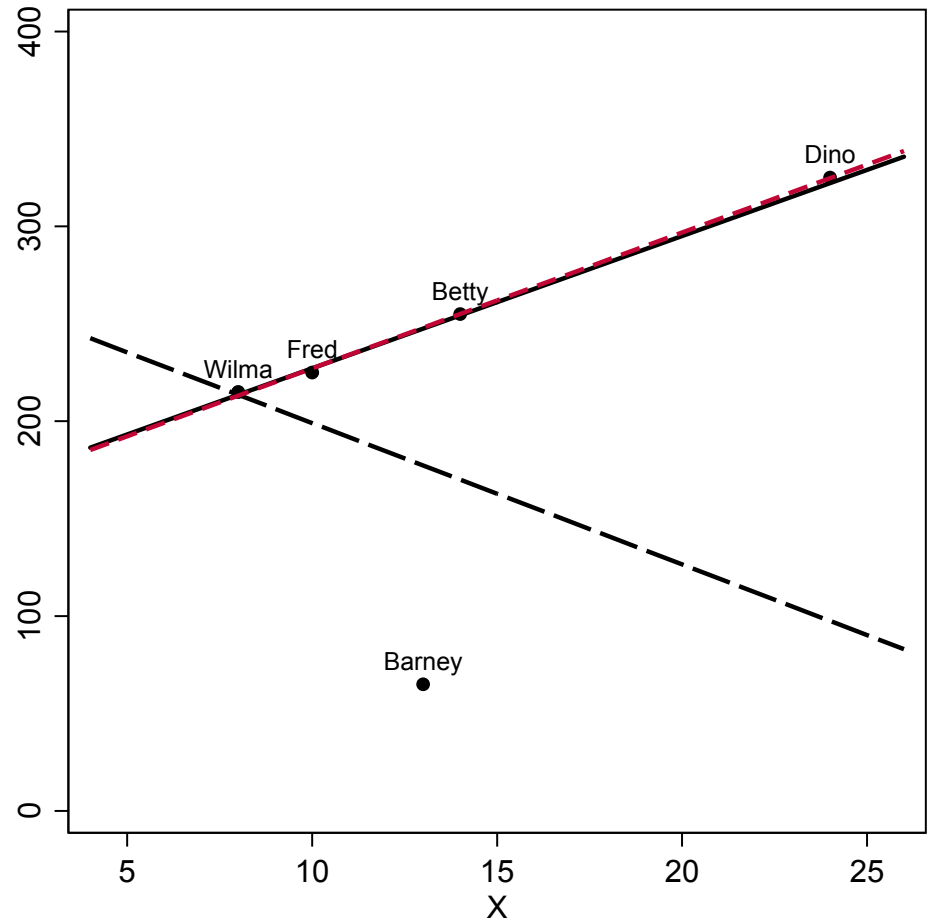
PLSC 503 – Spring 2023
Residuals, Model Fit, and Outliers

February 20, 2023

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
```

A Variance-Based Statistic

"COVRATIO":

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

For observation i:

- $\text{COVRATIO}_i > 1 \rightarrow$ *increased* precision of the estimates / *smaller* standard errors
- $\text{COVRATIO}_i < 1 \rightarrow$ *decreased* precision of the estimates / *larger* standard errors

Example: Federal Judicial Review, 1789-2018

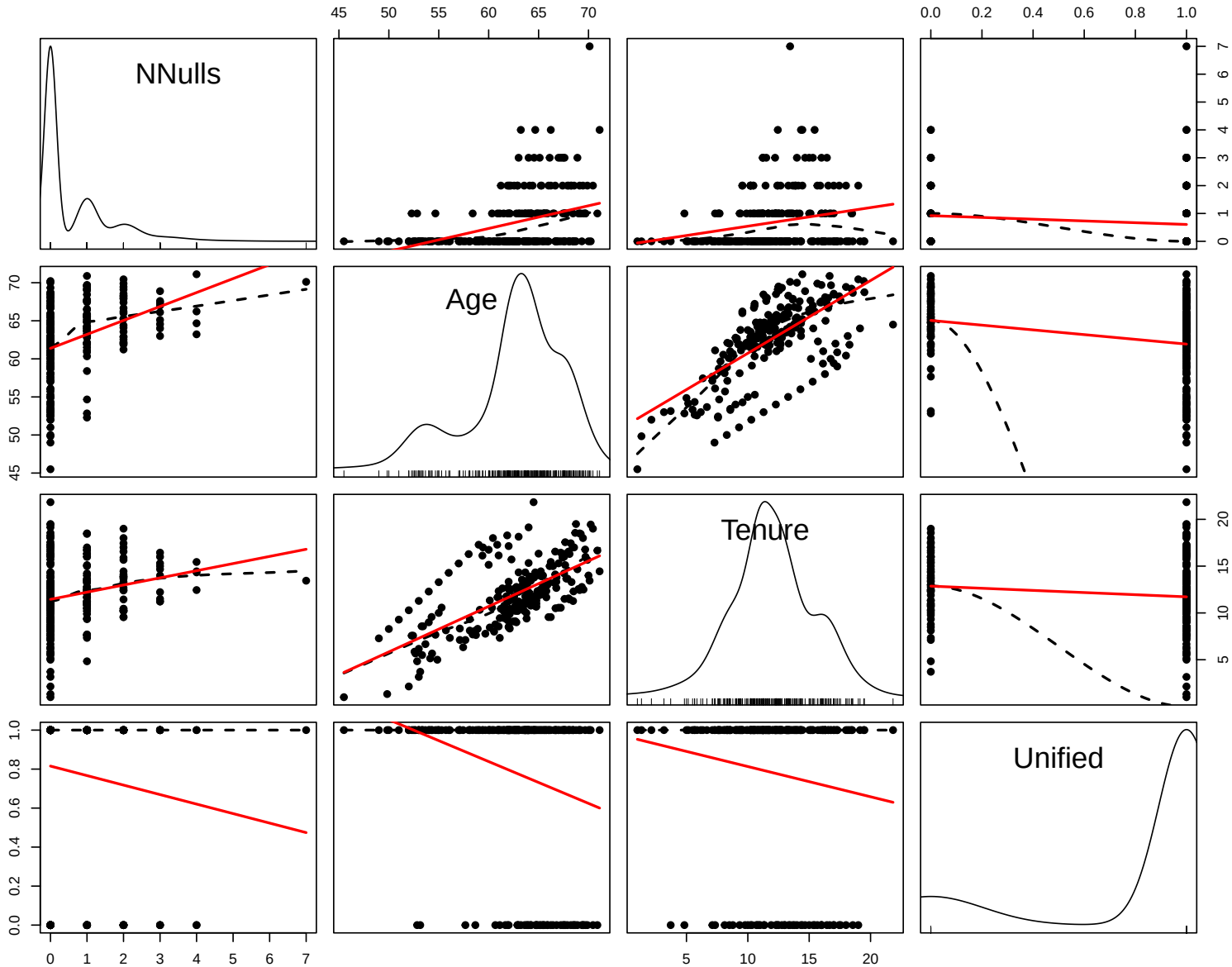
Dahl (1957):

- SCOTUS gets "out of step" with the other branches → judicial review
- Older / longer-serving justices will more likely to invalidate legislation

Data:

```
> summary(NewDahl)
      Year          NNulls            Age           Tenure        Unified
 Min.   :1789   Min.   :0.000   Min.   :45.5   Min.   : 1.0   Min.   :0.000
 1st Qu.:1846   1st Qu.:0.000   1st Qu.:60.7   1st Qu.:10.0   1st Qu.:1.000
 Median :1904   Median :0.000   Median :63.5   Median :11.8   Median :1.000
 Mean   :1904   Mean   :0.674   Mean   :62.6   Mean   :12.0   Mean   :0.783
 3rd Qu.:1961   3rd Qu.:1.000   3rd Qu.:66.0   3rd Qu.:14.1   3rd Qu.:1.000
 Max.   :2018   Max.   :7.000   Max.   :71.1   Max.   :21.8   Max.   :1.000
```

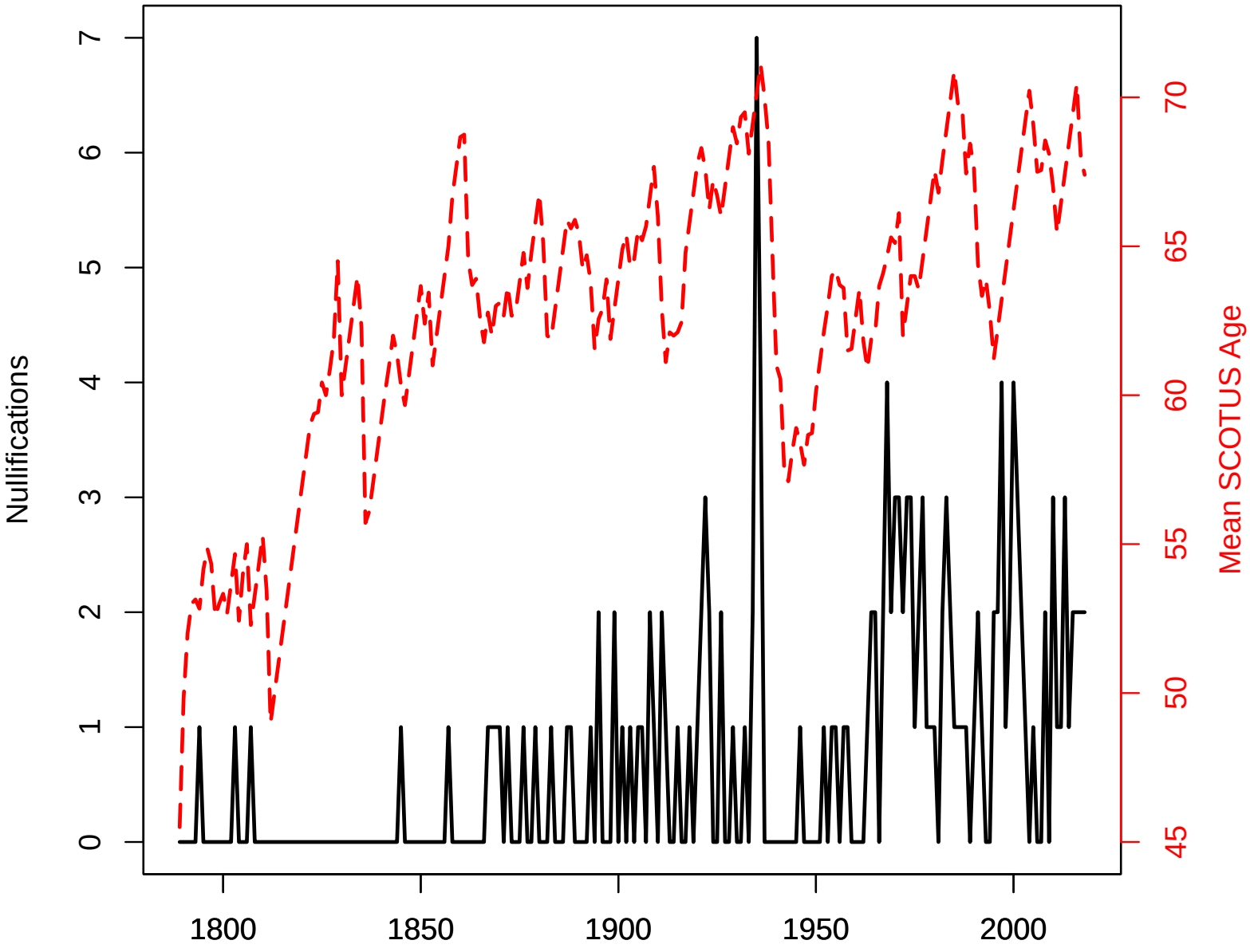
Example: Federal Judicial Review, 1789-2018



Basic Regression...

```
> Fit<-with(NewDahl, lm(NNulls~Age+Tenure+Unified))
> summary(Fit)

Call:
lm(formula = NNulls ~ Age + Tenure + Unified)

Residuals:
   Min     1Q Median     3Q    Max
-1.308 -0.700 -0.135  0.308  5.693

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -4.6833     1.0026   -4.67 0.0000051 ***
Age           0.0901     0.0181    4.97 0.0000013 ***
Tenure       -0.0201     0.0248   -0.81      0.42
Unified      -0.0573     0.1613   -0.36      0.72
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.973 on 226 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.152,Adjusted R-squared:  0.141
F-statistic: 13.6 on 3 and 226 DF,  p-value: 0.0000000365
```

Federal Judicial Review and Mean SCOTUS Age

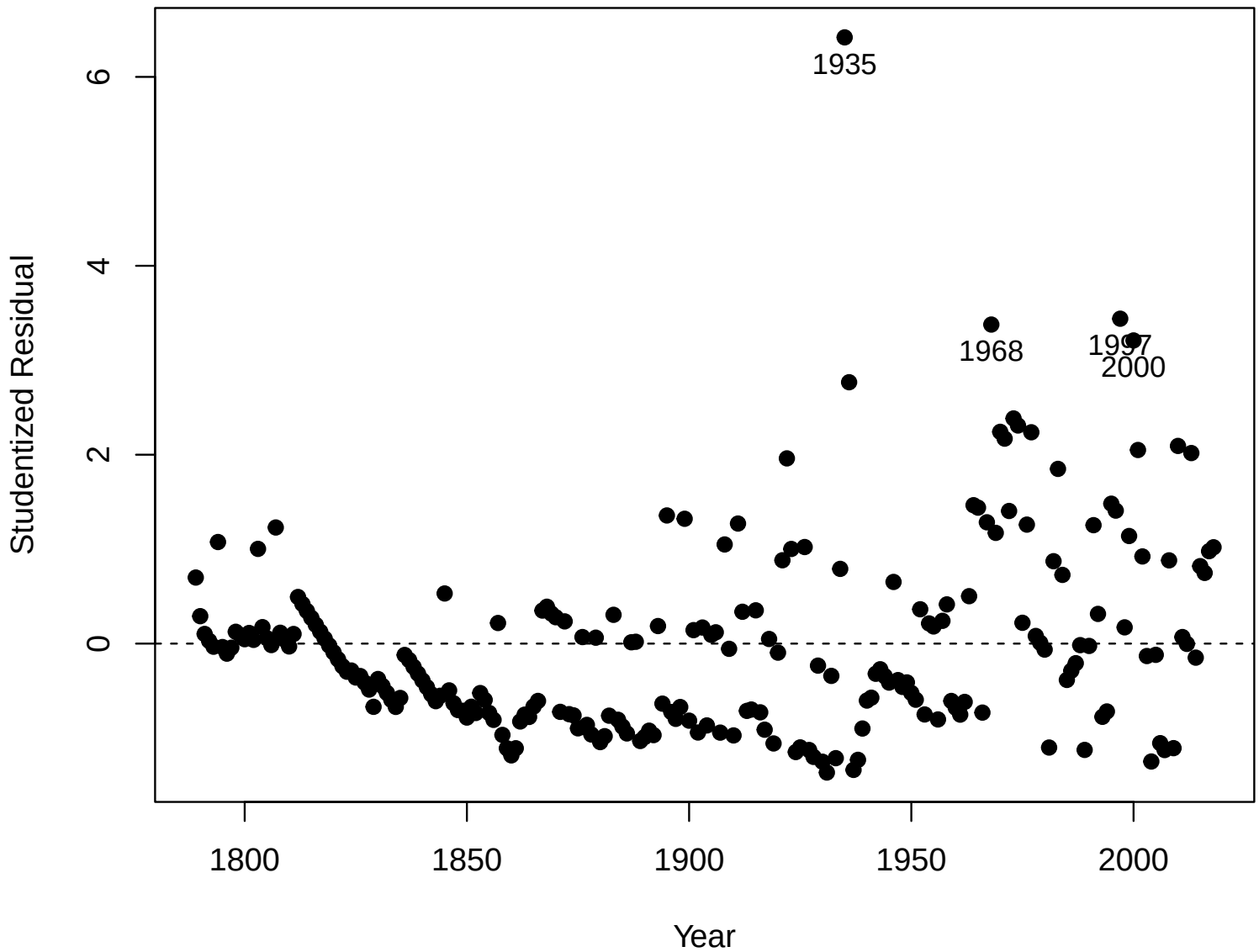


Residuals, etc.

Generate some statistics:

```
> FitResid<-with(NewDahl, (Fit$model$NNulls - predict(Fit)))
> 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

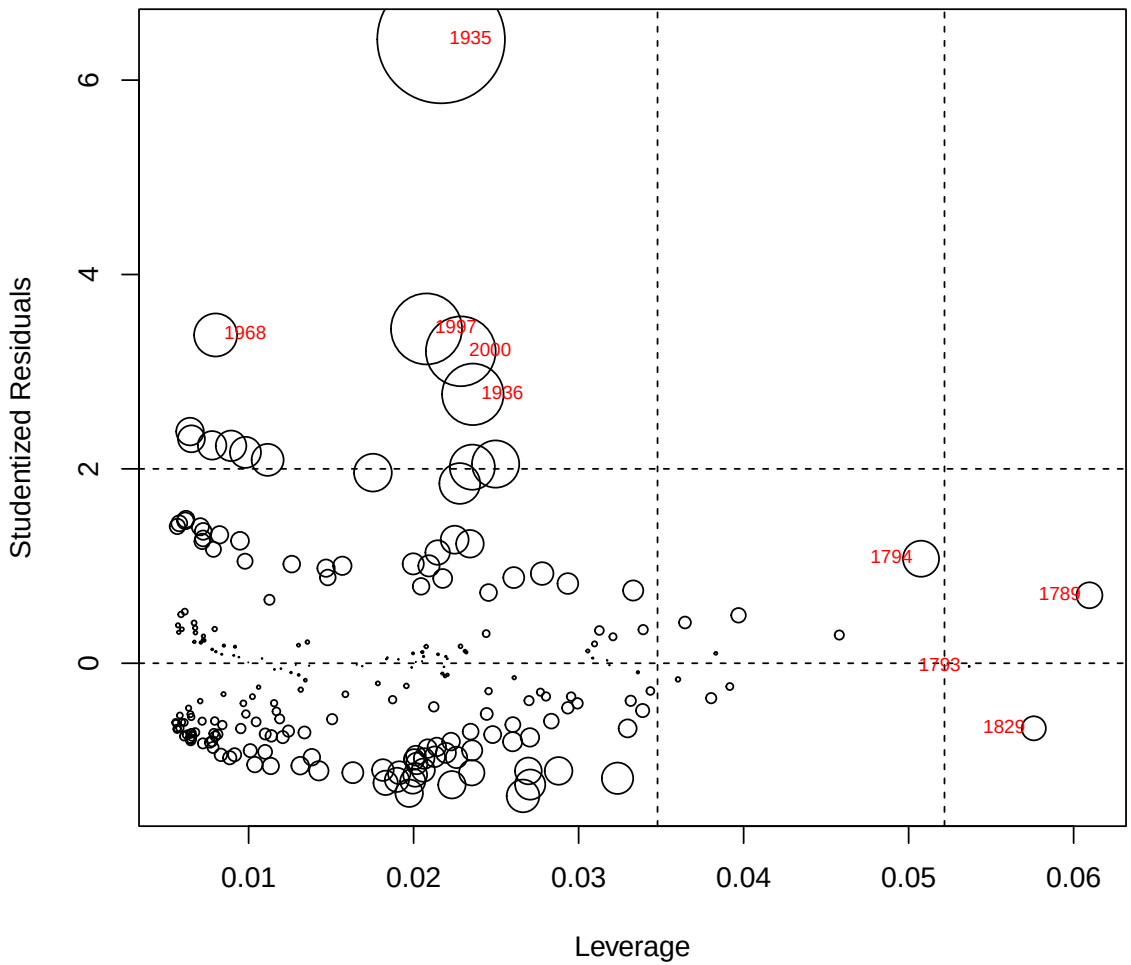


More About Studentized Residuals

```
> max(FitStudent)
[1] 6.418

> NewDahl$Year1935<-ifelse(NewDahl$Year==1935,1,0)
> summary(with(NewDahl, lm(NNulls~Age+Tenure+Unified+Year1935)))

Residuals:
   Min     1Q Median     3Q    Max
-1.250 -0.652 -0.122  0.302  3.247

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  -3.9298     0.9313   -4.22 0.00003546916 ***
Age           0.0768     0.0168    4.56 0.00000846697 ***
Tenure       -0.0113     0.0229   -0.50          0.62
Unified      -0.1210     0.1490   -0.81          0.42
Year1935      5.8186     0.9066    6.42 0.00000000081 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.897 on 225 degrees of freedom
Multiple R-squared:  0.284,Adjusted R-squared:  0.271
F-statistic: 22.3 on 4 and 225 DF,  p-value: 1.65e-15
```

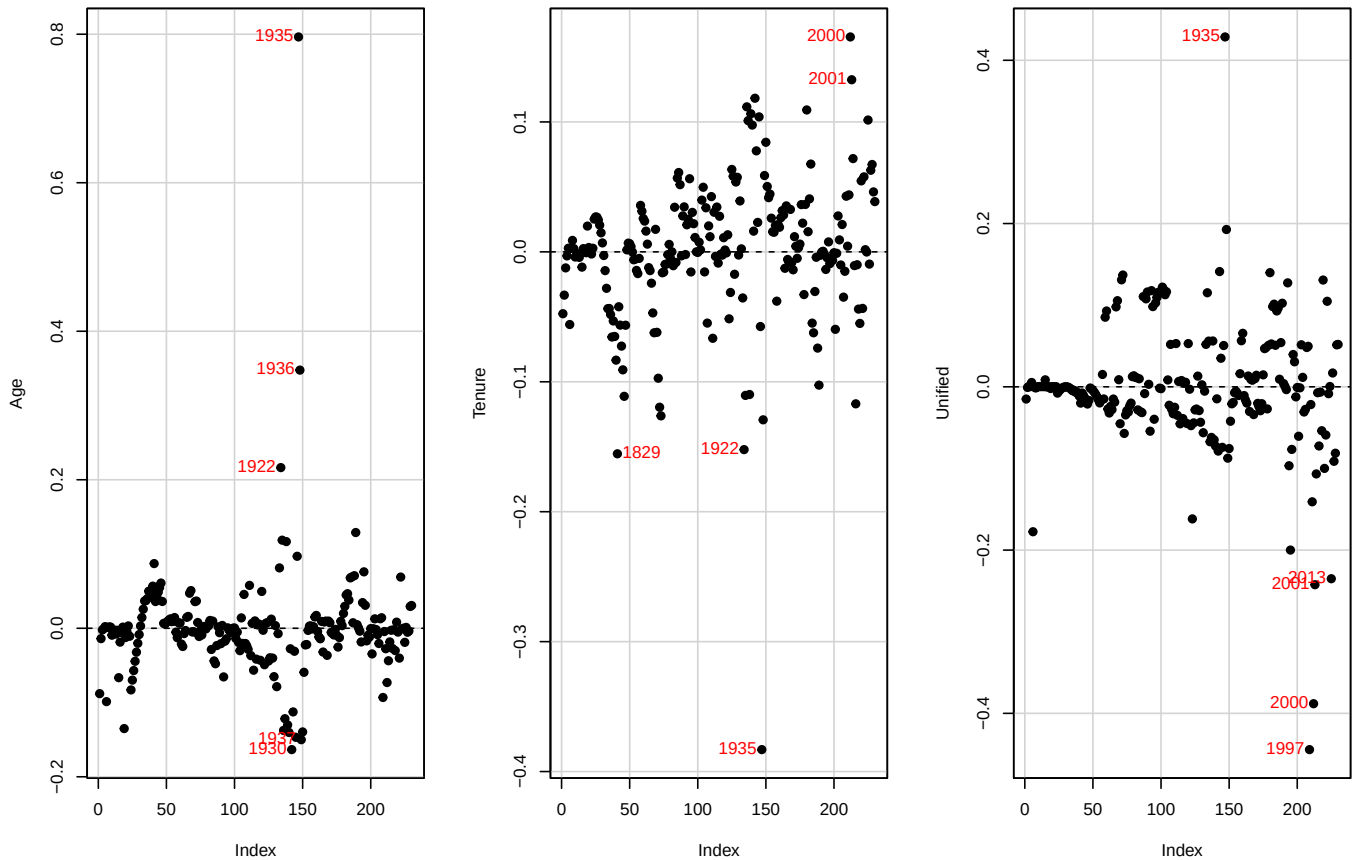
"Bubble Plot"

```
> influencePlot(Fit,id=list(method="noteworthy",n=4,cex=0.7,
                            labels=NewDahl$Year,col="red"),
                xlab="Leverage")
```



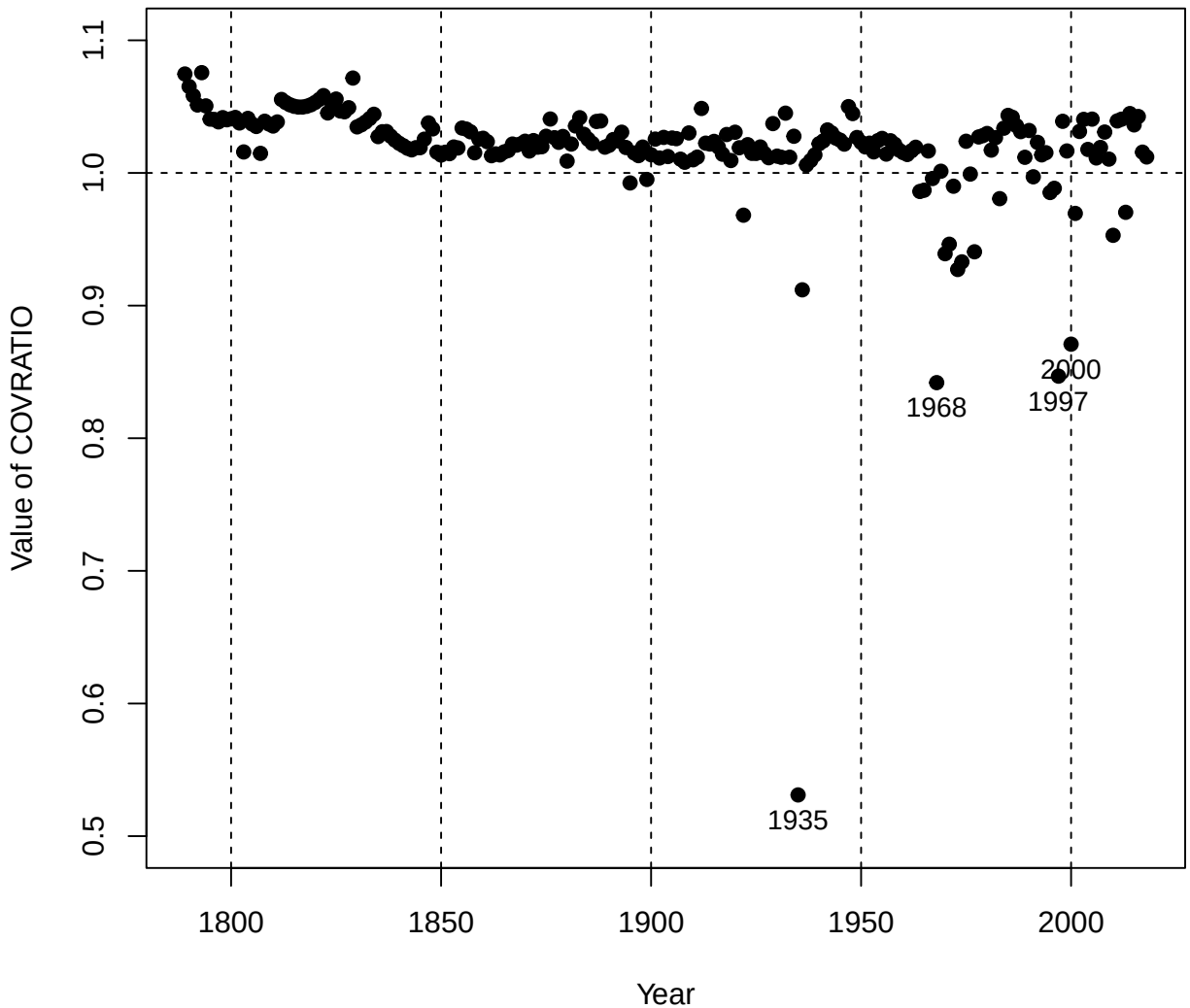
DFBETAS

```
> dfbetasPlots(Fit,id.n=5,id.col="red",main="",pch=19,
                 layout=c(1,3),labels=NewDahl$Year)
```



COVRATIO Plot

```
> plot(FitCOVRATIO~NewDahl$Year,pch=19,ylim=c(0.5,1.1),
     xlab="Year",ylab="Value of COVRATIO")
```

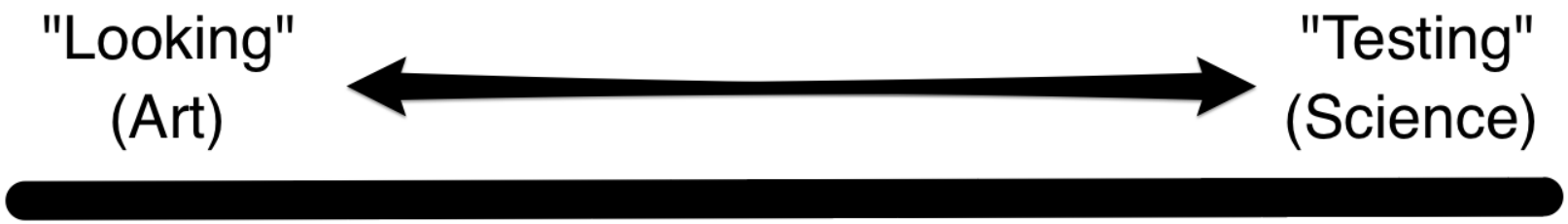


Sensitivity Analyses: Omitting Outliers

```
> out1<-c(1935) # one outlier
> LD2<-NewDahl[!(NewDahl$Year %in% out1),]
> out2<-c(1935,1968,1997,2000) # four outliers
> LD3<-NewDahl[!(NewDahl$Year %in% out2),]
> Fit2<-lm(NNulls~Age+Tenure+Unified,data=LD2)
> Fit3<-lm(NNulls~Age+Tenure+Unified,data=LD3)
```

	Dependent variable:		
	NNulls		
	(1)	(2)	(3)
Age	0.090***	0.077***	0.079***
	(0.018)	(0.017)	(0.015)
Tenure	−0.020	−0.011	−0.019
	(0.025)	(0.023)	(0.021)
Unified	−0.057	−0.121	−0.010
	(0.161)	(0.149)	(0.139)
Constant	−4.683***	−3.930***	−4.130***
	(1.003)	(0.931)	(0.855)
Observations	230	229	226
R^2	0.152	0.148	0.158
Adjusted R^2	0.141	0.137	0.147
Residual Std. Error	0.973 (df = 226)	0.897 (df = 225)	0.822 (df = 222)
F Statistic	13.550*** (df = 3; 226)	13.030*** (df = 3; 225)	13.930*** (df = 3; 222)
Note:			*p<0.1; **p<0.05; ***p<0.01

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.

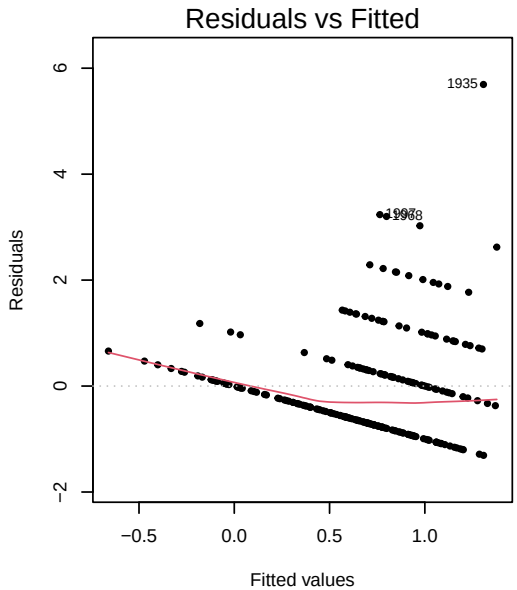
Tests for:

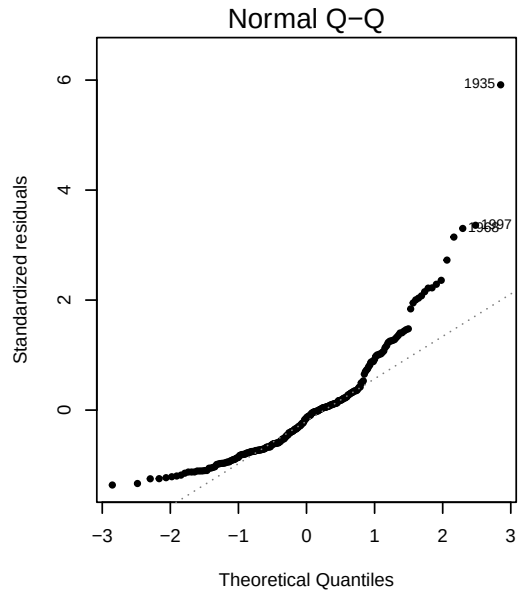
- Normality in $\hat{u}$s (via skewness & kurtosis tests)
- "Link function" (linearity / additivity)
- Constant variance and uncorrelatedness in $\hat{u}$s ("heteroskedasticity" test)

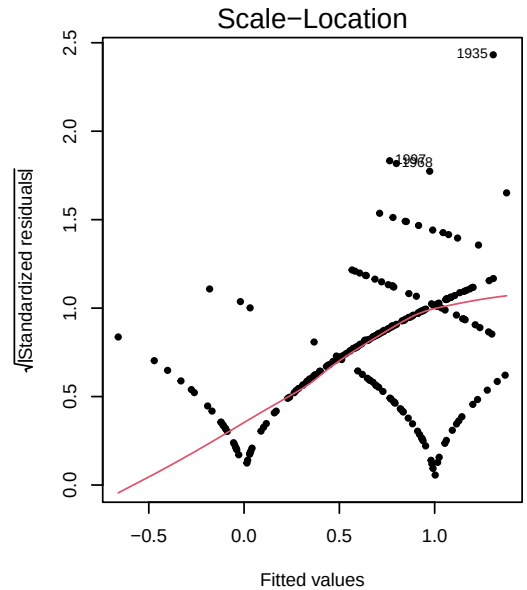
In Action

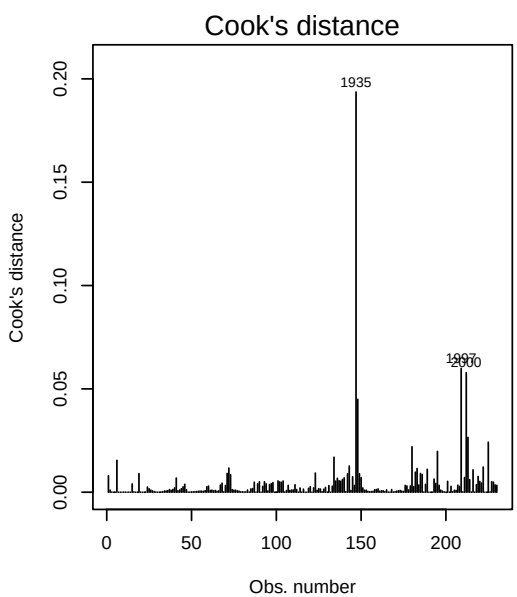
```
> Fit<-with(NewDahl, lm(NNulls~Age+Tenure+Unified))

> 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        454.87 0.00e+00 Assumptions NOT satisfied!
Skewness           122.09 0.00e+00 Assumptions NOT satisfied!
Kurtosis           283.21 0.00e+00 Assumptions NOT satisfied!
Link Function        5.35 2.07e-02 Assumptions NOT satisfied!
Heteroscedasticity  44.23 2.92e-11 Assumptions NOT satisfied!
```

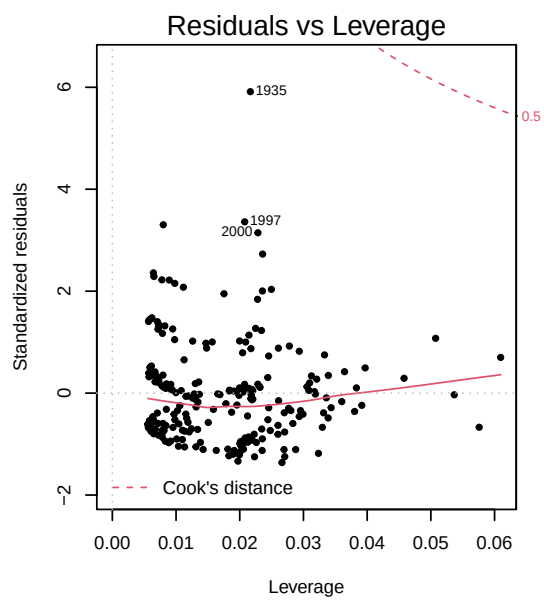
Another Approach: `plot(fit)`

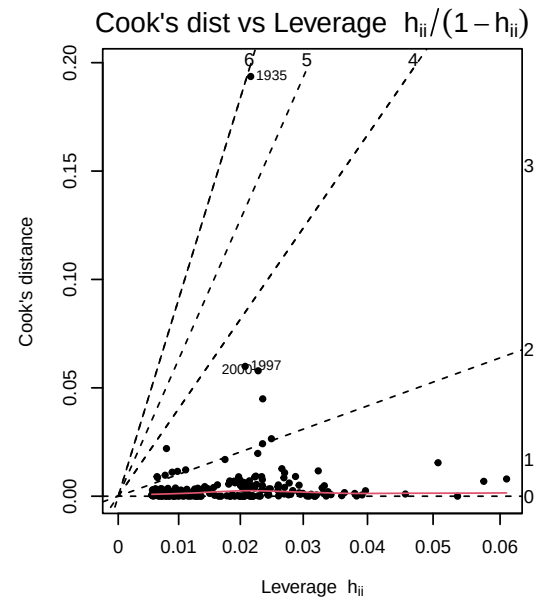




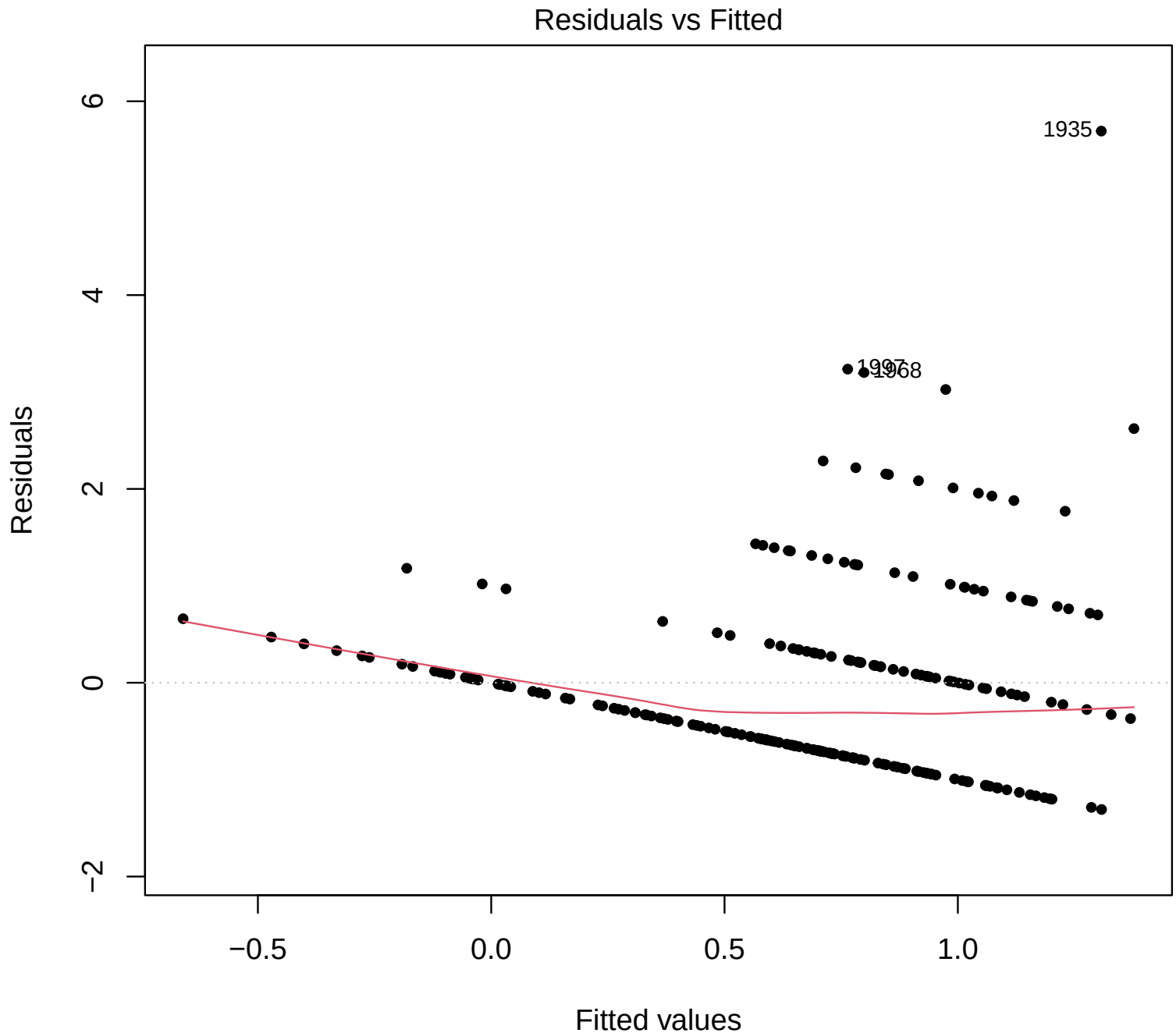




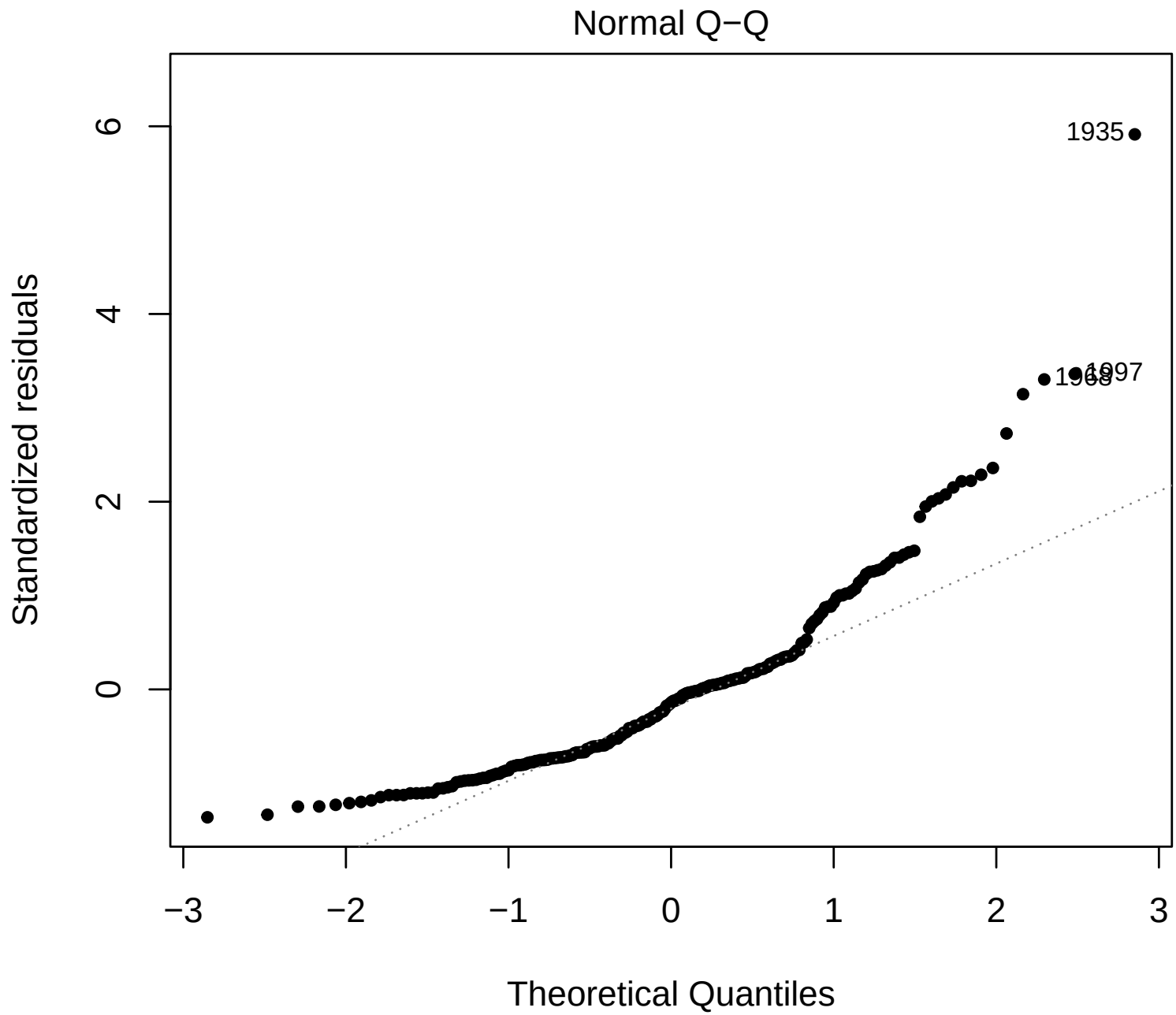




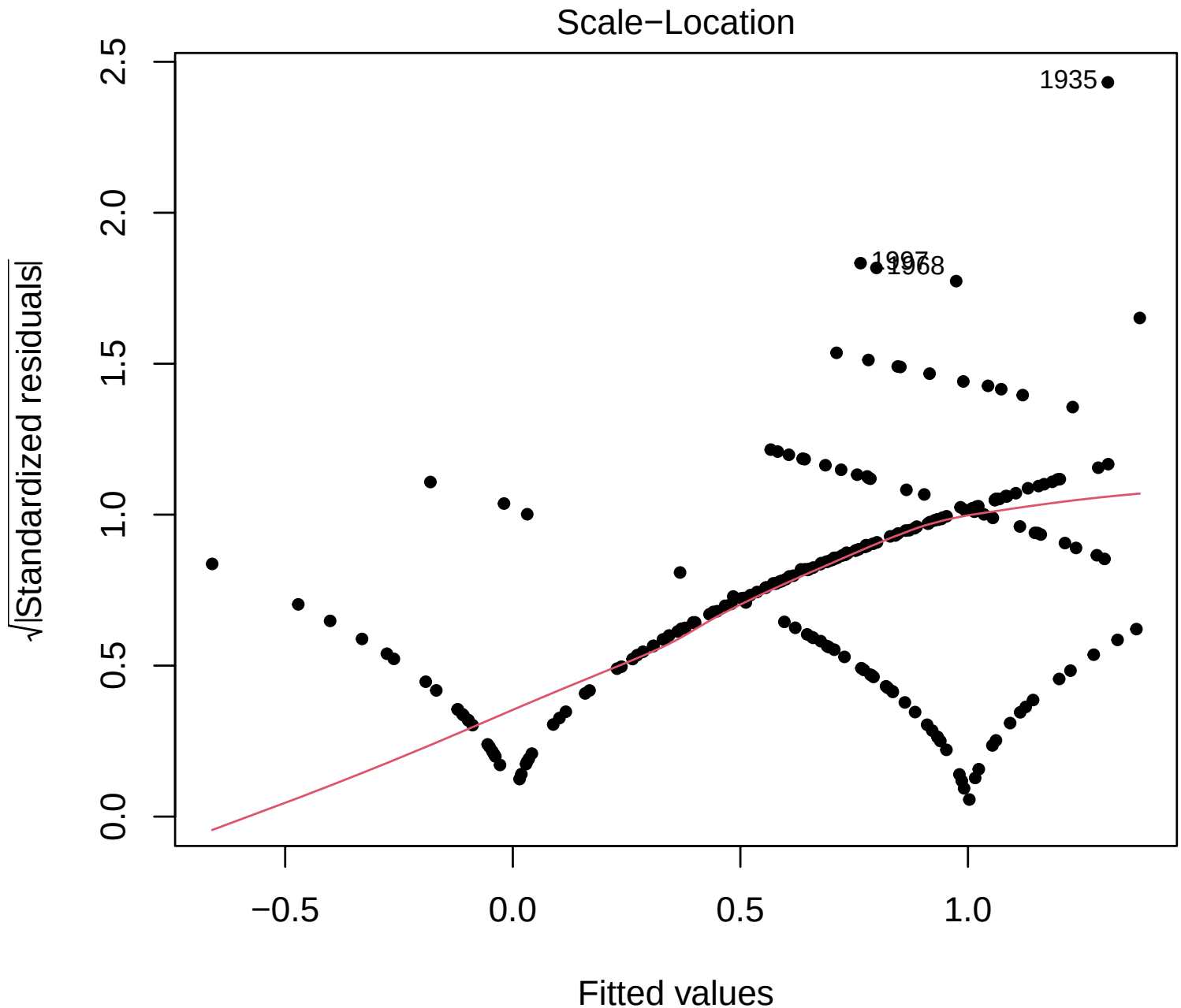
#1: Residuals vs. Fitted Values



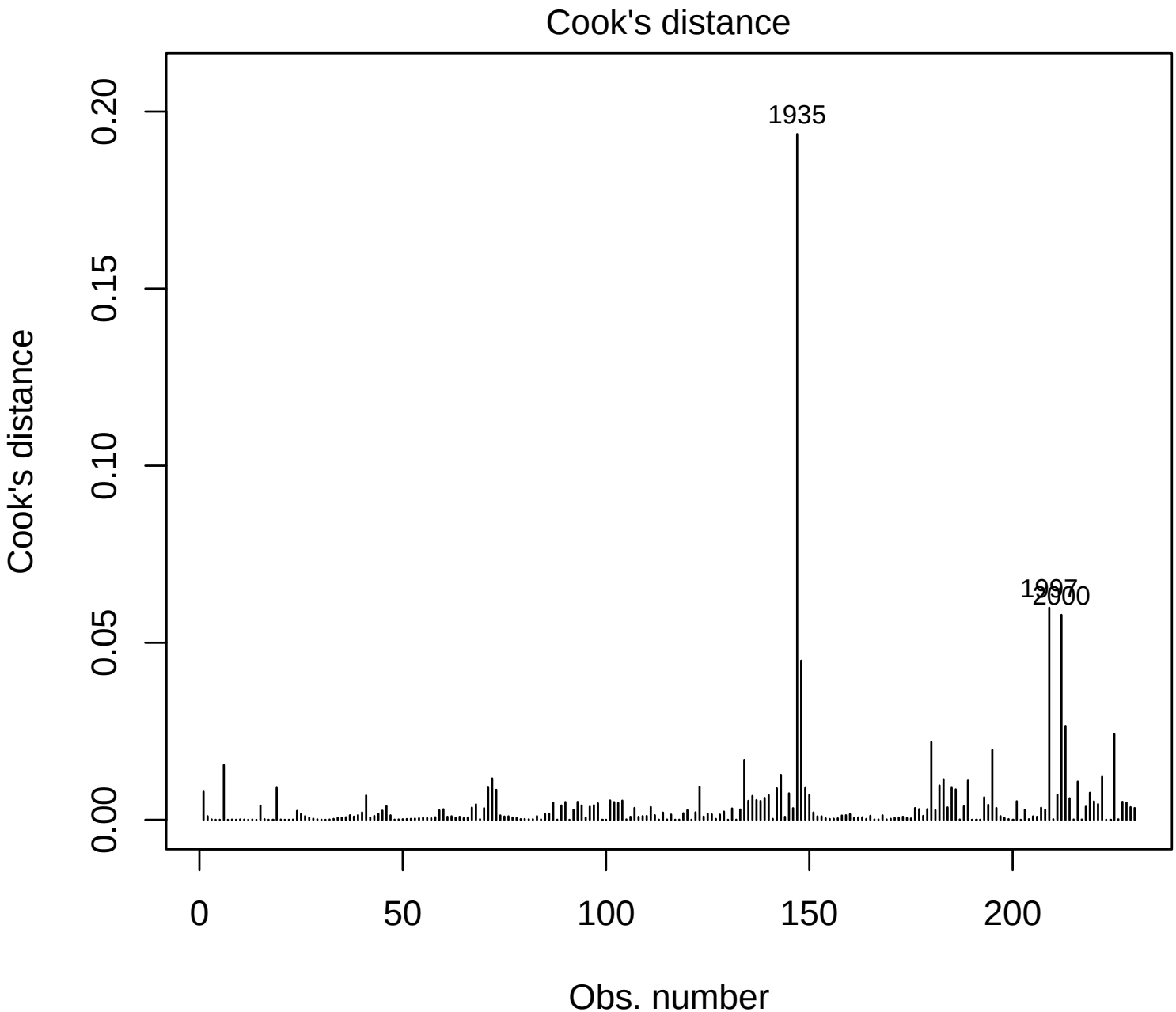
#2: Q-Q Plot of $\hat{u}$s



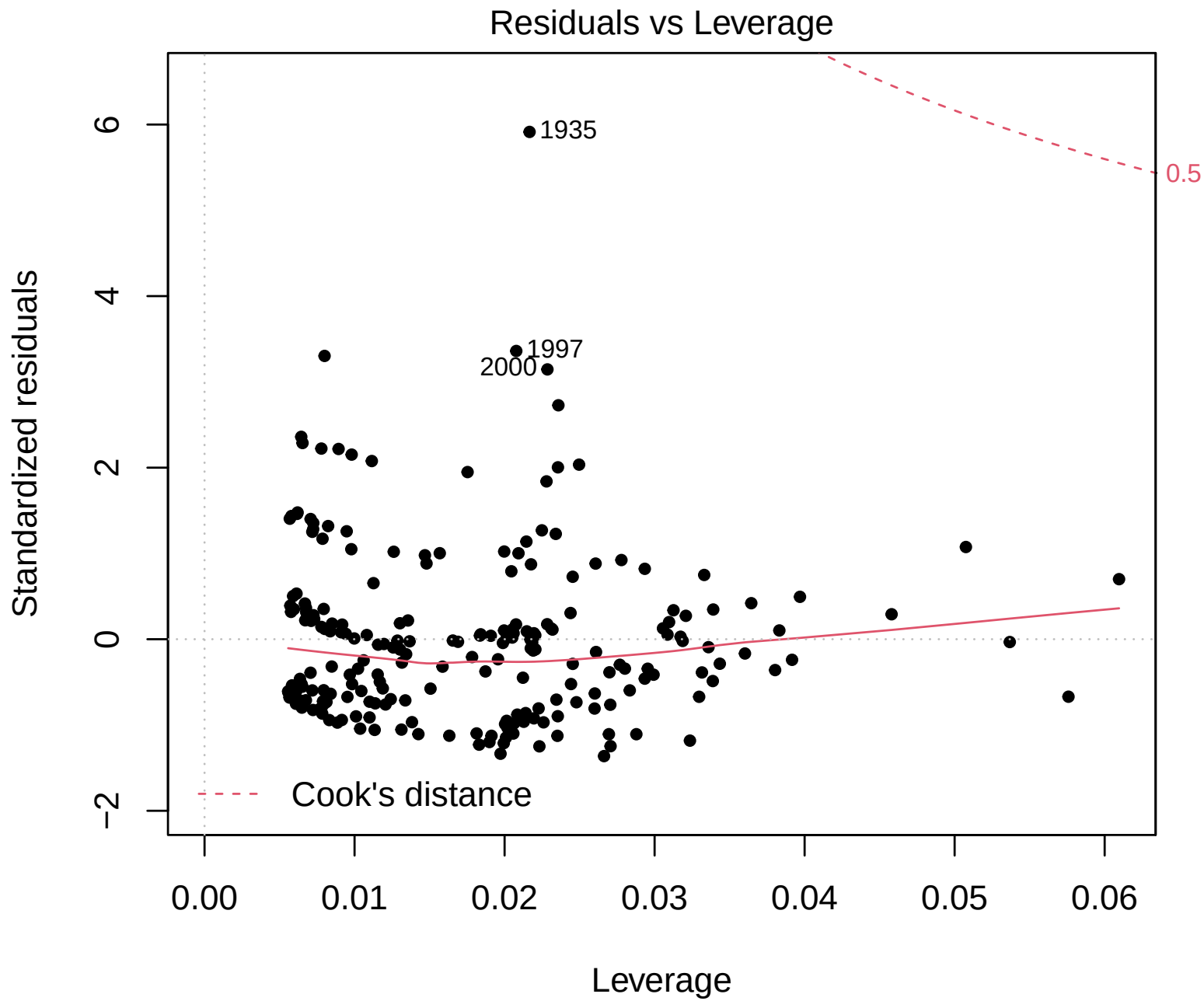
"Scale-Location" Plot



Cook's D


Cook's distance
1935
1997
2000
0.00
0.05
0.10
0.15
0.20
Cook's distance
0
50
100
150
200
Obs. number

Residuals vs. Leverage



Cook's D vs. Leverage