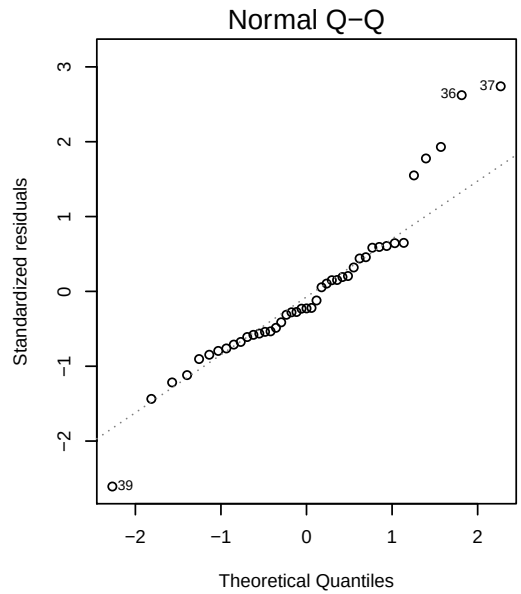
- Normality in $\hat{u}$s (via skewness & kurtosis tests)
- "Link function" (linearity / additivity)
- Constant variance and uncorrelatedness in $\hat{u}$s ("heteroskedasticity" test)

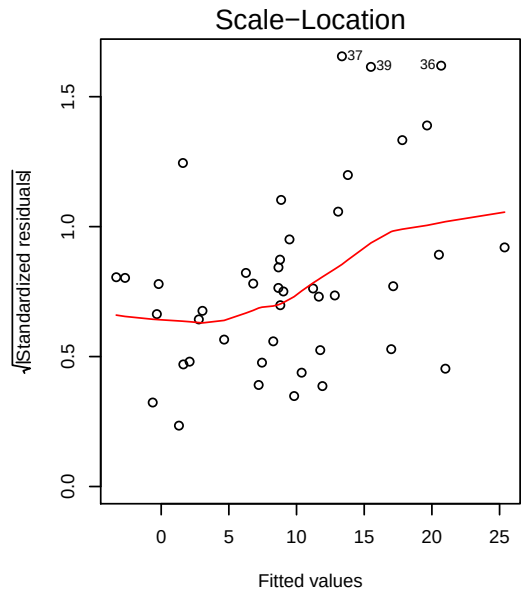
In Action

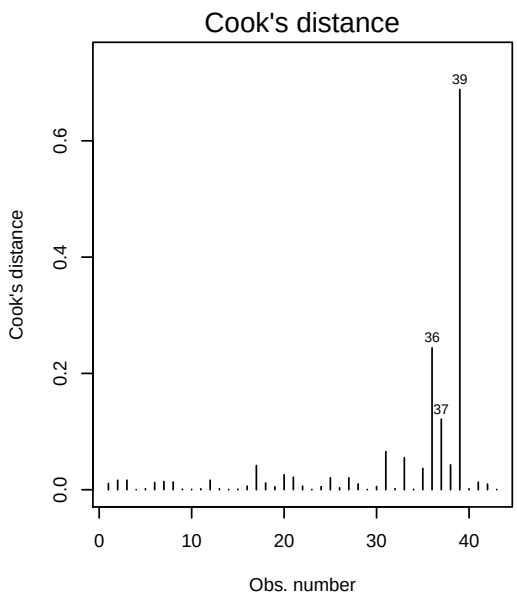
```
> Fit <- with(Africa, lm(adrate~gdppppd+muslperc+subsaharan+healthexp+
                literacy+internalwar))

> library(gvlma)
> Nope <- gvlma(Fit)
> display.gvlmatests(Nope)

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = Fit)

                     Value   p-value                   Decision
Global Stat         21.442 0.0002587 Assumptions NOT satisfied!
Skewness             5.720 0.0167698 Assumptions NOT satisfied!
Kurtosis             2.345 0.1256876    Assumptions acceptable.
Link Function        5.892 0.0152059 Assumptions NOT satisfied!
Heteroscedasticity   7.485 0.0062227 Assumptions NOT satisfied!
```

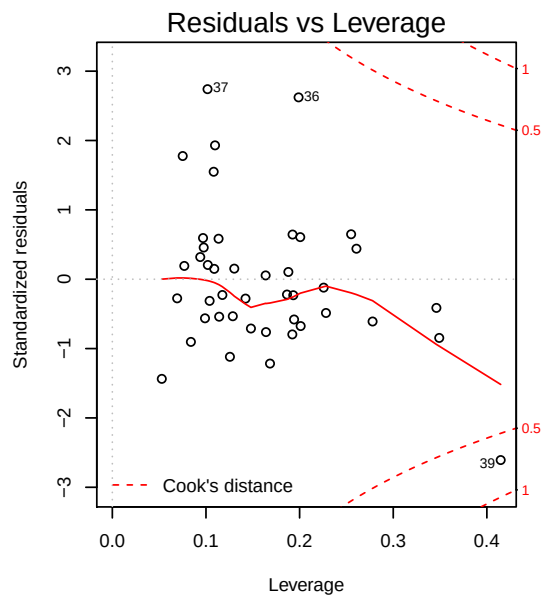
Another Approach: `plot(fit)`

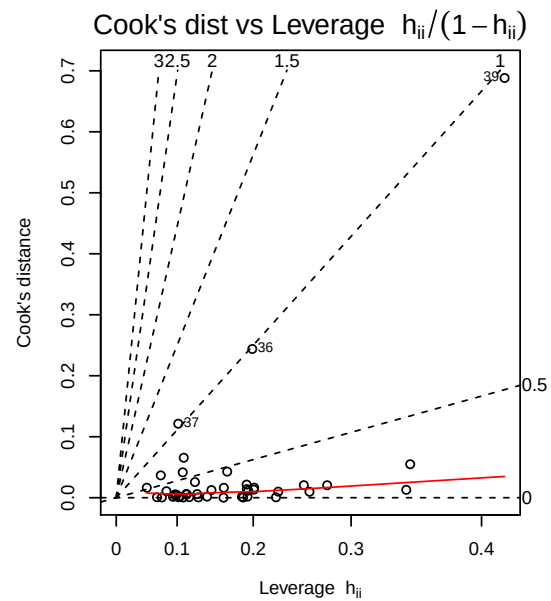




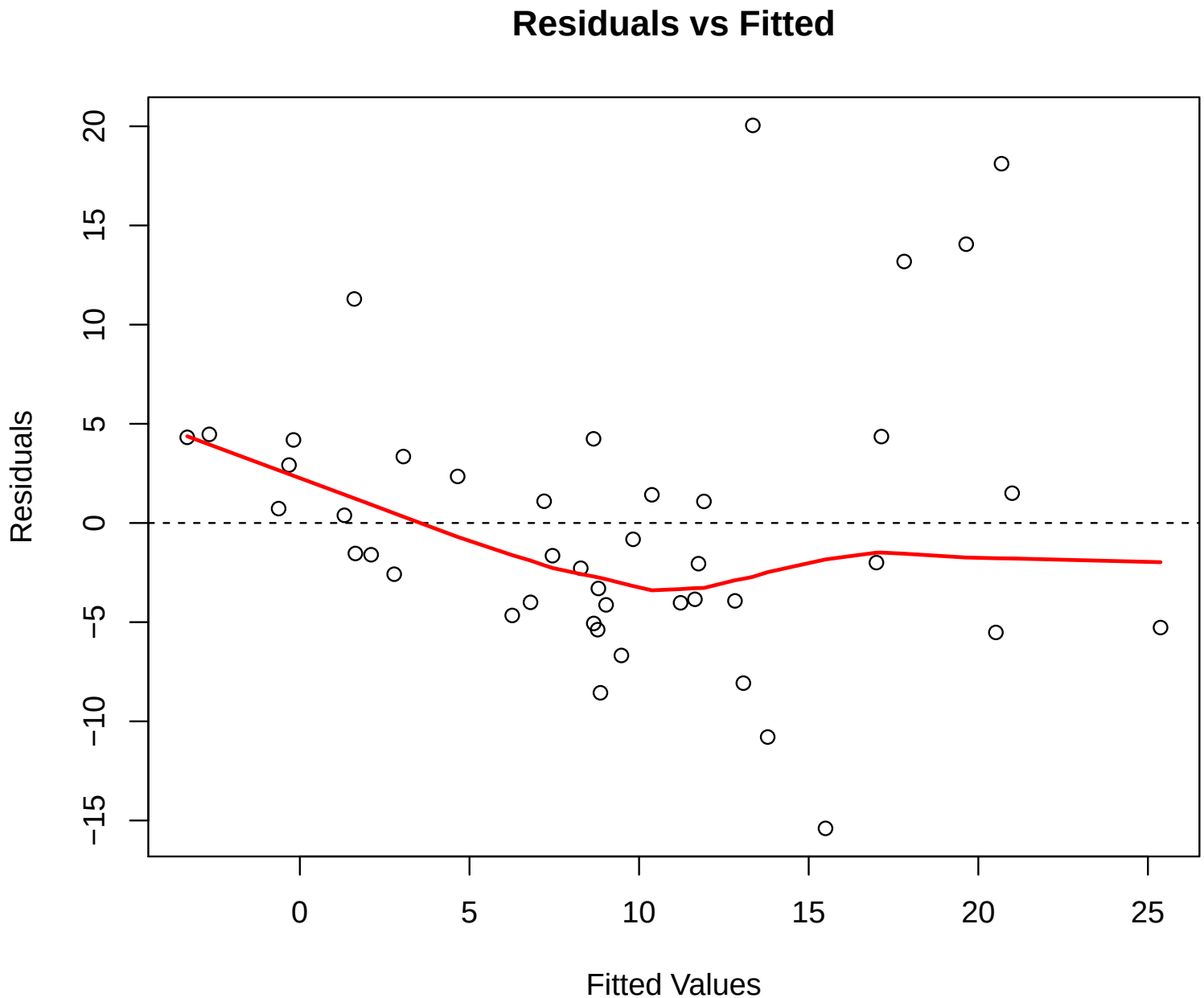




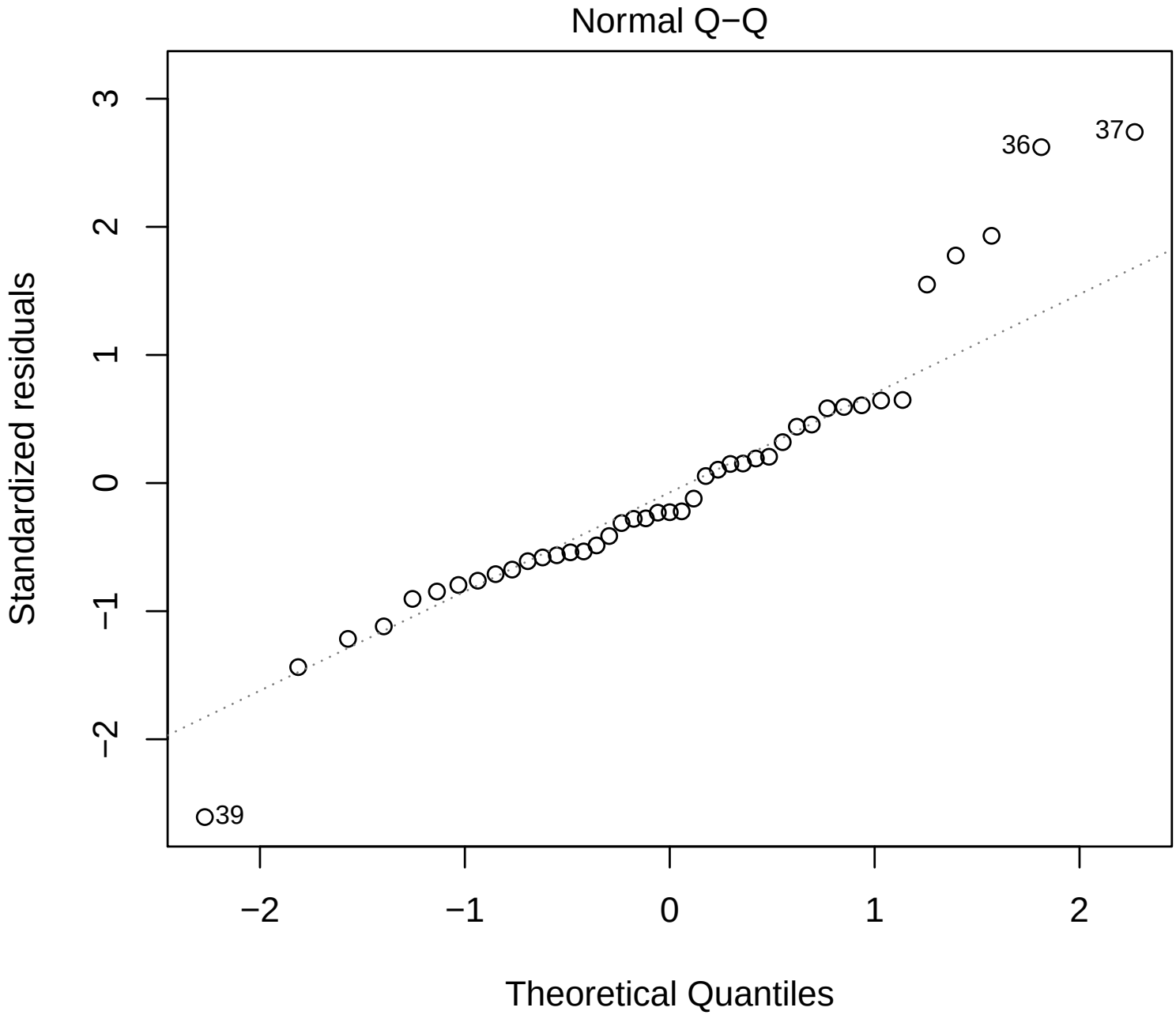




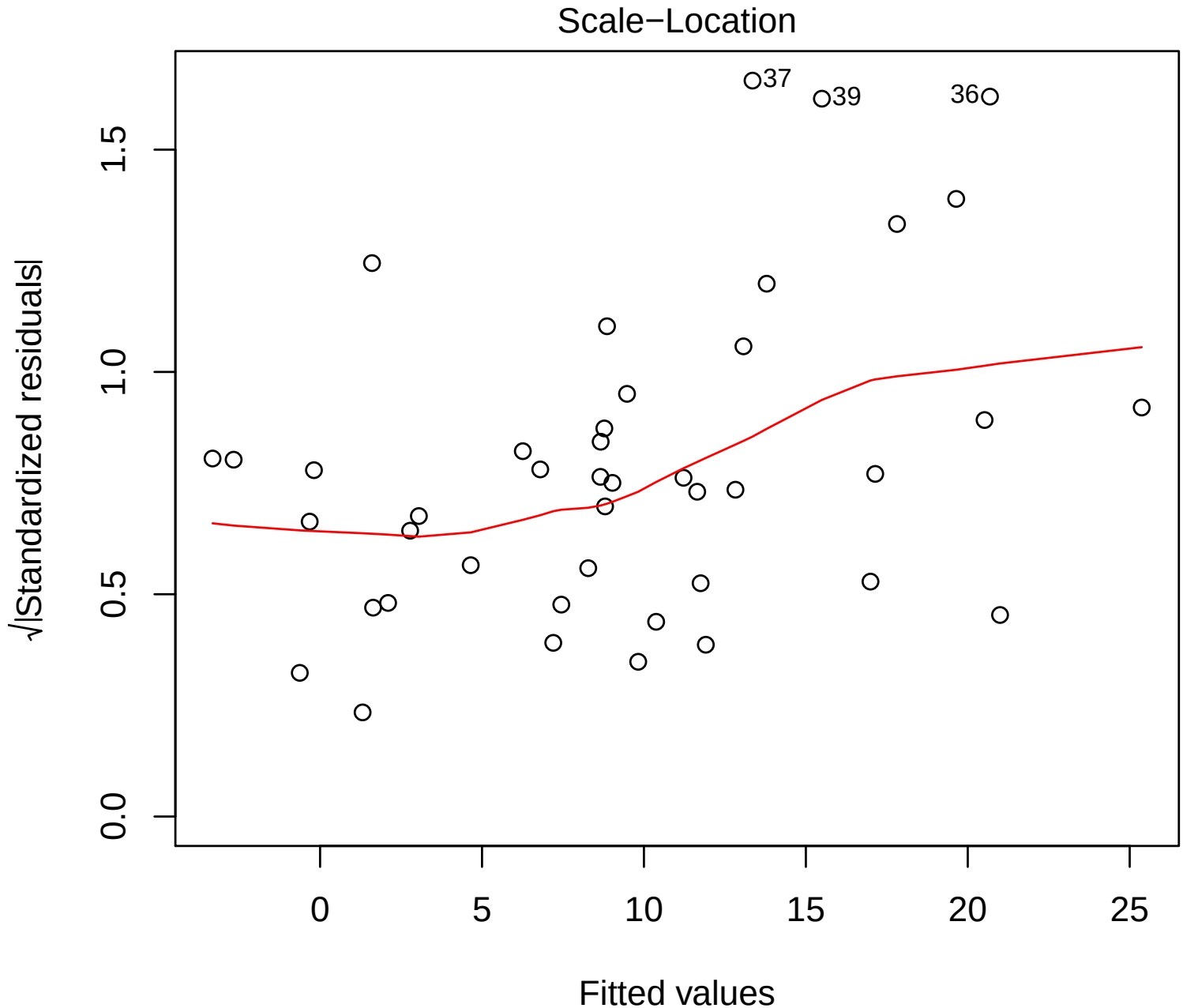
#1: Residuals vs. Fitted Values



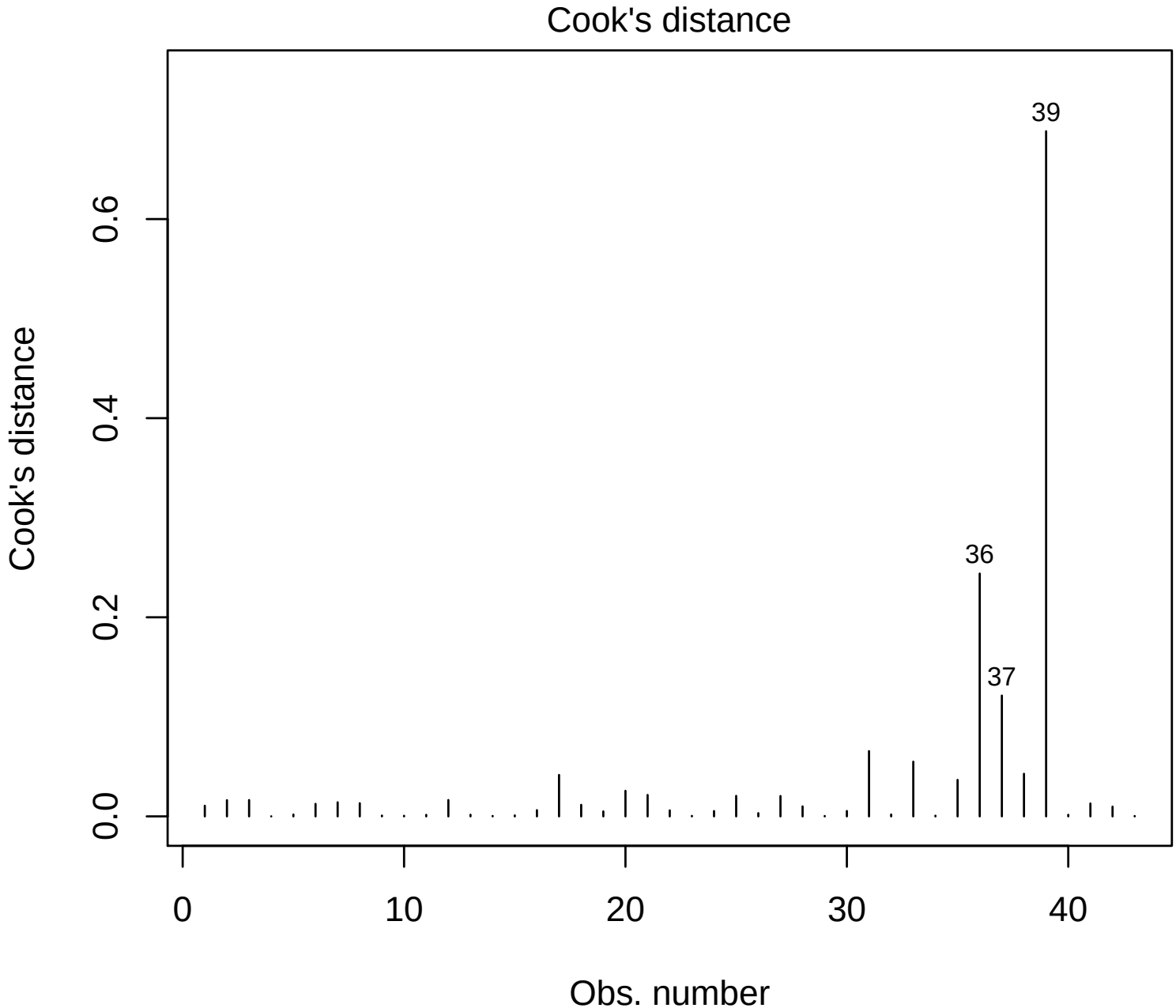
#2: Q-Q Plot of $\hat{u}$s



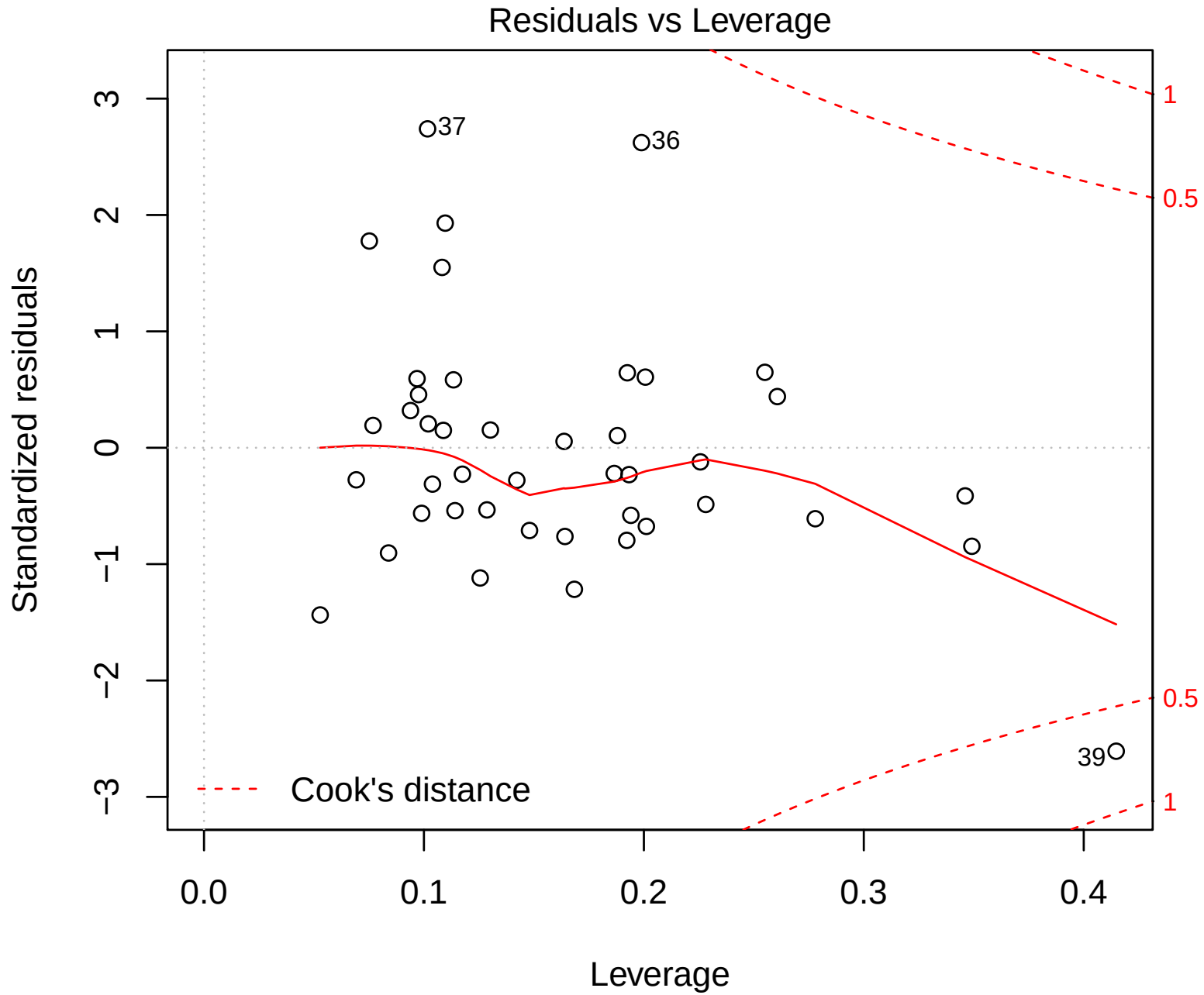
"Scale-Location" Plot



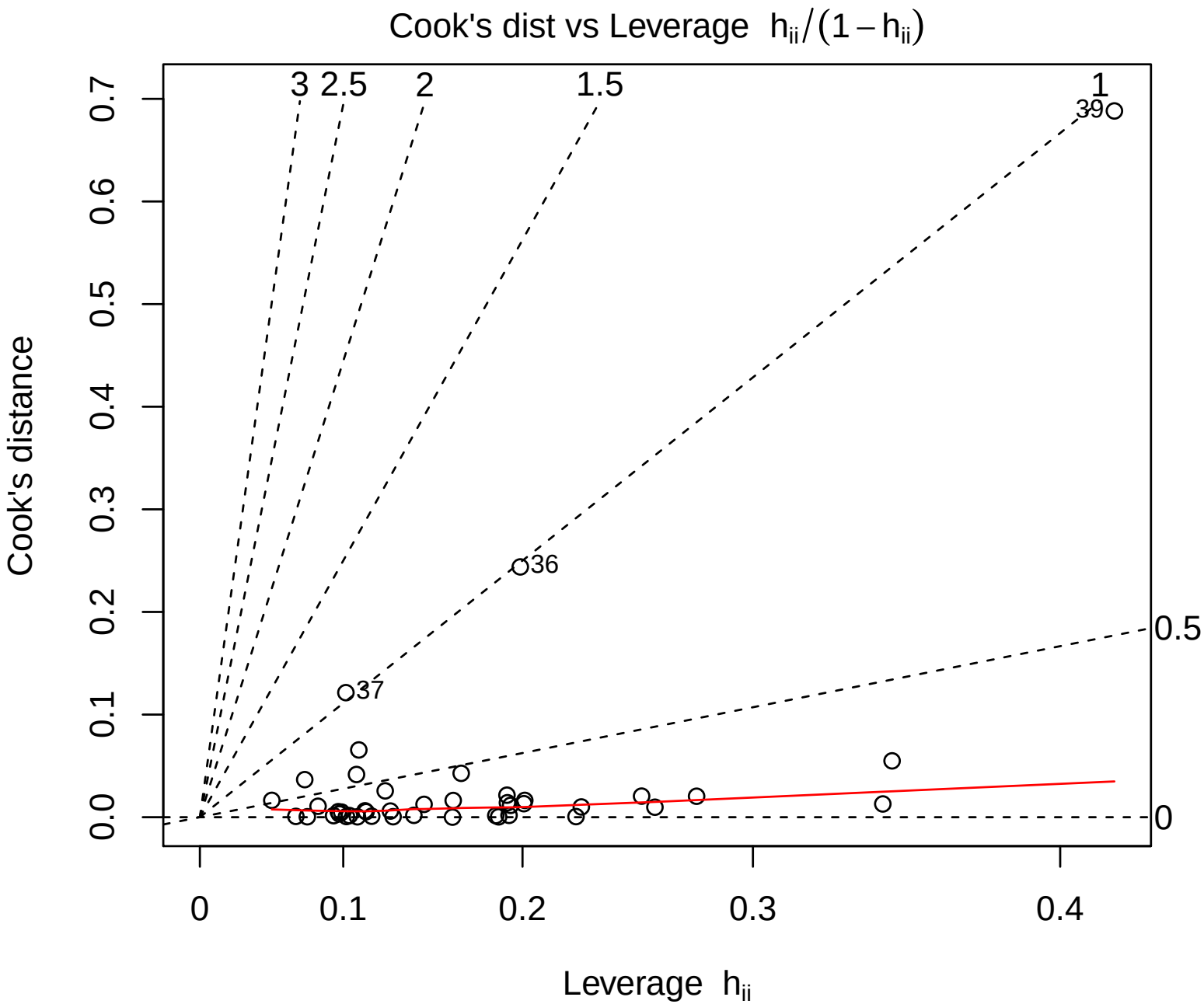
Cook's D


Cook's distance
Cook's distance
0.0
0.2
0.4
0.6
39
36
37
0
10
20
30
40
Obs. number

Residuals vs. Leverage



Cook's D vs. Leverage



BREAK

"Robust" Inference

Many of the assumptions of the classical linear regression model are necessary for accurate variance estimation, but not necessarily for unbiasedness. These are also commonly violated in practice. Under such circumstances, we often turn to "robust" methods for variance estimation.

- "Robust" (Huber / White) variance-covariance estimators
- Permutation / exact tests
- Jacknife
- **Bootstrap**
- Cross-Validation

The population is to the sample as the sample is to the bootstrap sample.

Practical (Nonparametric) Bootstrapping

- Draw one bootstrap sample of size N **with replacement** from the original data,
- Estimate the parameter(s) $\underset{k\times 1}{\tilde{\theta}}$,
- Repeat steps 1 and 2 R times, to get $\tilde{\theta}_r$, $r \in \{1, 2, ...R\}$, comprising elements $\tilde{\theta}_{rk}$,
- Examine the empirical characteristics of the resulting distribution(s) of $\tilde{\theta}_{rk}$.

Why Bootstrap?

- **It's intuitive.**
- **It's simple.**
- **It's robust.**

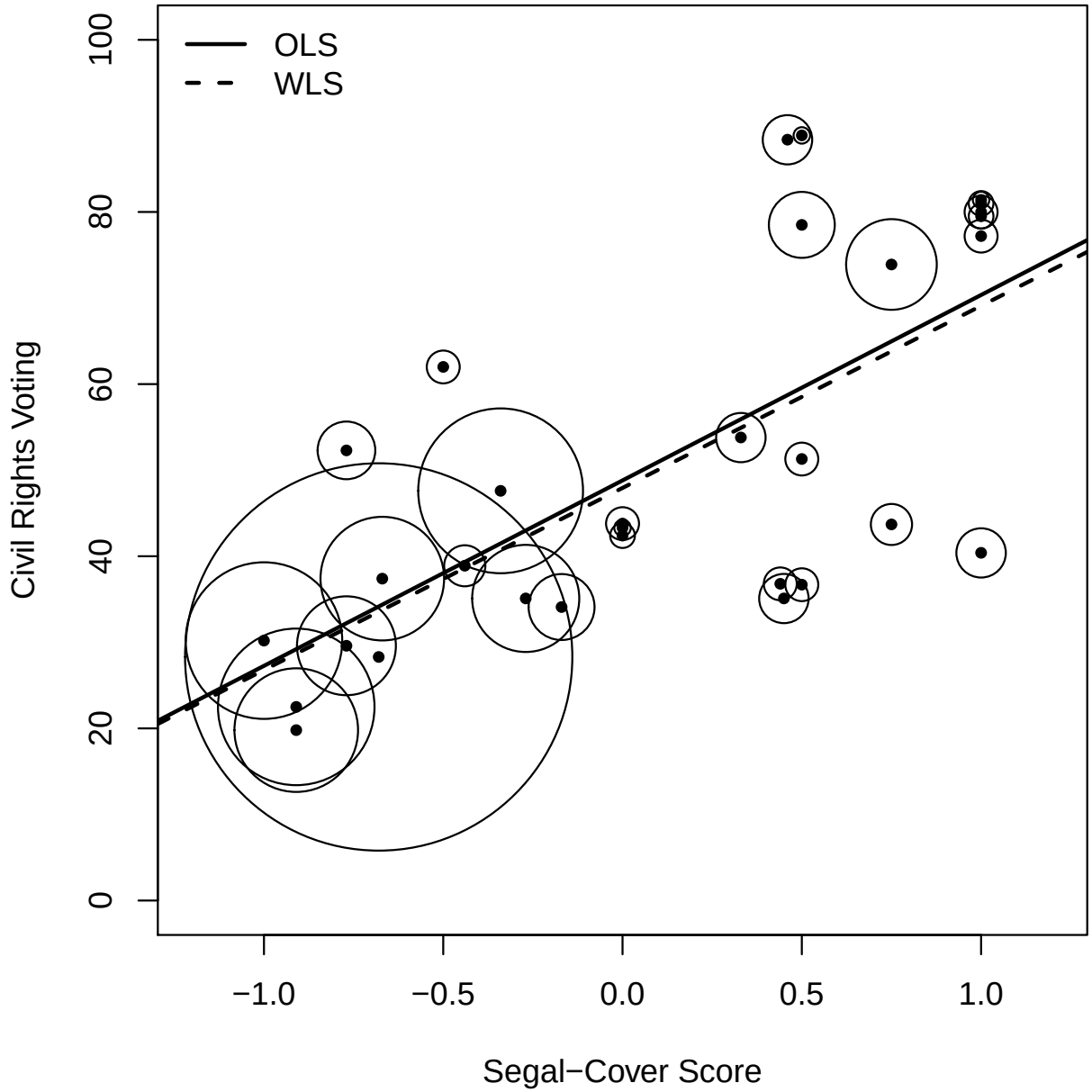
Bootstrap Example

```
> Justices<-read.csv("Justices.csv")
> attach(Justices)
> summary(Justices)
     name               score              civrts          econs
 Length:31          Min.   :-1.0000   Min.   :19.80   Min.   :34.60
 Class :character   1st Qu.:-0.4700   1st Qu.:35.90   1st Qu.:43.85
 Mode  :character   Median : 0.3300   Median :43.70   Median :50.20
                    Mean   : 0.1210   Mean   :51.42   Mean   :55.75
                    3rd Qu.: 0.6250   3rd Qu.:75.55   3rd Qu.:66.65
                    Max.   : 1.0000   Max.   :88.90   Max.   :81.70
  Neditorials          eratio           scoresq          lnNedit
 Min.   : 2.000   Min.   : 0.5000   Min.   :0.0000   Min.   :0.6931
 1st Qu.: 4.000   1st Qu.: 0.7083   1st Qu.:0.1936   1st Qu.:1.3863
 Median : 6.000   Median : 1.0000   Median :0.2500   Median :1.7918
 Mean   : 8.742   Mean   : 2.0242   Mean   :0.4599   Mean   :1.8442
 3rd Qu.:11.500   3rd Qu.: 2.5000   3rd Qu.:0.8281   3rd Qu.:2.4414
 Max.   :47.000   Max.   :11.7500   Max.   :1.0000   Max.   :3.8501
```

Bootstrap Example

```
> OLSfit<-with(Justices, lm(civrts~score))
> summary(OLSfit)

Call:
lm(formula = civrts ~ score)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   48.810      2.852  17.113  < 2e-16 ***
score         21.544      4.206   5.122 1.81e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 15.63 on 29 degrees of freedom
Multiple R-squared: 0.475,Adjusted R-squared: 0.4569
F-statistic: 26.24 on 1 and 29 DF,  p-value: 1.806e-05
```

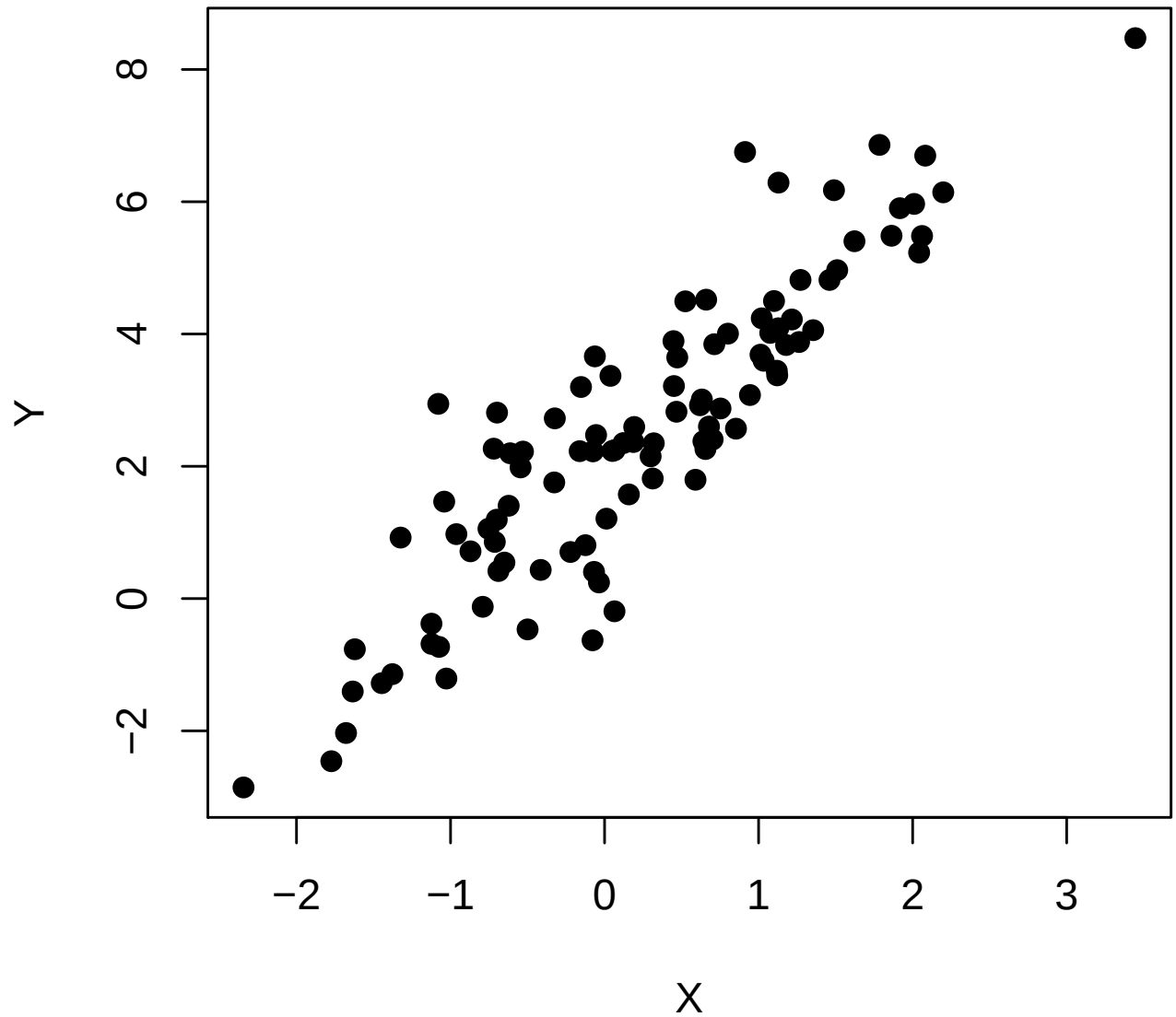
Figure 1: Plot of `civrts` Against `score`, Weighted by `Neditorials`



Bootstrapping: "By Hand"

```
N<-100
reps<-999

set.seed(1337)
X<-rnorm(N)
Y<-2+2*X+rnorm(N)
data<-data.frame(Y,X)
fitOLS<-lm(Y~X)
CI<-confint(fitOLS)

B0<-numeric(reps)
B1<-numeric(reps)

for (i in 1:reps) {
  temp<-data[sample(1:N,N,replace=TRUE),]
  temp.lm<-lm(Y~X,data=temp)
  B0[i]<-temp.lm$coefficients[1]
  B1[i]<-temp.lm$coefficients[2]
}

ByHandB0<-median(B0)
ByHandB1<-median(B1)
ByHandCI.B0<-quantile(B0,probs=c(0.025,0.975)) # <-- 95% c.i.s
ByHandCI.B1<-quantile(B1,probs=c(0.025,0.975))
```

Bootstrapping "By Hand"



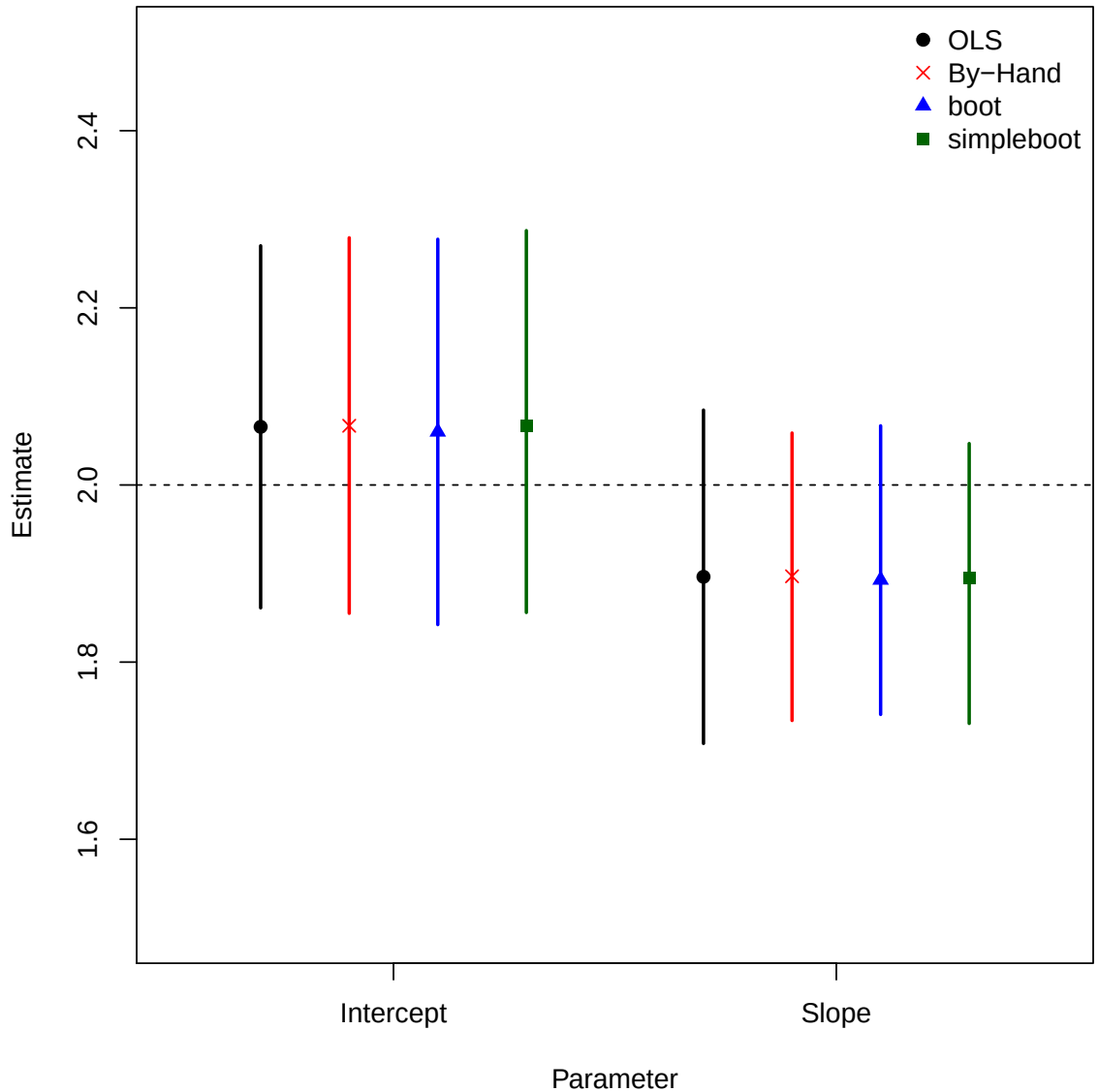
Bootstrapping Via boot

```
library(boot)

Bs<-function(formula, data, indices) { # <- regression function
    dat <- data[indices,]
    fit <- lm(formula, data=dat)
    return(coef(fit))
}

Boot.fit<-boot(data=data, statistic=Bs,
          R=reps, formula=Y~X)

BootB0<-median(Boot.fit$t[,1])
BootB1<-median(Boot.fit$t[,2])
BootCI.B0<-boot.ci(Boot.fit,type="basic",index=1)
BootCI.B1<-boot.ci(Boot.fit,type="basic",index=2)
```

Bootstrapping Via simpleboot

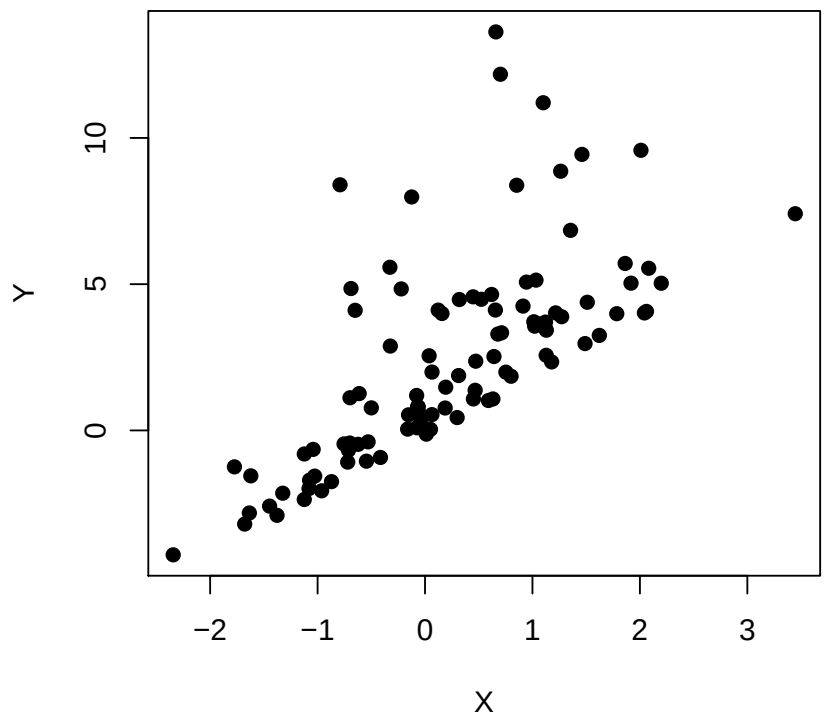
```
library(simpleboot)

Simple<-lm.boot(fitOLS,reps)
SimpleB0<-perc(Simple,.50)[1]
SimpleB1<-perc(Simple,.50)[2]
Simple.CIs<-perc(Simple,p=c(0.025,0.975))
```

Bootstrapping Results



Bootstrapping: Skewed Residuals

```
N<-100
reps<-999

set.seed(1337)
X<-rnorm(N)
ustar<-rchisq(N,2) # <- skewed u.s
Y<-2+2*X+(ustar-mean(ustar))
data<-data.frame(Y,X)
fitOLS<-lm(Y~X)
CI<-confint(fitOLS)
```



Skewed Residuals: Results