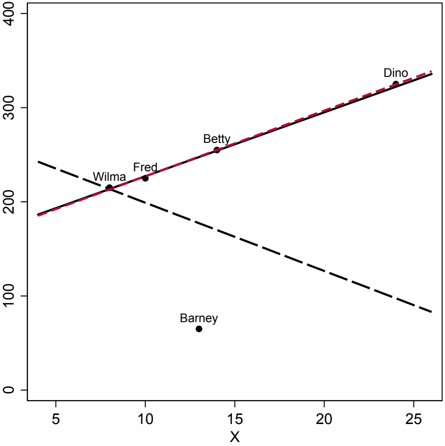
GSERM 2020
Regression for Publishing

June 17, 2020 (first 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:

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

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

"Standardized":

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

Residuals

"Studentized": define

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

Then:

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

"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
```

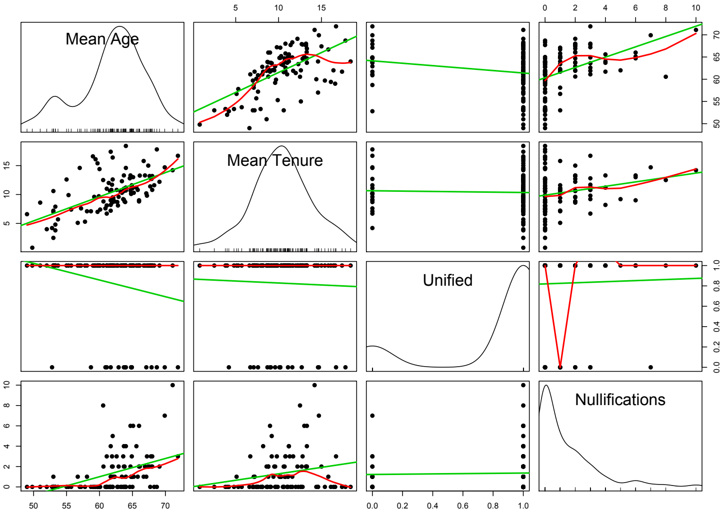
Variance

```
> # 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":

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

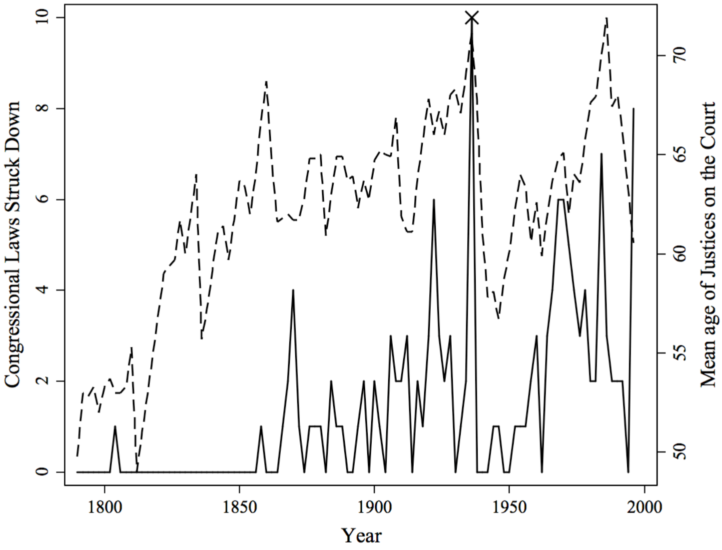
Example: Federal Judicial Review, 1789-1996



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.

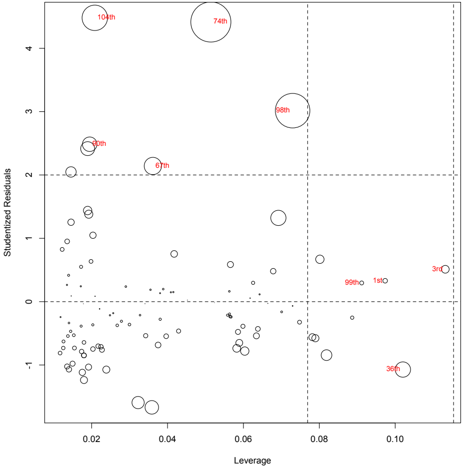
```
> FitResid<-(nulls - predict(Fit)) # residuals
> FitStandard<-rstandard(Fit) # standardized residuals
> FitStudent<-rstudent(Fit) # studentized residuals
> FitCooksD<-cooks.distance(Fit) # Cook's D
> FitDFBeta<-dfbeta(Fit) # DFBeta
> FitDFBetaS<-dfbetas(Fit) # DFBetaS
> FitCOVRATIO<-covratio(Fit)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

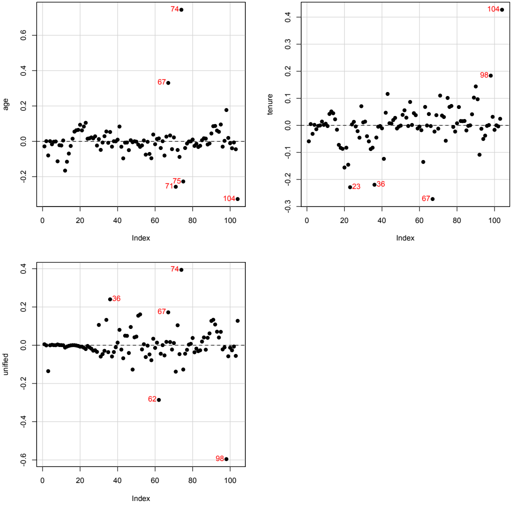
“Bubble Plot”

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



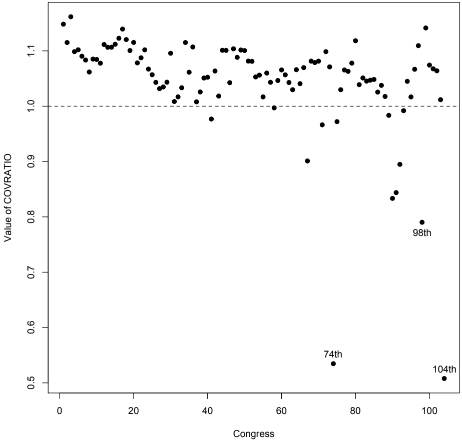
DFBETAS

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



COVRATIO Plot

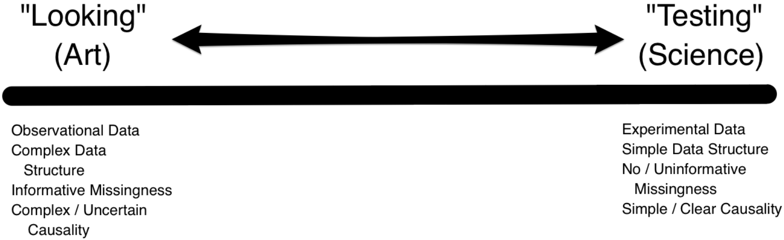
```
> plot(FitCOVRATIO~congress,pch=19,xlab="Congress",ylab="Value of COVRATIO")
> abline(h=1,lty=2)
```



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



One Approach

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

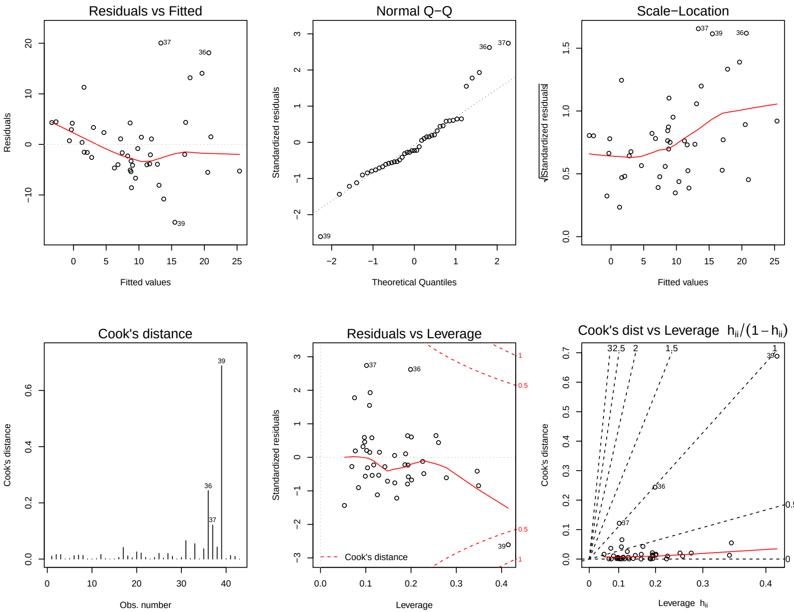
Tests for:

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

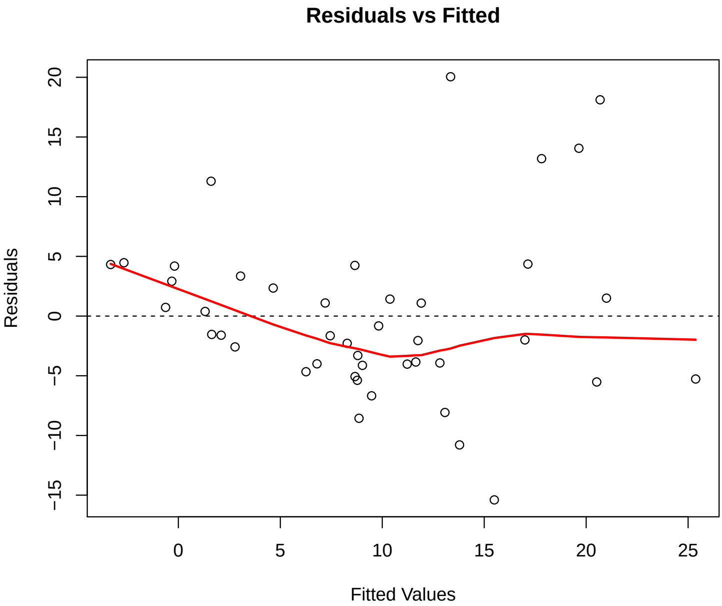
In Action

```
> 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!
```

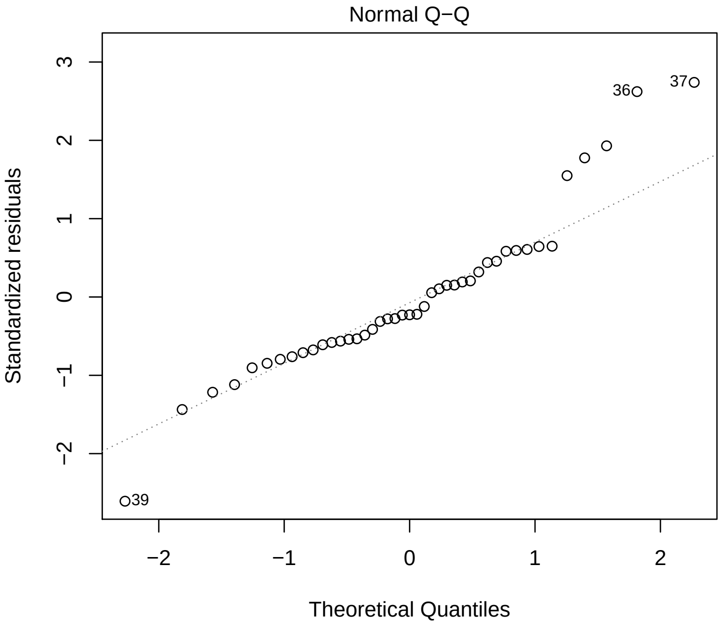
Another Approach: `plot(fit)`



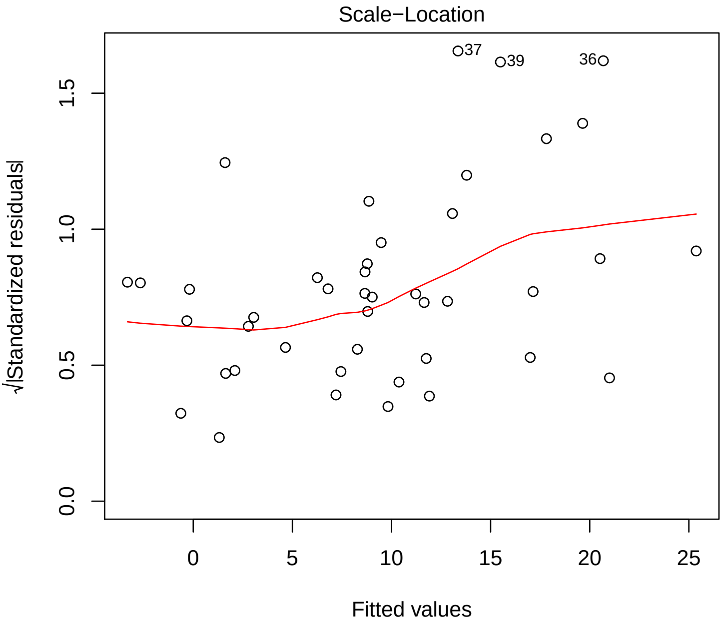
#1: Residuals vs. Fitted Values



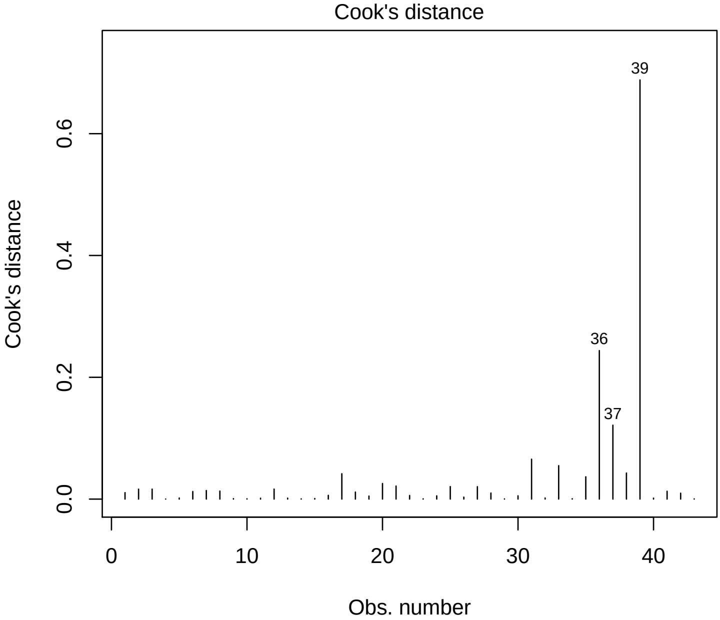
#2: Q-Q Plot of $\hat{u}$s



“Scale-Location” Plot

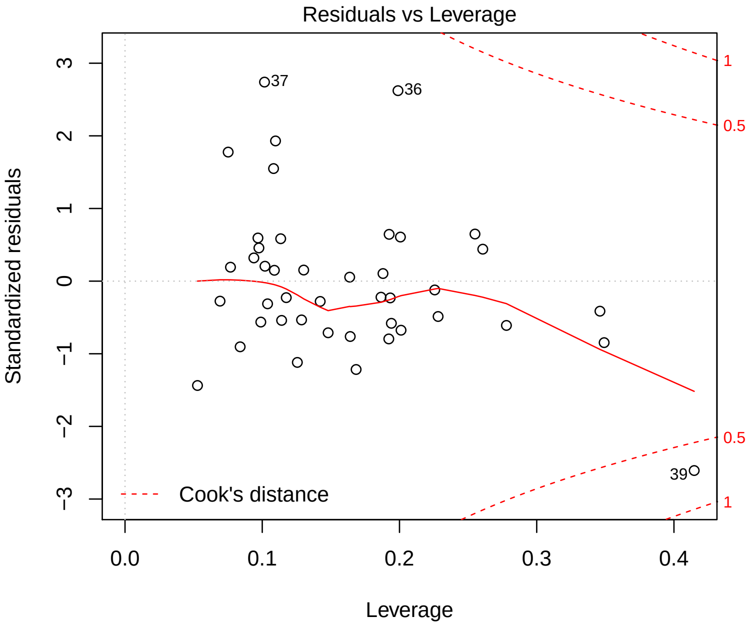


Cook's D

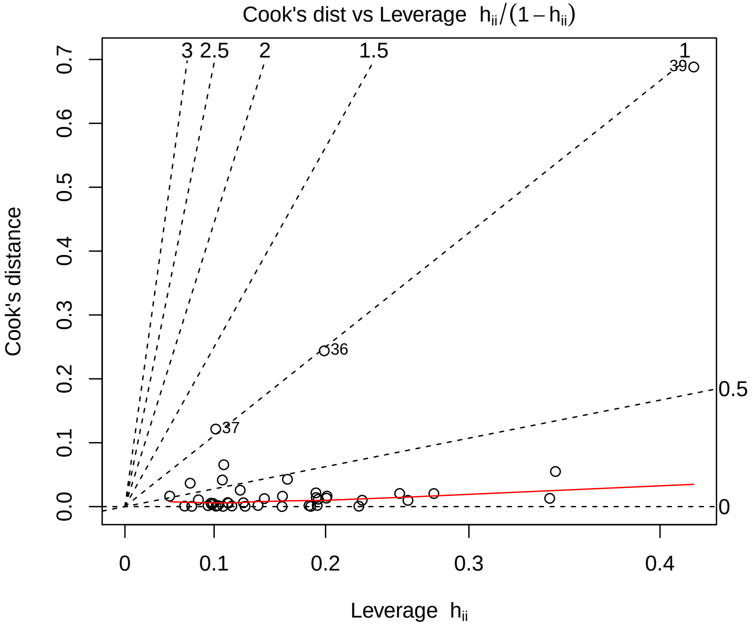


Cook's distance
Cook's distance
Obs. number
0.0
0.2
0.4
0.6
0
10
20
30
40
39
36
37

Residuals vs. Leverage



Cook's D vs. Leverage



Outliers?

```
> ASmall<-cbind(Africa[,3],Fit$model)

> ASmall[c(36,37,39),]
   Africa[, 3] adrate gdppppd muslperc  subsaharan
36    Botswana   38.8     7.8      0.0 Sub-Saharan
37   Swaziland   33.4     4.2     10.0 Sub-Saharan
39   Mauritius    0.1    10.8     16.6 Sub-Saharan

   healthexp literacy internalwar
36       6.6       78           0
37       3.3       80           0
39       3.4       85           0
```

“Variances”

Variances: Why We Care

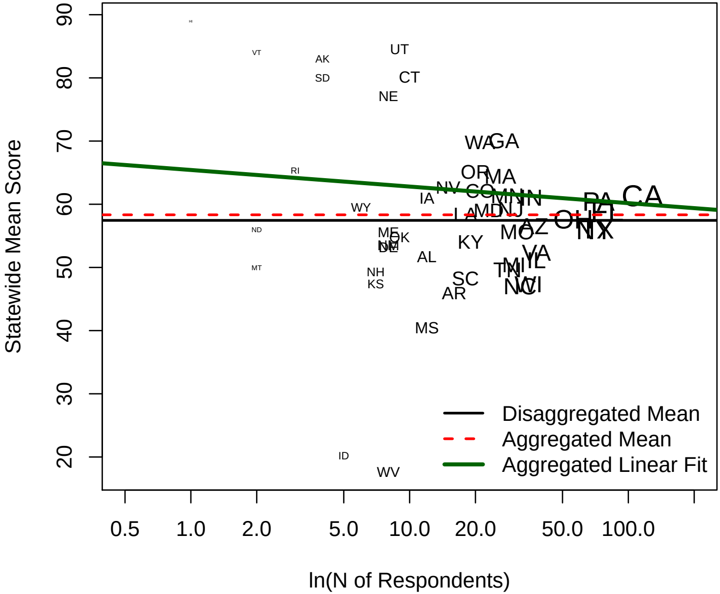
2016 ANES pilot study "feeling thermometer" toward gays and lesbians ($N = 1200$):

```
> summary(ANES$ftgay)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   40.50   54.00   57.45   88.50  100.00       1
```

Suppose we wanted to create aggregate measures, by state ($N = 51$). We would get:

```
> summary(StateFT)
    State               Nresp          meantherm
 Length:50          Min.   :  1.00   Min.   :17.62
 Class :character   1st Qu.:  8.00   1st Qu.:51.33
 Mode  :character   Median : 18.00   Median :57.11
                    Mean   : 24.00   Mean   :58.33
                    3rd Qu.: 30.75   3rd Qu.:62.55
                    Max.   :116.00   Max.   :89.00
```

Variances: Why We Care



Variances: A Generalization

Start with:

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + u_i$$

with:

$$\text{Var}(u_i) = \sigma^2/w_i$$

with w_{iu} <u>known</u>.

Weighted Least Squares

WLS now minimizes:

$$\text{RSS} = \sum_{i=1}^{N} w_i(Y_i - \mathbf{X}_i\boldsymbol{\beta}).$$

which gives:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{WLS} &= [\mathbf{X}'(\sigma^2\boldsymbol{\Omega})^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\sigma^2\boldsymbol{\Omega})^{-1}\mathbf{Y} \\ &= [\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y}\end{aligned}$$

where:

$$\mathbf{W} = \begin{bmatrix} \frac{\sigma^2}{w_1} & 0 & \cdots & 0 \\ 0 & \frac{\sigma^2}{w_2} & \cdots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{\sigma^2}{w_N} \end{bmatrix}$$

The variance-covariance matrix is:

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}_{WLS}) &= \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} \\ &\equiv (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1} \end{aligned}$$

A common case is:

$$\text{Var}(u_i) = \frac{\sigma^2}{N_i}$$

where N_i is the number of observations upon which (aggregate) observation i is based.

"Robust" Variance Estimators

Recall that, if $\sigma_i^2 \neq \sigma_j^2 \; \forall \; i \neq j$,

$$\begin{aligned} \text{Var}(\beta_{\text{Het.}}) &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{Q}\,(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

where $\mathbf{Q} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})$ and $\mathbf{W} = \sigma^2\boldsymbol{\Omega}$.

We can rewrite $\mathbf{Q}$ as

$$\begin{aligned} \mathbf{Q} &= \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}) \\ &= \sum_{i=1}^{N} \sigma_i^2 \mathbf{X}_i \mathbf{X}_i' \end{aligned}$$

Estimate $\hat{\mathbf{Q}}$ as:

$$\widehat{\mathbf{Q}} = \sum_{i=1}^{N} \hat{u}_i^2 \mathbf{X}_i \mathbf{X}_i'$$

Yields:

$$\begin{aligned}
\widehat{\text{Var}(\beta)}_{\text{Robust}} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\widehat{\mathbf{Q}}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\left[\mathbf{X}'\left(\sum_{i=1}^{N} \hat{u}_i^2 \mathbf{X}_i \mathbf{X}_i'\right)^{-1}\mathbf{X}\right](\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

"Robust" VCV estimates:

- are heteroscedasticity-consistent, but
- are biased in small samples, and
- are less efficient than "naive" estimates when $\mathrm{Var}(u) = \sigma^2 \mathbf{I}$.

"Clustering"

Huber / White | **?????????** | WLS / GLS

I know very little about my error variances...

I know a great deal about my error variances...

A common case:

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + u_{ij}$$

with

$$\sigma^2_{ij} = \sigma^2_{ik}.$$

"Robust, clustered" estimator:

$$\widehat{\text{Var}(\boldsymbol{\beta})}_{\text{Clustered}} = (\mathbf{X}'\mathbf{X})^{-1}\left\{\mathbf{X}'\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{n_j}\hat{u}_{ij}^2\mathbf{X}_{ij}\mathbf{X}_{ij}'\right)\right]^{-1}\mathbf{X}\right\}(\mathbf{X}'\mathbf{X})^{-1}$$

Robust / Clustered SEs: A Simulation

```
url_robust <- "https://raw.githubusercontent.com/IsidoreBeautrelet/economictheoryblog/master/robust_summary.R"
eval(parse(text = getURL(url_robust, ssl.verifypeer = FALSE)),
     envir=.GlobalEnv)

> set.seed(7222009)
> X <- rnorm(10)
> Y <- 1 + X + rnorm(10)
> df10 <- data.frame(ID=seq(1:10),X=X,Y=Y)
>
> fit10 <- lm(Y~X,data=df10)
> summary(fit10)

Residuals:
     Min       1Q   Median       3Q      Max
-1.12328 -0.65321 -0.05073  0.43937  1.81661

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8438     0.3020   2.794   0.0234 *
X             0.3834     0.3938   0.974   0.3588
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9313 on 8 degrees of freedom
Multiple R-squared:  0.1059,Adjusted R-squared:  -0.005832
F-statistic: 0.9478 on 1 and 8 DF,  p-value: 0.3588

> rob10 <- vcovHC(fit10,type="HC1")
> sqrt(diag(rob10))
(Intercept)           X
  0.2932735   0.2859552
```

Robust / Clustered SEs: A Simulation (continued)

```
> # "Clone" each observation 100 times
>
> df1K <- df10[rep(seq_len(nrow(df10)), each=100),]
> df1K <- pdata.frame(df1K, index="ID")
>
> fit1K <- lm(Y~X,data=df1K)
> summary(fit1K)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.84383    0.02704   31.20   <2e-16 ***
X            0.38341    0.03526   10.87   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8338 on 998 degrees of freedom
Multiple R-squared:  0.1059,Adjusted R-squared:  0.105
F-statistic: 118.2 on 1 and 998 DF,  p-value: < 2.2e-16

> summary(fit1K, cluster="ID")

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8438     0.2766   3.050  0.00235 **
X             0.3834     0.2697   1.421  0.15551
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8338 on 998 degrees of freedom
Multiple R-squared:  0.1059,Adjusted R-squared:  0.105
F-statistic:  2.02 on 1 and 9 DF,  p-value: 0.1889
```

“Real-Data” 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
```

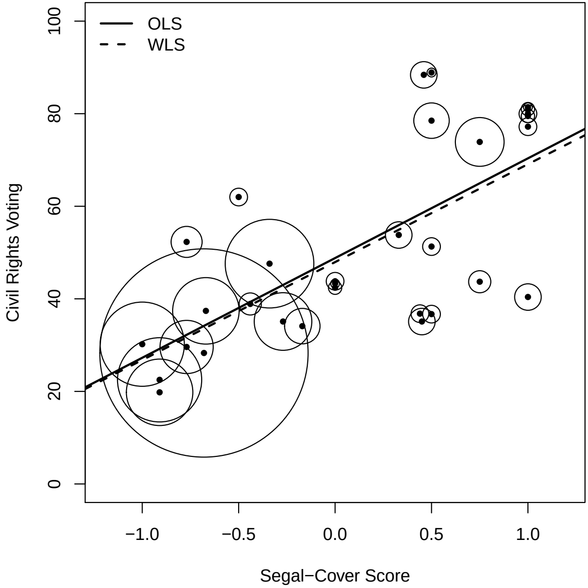
OLS...

```
> 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
```

WLS, Weighting by ln(N of Editorials)

```
> WLSfit<-with(Justices, lm(civrts~score,weights=lnNedit))
> summary(WLSfit)

Call:
lm(formula = civrts ~ score, weights = lnNedit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   47.936      2.600  18.439  < 2e-16 ***
score         21.158      3.797   5.572 5.18e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 19.59 on 29 degrees of freedom
Multiple R-squared: 0.5171,Adjusted R-squared: 0.5004
F-statistic: 31.05 on 1 and 29 DF,  p-value: 5.179e-06
```

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



“Robust” Standard Errors

```
> library(car)
> hccm(OLSfit, type="hc1")
            (Intercept)     score
(Intercept)    6.963921  2.929622
score          2.929622 13.931212


> library(rms)
> OLSfit2<-ols(civrts~score, x=TRUE, y=TRUE)
> RobSEs<-robcov(OLSfit2)
> RobSEs

Linear Regression Model

ols(formula = civrts ~ score, x = TRUE, y = TRUE)

         n Model L.R.       d.f.         R2      Sigma
        31      19.97          1      0.475      15.63

Residuals:
    Min      1Q  Median      3Q     Max
-29.954  -8.088  -2.120   9.396  29.680

Coefficients:
          Value Std. Error      t  Pr(>|t|)
Intercept 48.81      2.552 19.123 0.000e+00
score     21.54      3.610  5.968 1.739e-06

Residual standard error: 15.63 on 29 degrees of freedom
Adjusted R-Squared: 0.4569
```