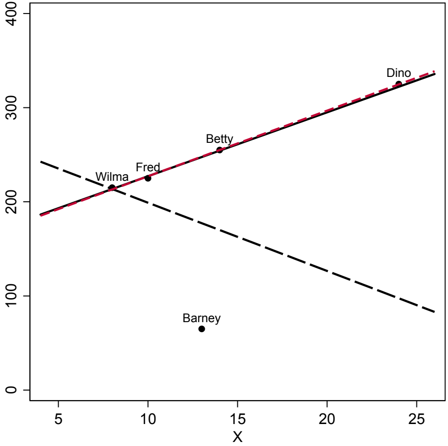
GSERM 2023

Regression for Publishing

June 21, 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

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

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

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

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

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:

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

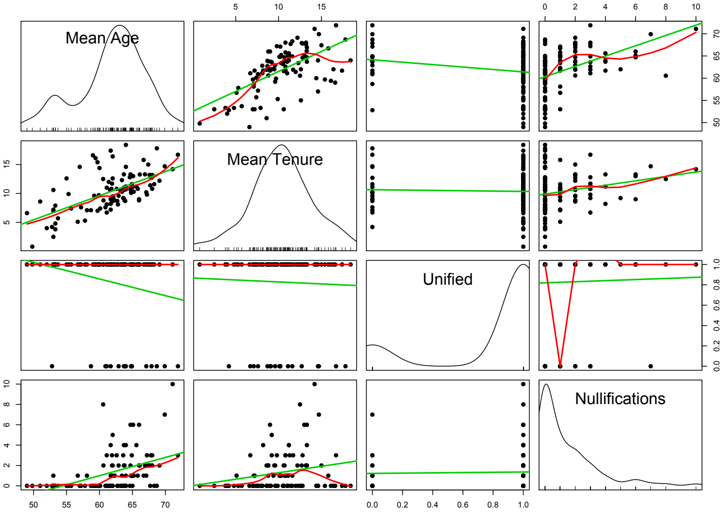
```
> # No Barney OR Dino...
> summary(lm(Y~X,data=subset(flintstones,name!="Dino" & name!="Barney")))

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 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 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":

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

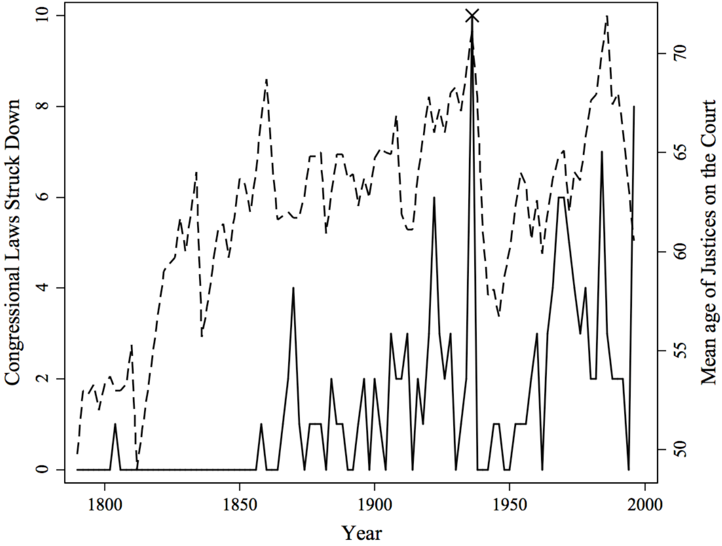
Example: Federal Judicial Review, 1789-1996



A Regression...

```
> 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



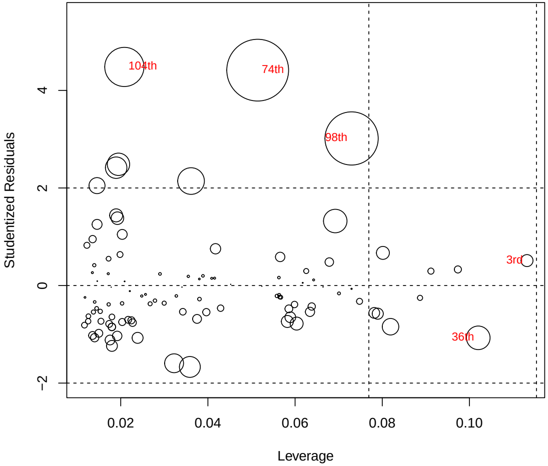
```
> FitResid<-(nulls - predict(Fit)) # residuals
> FitStandard<-rstandard(Fit) # standardized residuals
> FitStudent<-rstudent(Fit) # studentized residuals
> FitCooksD<-cooks.distance(Fit) # Cook's D
> FitDFBeta<-dfbeta(Fit) # DFBeta
> FitDFBetaS<-dfbetas(Fit) # DFBetaS
> FitCOVRATIO<-covratio(Fit)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

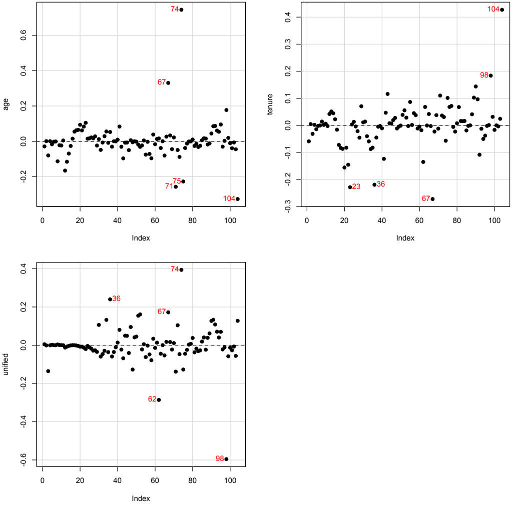
“Bubble Plot”

```
> influencePlot(Fit,id=list(method="noteworthy",n=2,cex=0.8,col="red",
                location="lr",labels=LittleDahl$Congress),fill=FALSE,
                xlab="Leverage",ylim=c(-2,5.5))
```



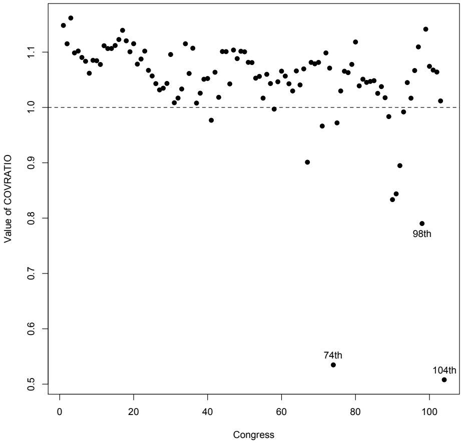
DFBETAS

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



COVRATIO Plot

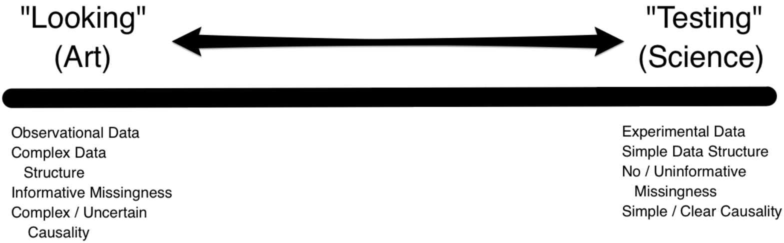
```
> plot(FitCOVRATIO~congress,pch=19,xlab="Congress",ylab="Value of COVRATIO")
> abline(h=1,lty=2)
```



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.

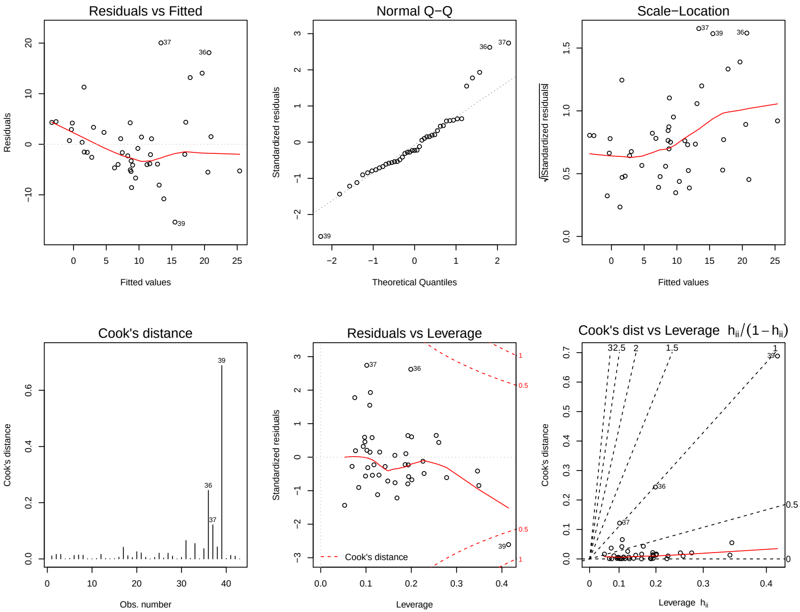
Tests for:

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

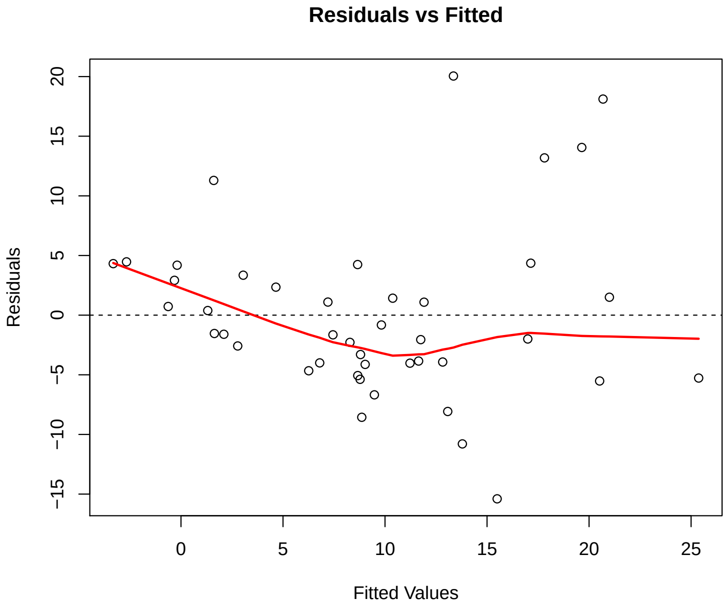
In Action

```
> 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!
```

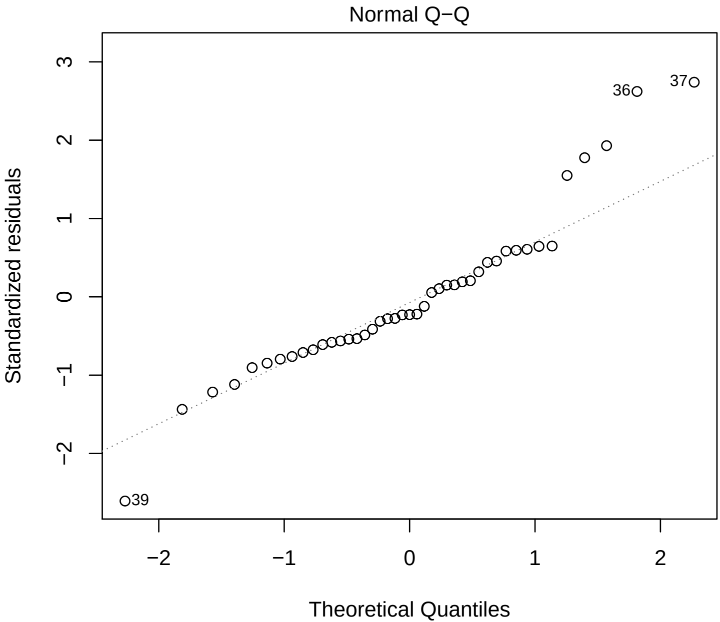
Another Approach: `plot(fit)`



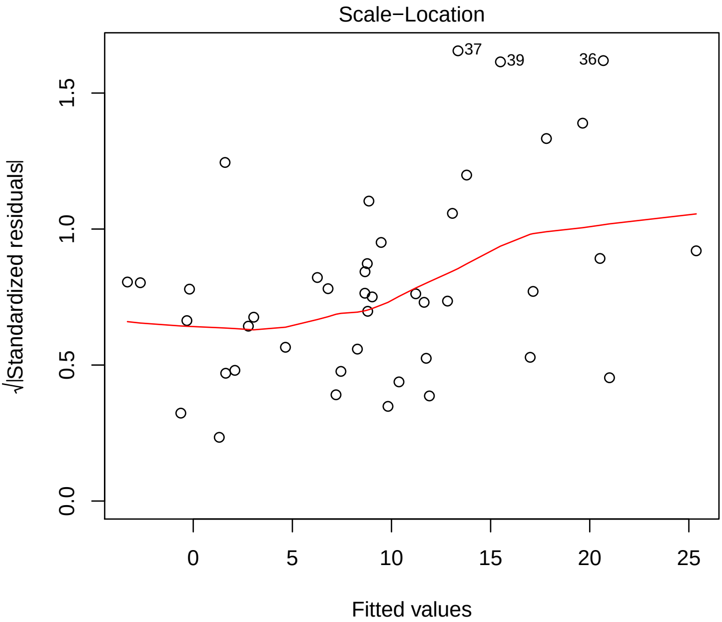
#1: Residuals vs. Fitted Values



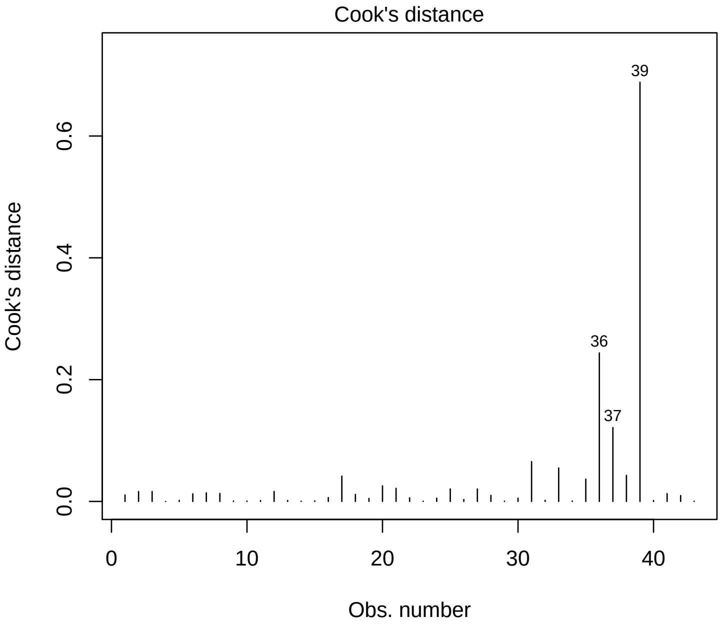
#2: Q-Q Plot of $\hat{u}$s



“Scale-Location” Plot

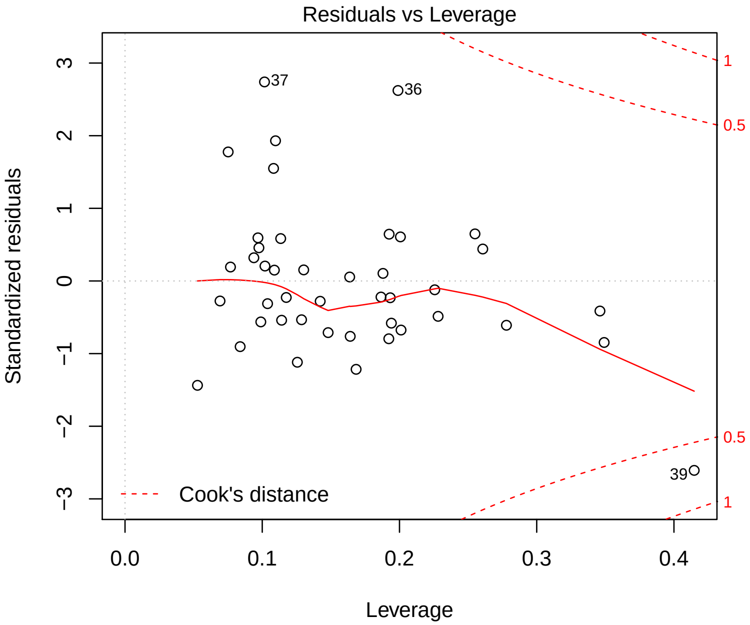


Cook's D

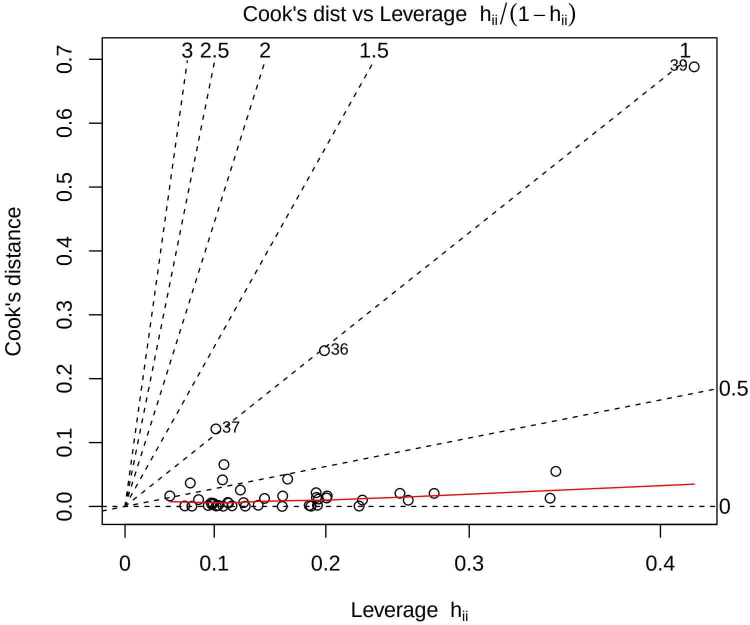


Cook's distance
39
36
37
0.0
0.2
0.4
0.6
Cook's distance
0
10
20
30
40
Obs. number

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

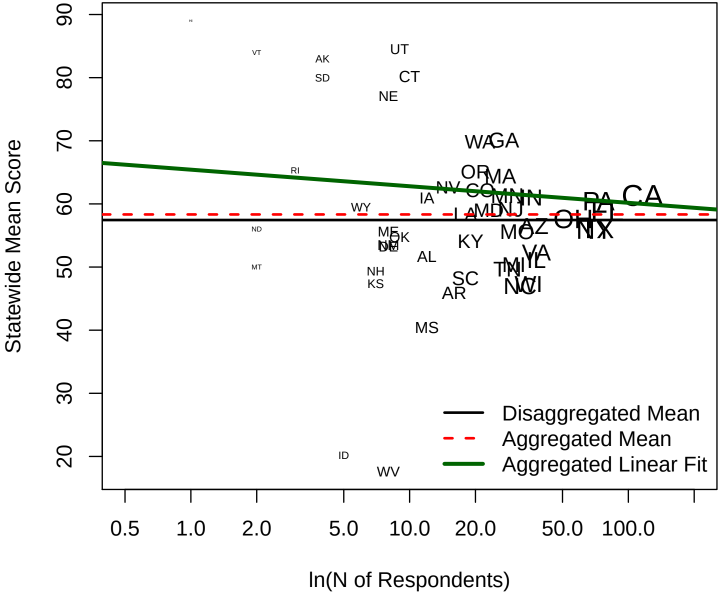
2016 ANES pilot study "feeling thermometer" toward gays and lesbians ($N = 1200$):

```
> 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



Start with:

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

with:

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

with w_{iu} known.

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:

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

A common case is:

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

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 $\text{Var}(u) = \sigma^2 \mathbf{I}$.

“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
```

```
> 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  
```

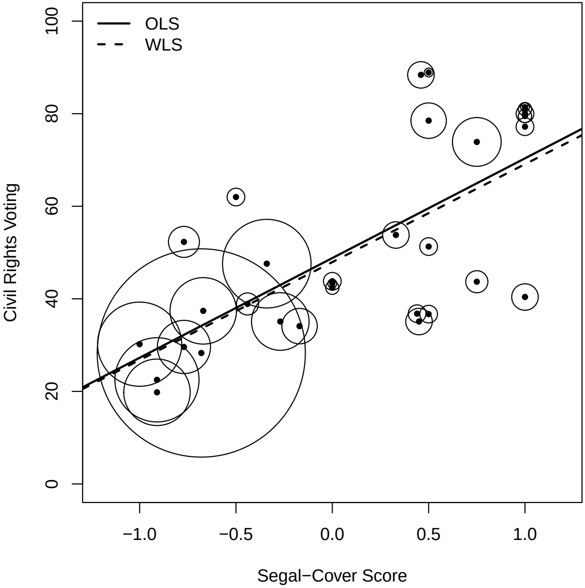
OLS...

```
> 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`



```
> 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
```

Models for Binary Responses

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

$$\begin{aligned} Y_i &= 0 \text{ if } Y_i^* < 0 \\ Y_i &= 1 \text{ if } Y_i^* \geq 0 \end{aligned}$$

So:

$$\begin{aligned} \Pr(Y_i = 1) &= \Pr(Y_i^* \geq 0) \\ &= \Pr(\mathbf{X}_i\boldsymbol{\beta} + u_i \geq 0) \\ &= \Pr(u_i \geq -\mathbf{X}_i\beta) \\ &= \Pr(u_i \leq \mathbf{X}_i\beta) \\ &= \int_{-\infty}^{\mathbf{X}_i\boldsymbol{\beta}} f(u)du \end{aligned}$$

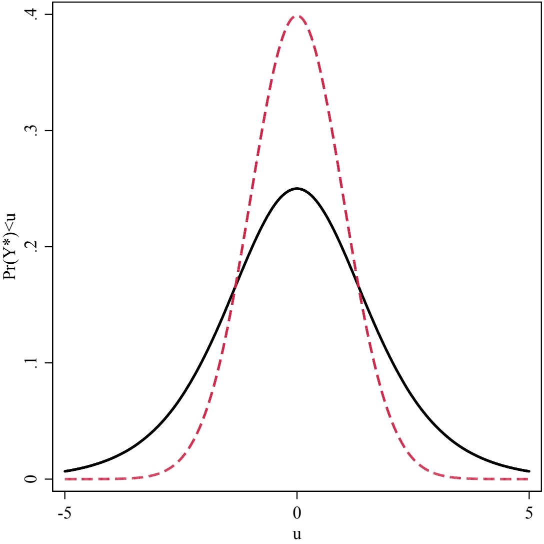
"Standard logistic" PDF:

$$\Pr(u) \equiv \lambda(u) = \frac{\exp(u)}{[1 + \exp(u)]^2}$$

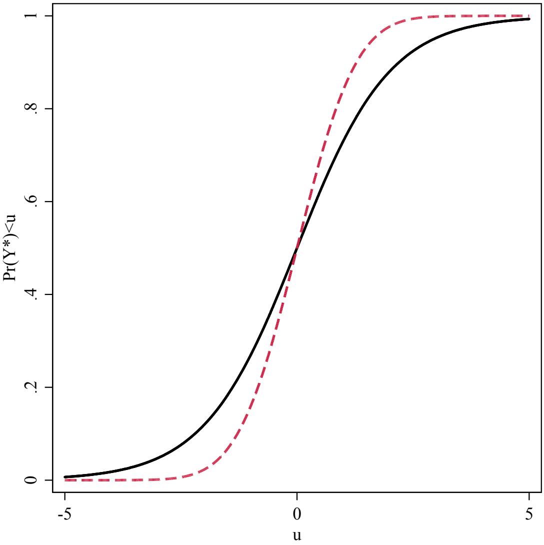
CDF:

$$\begin{aligned}
\Lambda(u) &= \int \lambda(u) du \\
&= \frac{\exp(u)}{1 + \exp(u)} \\
&= \frac{1}{1 + \exp(-u)}
\end{aligned}$$

Standard Normal and Logistic PDFs



Standard Normal and Logistic CDFs



- $\lambda(u) = 1 - \lambda(-u)$
- $\Lambda(u) = 1 - \Lambda(-u)$
- $\mathrm{Var}(u) = \frac{\pi^2}{3} \approx 3.29$

Logistic $\rightarrow$ "Logit"

$$
\begin{aligned}
\Pr(Y_i = 1) &= \Pr(Y_i^* > 0) \\
&= \Pr(u_i \leq \mathbf{X}_i \boldsymbol{\beta}) \\
&= \Lambda(\mathbf{X}_i \boldsymbol{\beta}) \\
&= \frac{\exp(\mathbf{X}_i \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i \boldsymbol{\beta})} \\
\text{(equivalently)} &= \frac{1}{1 + \exp(-\mathbf{X}_i \boldsymbol{\beta})}
\end{aligned}
$$

Likelihoods

$$L_i = \left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)^{Y_i}\left[1-\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]^{1-Y_i}$$

$$L = \prod_{i=1}^{N}\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)^{Y_i}\left[1-\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]^{1-Y_i}$$

$$\begin{aligned}\ln L \quad = \quad & \sum_{i=1}^{N} Y_i \ln\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right) + \\ & (1-Y_i)\ln\left[1-\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]\end{aligned}$$

Digression: Logit as an Odds Model

$$\text{Odds}(Z) \equiv \Omega(Z) = \frac{\Pr(Z)}{1 - \Pr(Z)}.$$

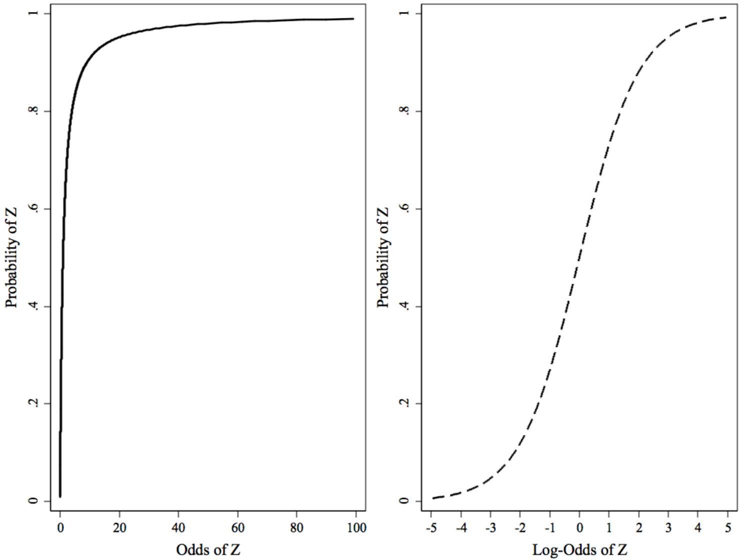
$$\ln[\Omega(Z)] = \ln\left[\frac{\Pr(Z)}{1 - \Pr(Z)}\right]$$

$$\ln[\Omega(Z_i)] = \mathbf{X}_i\boldsymbol{\beta}$$

$$\begin{aligned}\Omega(Z_i) &= \frac{\Pr(Z)}{1 - \Pr(Z)} \\ &= \exp(\mathbf{X}_i\boldsymbol{\beta})\end{aligned}$$

$$\Pr(Z_i) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}$$

Visualizing Log-Odds



Probit: Y Be Normal?

$$\Pr(u) \equiv \phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

$$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$$

Normal → "Probit"

$$\begin{aligned}\Pr(Y_i = 1) &= \Phi(\mathbf{X}_i\boldsymbol{\beta}) \\ &= \int_{-\infty}^{\mathbf{X}_i\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{X}_i\boldsymbol{\beta})^2}{2}\right) d\mathbf{X}_i\boldsymbol{\beta}\end{aligned}$$

$$L = \prod_{i=1}^{N} \left[\Phi(\mathbf{X}_i\boldsymbol{\beta})\right]^{Y_i} \left[1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})\right]^{(1-Y_i)}$$

$$\ln L = \sum_{i=1}^{N} Y_i \ln\Phi(\mathbf{X}_i\boldsymbol{\beta}) + (1 - Y_i) \ln[1 - \Phi(\mathbf{X}_i\boldsymbol{\beta})]$$

Three things:

- Similar in many respects
- $\hat{\beta}_{\text{logit}} \approx \hat{\beta}_{\text{probit}}$, s.e.s are proportional
- Never use probit.

What About Linear Regression?

Linear regression w / binary Y = **"Linear Probability Model" (LPM)**

Various thoughts:

- Issues:
  - Model misspecification $\rightarrow$ bias, inconsistency
  - Creates heteroscedasticity
  - Can yield predicted values outside $(0, 1)$
- The rehabilitation of the LPM:
  - "Logit is hard" / "OLS is awesome" / "It doesn't matter anyway"
  - More-or-less entirely due to (famous) economists
  - Examples: here, here, etc.
- Takeaway: **Pay attention to what people in your discipline / field are doing.**

Example: House Voting on NAFTA

- `vote` – Whether (=1) or not (=0) the House member in question voted in favor of NAFTA.
- `democrat` – Whether the House member in question is a Democrat (=1) or a Republican (=0).
- `pcthispc` – The percentage of the House member's district who are of Latino/hispanic origin.
- `cope93` – The 1993 AFL-CIO (COPE) voting score of the member in question; this variable ranges from 0 to 100, with higher scores indicating more pro-labor positions.
- `DemXCOPE` – The multiplicative interaction of `democrat` and `cope93`.

$$\Pr(\texttt{vote}_i = 1) \quad = \quad f[\beta_0 + \beta_1(\texttt{democrat}_i) + \beta_2(\texttt{pcthispc}_i) + \beta_3(\texttt{cope93}_i) + \beta_4(\texttt{democrat}_i \times \texttt{cope93}_i) + u_i]$$

```
> summary(nafta)
     vote           democrat         pcthispc        cope93          DemXCOPE
 Min.   :0.0000   Min.   :0.0000   Min.   : 0.0   Min.   :  0.00   Min.   :  0.00
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 1.0   1st Qu.: 17.00   1st Qu.:  0.00
 Median :1.0000   Median :1.0000   Median : 3.0   Median : 81.00   Median : 75.00
 Mean   :0.5392   Mean   :0.5853   Mean   : 8.8   Mean   : 60.18   Mean   : 51.65
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:10.0   3rd Qu.:100.00   3rd Qu.:100.00
 Max.   :1.0000   Max.   :1.0000   Max.   :83.0   Max.   :100.00   Max.   :100.00
```

$$\Pr(Y_i = 1) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}$$

or

$$\Pr(Y_i = 1) = \Phi(\mathbf{X}_i\boldsymbol{\beta})$$

Probit Estimates

```
> NAFTA.GLM.probit<-glm(vote~democrat+pcthispc+cope93+DemXCOPE,
  NAFTA,family=binomial(link="probit"))
> summary(NAFTA.GLM.probit)

Call:
glm(formula = vote ~ democrat + pcthispc + cope93 + DemXCOPE,
    family = binomial(link = "probit"))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.07761    0.15339    7.03  2.1e-12 ***
democrat     3.03359    0.73884    4.11  4.0e-05 ***
pcthispc     0.01279    0.00467    2.74   0.0062 **
cope93      -0.02201    0.00440   -5.00  5.8e-07 ***
DemXCOPE    -0.02888    0.00903   -3.20   0.0014 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 598.99  on 433  degrees of freedom
Residual deviance: 441.06  on 429  degrees of freedom
AIC: 451.1
```

Logit Estimates

```
> NAFTA.GLM.logit<-glm(vote~democrat+pcthispc+cope93+DemXCOPE,NAFTA,family=binomial)
> summary(NAFTA.GLM.logit)

Call:
glm(formula = vote ~ democrat + pcthispc + cope93 + DemXCOPE,
    family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79164    0.27544    6.50  7.8e-11 ***
democrat     6.86556    1.54729    4.44  9.1e-06 ***
pcthispc     0.02091    0.00794    2.63  0.00846 **
cope93      -0.03650    0.00760   -4.80  1.6e-06 ***
DemXCOPE    -0.06705    0.01820   -3.68  0.00023 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 598.99  on 433  degrees of freedom
Residual deviance: 436.83  on 429  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 446.8
```

NAFTA Model Results

Probit / Logit / OLS Models of the NAFTA Vote

	NAFTA Vote		
	Probit	Logit	OLS
(Constant)	1.08***	1.79***	0.86***
	(0.15)	(0.28)	(0.04)
Democratic Member	3.03***	6.87***	0.74***
	(0.74)	(1.55)	(0.14)
Hispanic Percent	0.01***	0.02***	0.004***
	(0.005)	(0.01)	(0.001)
COPE Score	−0.02***	−0.04***	−0.01***
	(0.004)	(0.01)	(0.001)
Democratic Member x COPE Score	−0.03***	−0.07***	−0.01***
	(0.01)	(0.02)	(0.002)
Observations	434	434	434
R^2			0.31
Adjusted R^2			0.31
Log Likelihood	−220.53	−218.41	
Akaike Inf. Crit.	451.06	446.83	
Residual Std. Error			0.42 (df = 429)
F Statistic			49.17*** (df = 4; 429)
Note:			*p<0.1; **p<0.05; ***p<0.01

Log-Likelihoods, "Deviance," etc.

- Reports "deviances":
  - "Residual" deviance $= 2(\ln L_S - \ln L_M)$
  - "Null" deviance $= 2(\ln L_S - \ln L_N)$
  - stored in `object$deviance` and `object$null.deviance`
- So:

$$\begin{aligned} LR_{\boldsymbol{\beta}=\mathbf{0}} &= 2(\ln L_M - \ln L_N) \\ &= \text{"Null" deviance} - \text{"Residual" deviance} \end{aligned}$$

```
> NAFTA.GLM.logit$null.deviance - NAFTA.GLM.logit$deviance
[1] 162.1577
```

```
. logit vote democrat pcthispc cope93 DemXCOPE

Logistic regression                               Number of obs   =        434
                                                  LR chi2(4)      =     162.16 <---
                                                  Prob > chi2     =     0.0000
Log likelihood = -218.41388                       Pseudo R2       =     0.2707

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    democrat |   6.865556   1.547357     4.44   0.000     3.832792    9.898319
    pcthispc |   .0209106    .007941     2.63   0.008     .0053466    .0364747
      cope93 |  -.0365007   .0075976    -4.80   0.000    -.0513917   -.0216097
    DemXCOPE |  -.0670544   .0182039    -3.68   0.000    -.1027334   -.0313754
       _cons |    1.79164   .2754383     6.50   0.000     1.251791    2.331489
------------------------------------------------------------------------------
```

Interpretation: "Signs-n-Significance"

For both logit and probit:

- $\hat{\beta}_k > 0 \leftrightarrow \frac{\partial \Pr(Y=1)}{\partial X_k} > 0$
- $\hat{\beta}_k < 0 \leftrightarrow \frac{\partial \Pr(Y=1)}{\partial X_k} < 0$
- $\frac{\hat{\beta}_k}{\hat{\sigma}_k} \sim N(0,1)$

Interactions:

$$\hat{\beta}_{\texttt{cope93}|\texttt{democrat=1}} \equiv \hat{\psi}_{\texttt{cope93}} = \hat{\beta}_3 + \hat{\beta}_4$$

$$\text{s.e.}(\hat{\beta}_{\texttt{cope93}|\texttt{democrat=1}}) = \sqrt{\text{Var}(\hat{\beta}_3) + (\texttt{democrat})^2\text{Var}(\hat{\beta}_4) + 2\,(\texttt{democrat})\,\text{Cov}(\hat{\beta}_3, \hat{\beta}_4)}$$

$\hat{\psi}_{\texttt{cope93}}$ point estimate:

```
> NAFTA.GLM.logit$coeff[4]+ NAFTA.GLM.logit$coeff[5]

    cope93
-0.1035551
```

z-score ("by hand"):

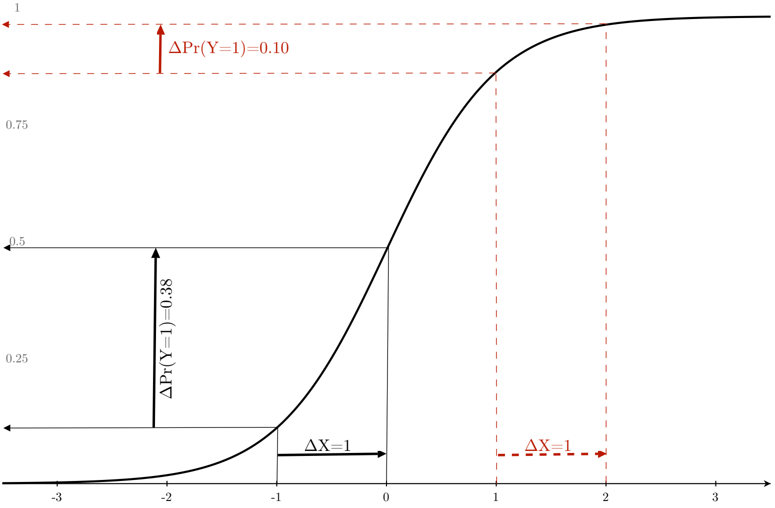
```
> (NAFTA.GLM.logit $coeff[4]+ NAFTA.GLM.logit $coeff[5]) / (sqrt(vcov(NAFTA.GLM.logit)[4,4] +
  (1)^2*vcov(NAFTA.GLM.logit)[5,5] + 2*1*vcov(NAFTA.GLM.logit)[4,5]))

   cope93
-6.245699
```

(Or use car...)

```
> library(car)
> linear.hypothesis(NAFTA.GLM.logit,"cope93+DemXCOPE=0")
Linear hypothesis test

Hypothesis:
cope93 + DemXCOPE = 0

Model 1: vote ~ democrat + pcthispc + cope93 + DemXCOPE
Model 2: restricted model

  Res.Df Df  Chisq Pr(>Chisq)
1    429
2    430 -1 39.009  4.219e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

$$
\begin{aligned}
\widehat{\Pr(Y_i = 1)} &= F(\mathbf{X}_i\hat{\boldsymbol{\beta}}) \\
&= \frac{\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}})} \text{ for logit,} \\
&= \Phi(\mathbf{X}_i\hat{\boldsymbol{\beta}}) \text{ for probit.}
\end{aligned}
$$

Predicted Probabilities Illustrated



Predicted Probabilities: Standard Errors

$$\begin{aligned}\text{Var}[\widehat{\Pr(Y_i = 1)})] &= \left[\frac{\partial F(\mathbf{X}_i\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}}\right]' \hat{\mathbf{V}} \left[\frac{\partial F(\mathbf{X}_i\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}}\right] \\ &= [f(\mathbf{X}_i\hat{\boldsymbol{\beta}})]^2 \mathbf{X}_i'\hat{\mathbf{V}}\mathbf{X}_i\end{aligned}$$

So,

$$\text{s.e.}[\widehat{\Pr(Y_i = 1)})] = \sqrt{[f(\mathbf{X}_i\hat{\boldsymbol{\beta}})]^2 \mathbf{X}_i'\hat{\mathbf{V}}\mathbf{X}_i}$$

$$\hat{\Delta}\Pr(Y=1)_{\mathbf{X}_A \to \mathbf{X}_B} = \frac{\exp(\mathbf{X}_B\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{X}_B\hat{\boldsymbol{\beta}})} - \frac{\exp(\mathbf{X}_A\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{X}_A\hat{\boldsymbol{\beta}})}$$

or

$$= \Phi(\mathbf{X}_B\hat{\boldsymbol{\beta}}) - \Phi(\mathbf{X}_A\hat{\boldsymbol{\beta}})$$

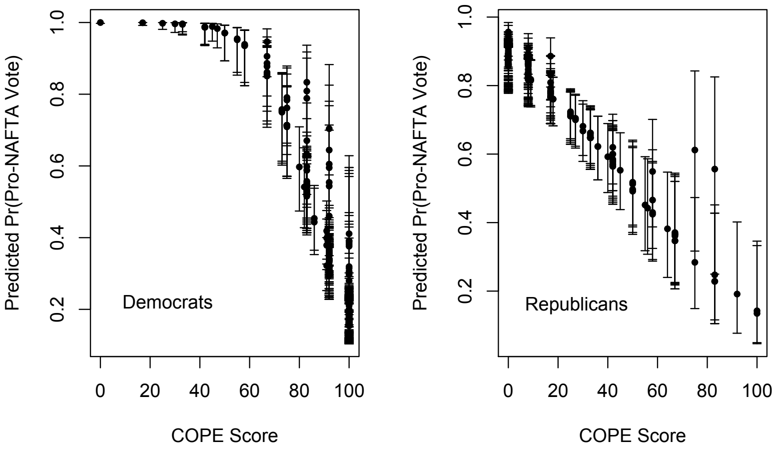
Standard errors obtainable via delta method, bootstrap, etc...

In-Sample Predictions

```
> preds<-NAFTA.GLM.logit$fitted.values

> hats<-predict(NAFTA.GLM.logit,se.fit=TRUE)
> hats
$fit
         1          2          3          4 ...
 9.01267619  7.25223902  6.11013844  5.57444635 ...
 ...
 $se.fit
        1         2         3         4 ...
1.5331506 1.2531475 1.1106989 0.9894208 ...


> XBUB<-hats$fit + (1.96*hats$se.fit)
> XBLB<-hats$fit - (1.96*hats$se.fit)
> plotdata<-cbind(as.data.frame(hats),XBUB,XBLB)
> plotdata<-data.frame(lapply(plotdata,binomial(link="logit")$linkinv))
```

```
...
> par(mfrow=c(1,2))
> library(plotrix)
> plotCI(cope93[democrat==1],plotdata$fit[democrat==1],
  ui=plotdata$XBUB[democrat==1],li=plotdata$XBLB[democrat==1],pch=20,
  xlab="COPE Score",ylab="Predicted Pr(Pro-NAFTA Vote)")
> text(locator(1),label="Democrats")
> plotCI(cope93[democrat==0],plotdata$fit[democrat==0],
  ui=plotdata$XBUB[democrat==0],li=plotdata$XBLB[democrat==0],pch=20,
  xlab="COPE Score",ylab="Predicted Pr(Pro-NAFTA Vote)")
> text(locator(1),label="Republicans")
```

In-Sample Predictions



"Fake" data:

```
> sim.data<-data.frame(pcthispc=mean(nafta$pcthispc),democrat=rep(0:1,101),
  cope93=seq(from=0,to=100,length.out=101))
> sim.data$DemXCOPE<-sim.data$democrat*sim.data$cope93
```

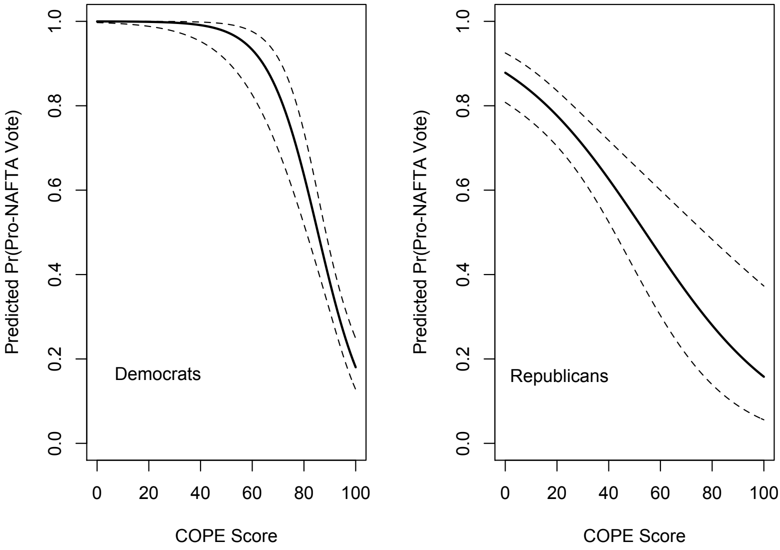
Generate predictions:

```
> OutHats<-predict(NAFTA.GLM.logit,se.fit=TRUE,newdata=sim.data)
> OutHatsUB<-OutHats$fit+(1.96*OutHats$se.fit)
> OutHatsLB<-OutHats$fit-(1.96*OutHats$se.fit)
> OutHats<-cbind(as.data.frame(OutHats),OutHatsUB,OutHatsLB)
> OutHats<-data.frame(lapply(OutHats,binomial(link="logit")$linkinv))
```

Plotting...

```
> par(mfrow=c(1,2))
> both<-cbind(sim.data,OutHats)
> both<-both[order(both$cope93,both$democrat),]

> plot(both$cope93[democrat==1],both$fit[democrat==1],t="l",lwd=2,ylim=c(0,1),
  xlab="COPE Score",ylab="Predicted Pr(Pro-NAFTA Vote)")
> lines(both$cope93[democrat==1],both$OutHatsUB[democrat==1],lty=2)
> lines(both$cope93[democrat==1],both$OutHatsLB[democrat==1],lty=2)
> text(locator(1),label="Democrats")

> plot(both$cope93[democrat==0],both$fit[democrat==0],t="l",lwd=2,ylim=c(0,1),
  xlab="COPE Score",ylab="Predicted Pr(Pro-NAFTA Vote)")
> lines(both$cope93[democrat==0],both$OutHatsUB[democrat==0],lty=2)
> lines(both$cope93[democrat==0],both$OutHatsLB[democrat==0],lty=2)
> text(locator(1),label="Republicans")
```

Out-of-Sample Predictions



Odds Ratios

$$\ln \Omega(\mathbf{X}) = \ln \left[\frac{\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}}{1 - \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}} \right] = \mathbf{X}\beta$$

$$\frac{\partial \ln \Omega}{\partial \mathbf{X}} = \boldsymbol{\beta}$$

Means:

$$\frac{\Omega(X_k + 1)}{\Omega(X_k)} = \exp(\hat{\beta}_k)$$

More generally,

$$\frac{\Omega(X_k + \delta)}{\Omega(X_k)} = \exp(\hat{\beta}_k \delta)$$

$$\text{Percentage Change} = 100[\exp(\hat{\beta}_k \delta) - 1]$$

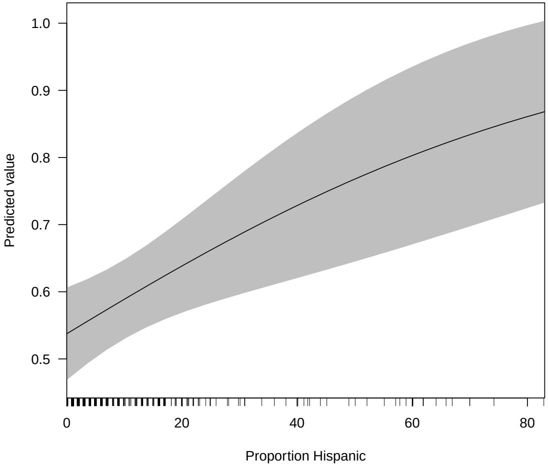
```
> lreg.or <- function(model)
+           {
+           coeffs <- coef(summary(NAFTA.GLM.logit))
+           lci <- exp(coeffs[ ,1] - 1.96 * coeffs[ ,2])
+           or <- exp(coeffs[ ,1])
+           uci <- exp(coeffs[ ,1] + 1.96 * coeffs[ ,2])
+           lreg.or <- cbind(lci, or, uci)
+           lreg.or
+           }

> lreg.or(NAFTA.GLM.fit)
                lci       or       uci
(Intercept)  3.4966   5.9993 1.029e+01
democrat    46.1944 958.6783 1.990e+04
pcthispc     1.0054   1.0211 1.037e+00
cope93       0.9499   0.9642 9.786e-01
DemXCOPE     0.9024   0.9351 9.691e-01
```

Example text:

- "A one percent increase in the percent Hispanic in a district is associated with a $\{[\exp(1 \times 0.021) = 1.0054 - 1] \times 100 =\}$ 0.5 percent *increase* in the odds of that member's support for NAFTA."
- "A ten percent increase in the percent Hispanic in a district is associated with a $\{[\exp(10 \times 0.021) = 1.234 - 1] \times 100 =\}$ 23.4 percent *increase* in the odds of that member's support for NAFTA."
- "*Among Republicans*, one percent increase in a member's COPE score is associated with a $\{[\exp(1 \times -0.036) = 0.965 - 1] \times 100 =\}$ 3.5 percent *decrease* in the odds of that member's support for NAFTA."

Single-Variable Example (using `cplot`)

```
> cplot(NAFTA.fit,"PropHisp",xlab="Proportion Hispanic")
```



- **Proportional reduction in error (PRE)**
- Pseudo-R^2,
- ROC curves.

lation, etc.

PRE:

$$\text{PRE} = \frac{N_{MC} - N_{NC}}{N - N_{NC}}$$

- N_{NC} = number correct under the "null model,"
- N_{MC} = number correct under the estimated model,
- N = total number of observations.

```
> table(NAFTA$vote)

  0   1
200 234


> table(NAFTA.GLM.logit$fitted.values>0.5,nafta$vote==1)

        FALSE TRUE
  FALSE   148   49
  TRUE     52  185
```

$$
\begin{aligned}
\text{PRE} &= \frac{N_{MC} - N_{NC}}{N - N_{NC}} \\
&= \frac{(148 + 185) - 234}{434 - 234} \\
&= \frac{99}{200} \\
&= \mathbf{0.495}
\end{aligned}
$$

Example text:

"The model yielded a 49.5 percent proportional reduction in in-sample prediction error."

Concepts:

- *Sensitivity* (or "true positive rate")
  - The proportion of all actual positives that were predicted correctly
  - Sensitivity $= \frac{\text{TP}}{\text{TP} + \text{FN}}$

- *Specificity* (or "true negative rate")
  - The proportion of all actual negatives that were predicted correctly
  - Specificity $= \frac{\text{TN}}{\text{TN} + \text{FP}}$

- False positive rate $= 1-$*Specificity*
- False negative rate $= 1-$*Sensitivity*

Varying τ

Suppose we set $\tau = 0.00001$. Then:

- We would essentially *always* predict $\hat{Y}_i = 1$, which means
- ...we would always correctly predict all the actual positives (maximize TPs), but
- ...we'd also always get every actual negative wrong (maximize FPs).

Similarly, if we set $\tau = 0.99999$. Then:

- We would essentially *always* predict $\hat{Y}_i = 0$, which means
- ...we would always correctly predict all the actual negatives (maximize TNs), but
- ...also always get every actual positive wrong (maximize FNs).

Values of τ between the extremes trade off true positives for false positives; as τ increases, we have fewer of the former and more of the latter.

NAFTA Examples

```
> # Tau = 0.2:

> Hats02<-ifelse(NAFTA.fit$fitted.values>0.2,1,0)
> CrossTable(NAFTA$Vote,Hats02,prop.r=FALSE,prop.c=FALSE,
             prop.t=FALSE,prop.chisq=FALSE)

             | Hats02
  NAFTA$Vote |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |        96 |       104 |       200 |
-------------|-----------|-----------|-----------|
           1 |         1 |       233 |       234 |
-------------|-----------|-----------|-----------|
Column Total |        97 |       337 |       434 |
-------------|-----------|-----------|-----------|

TPR = 233/234 = 0.996
FPR = 104/200 = 0.520


> # Tau = 0.8:

> Hats08<-ifelse(NAFTA.fit$fitted.values>0.8,1,0)
> CrossTable(NAFTA$Vote,Hats08,prop.r=FALSE,prop.c=FALSE,
             prop.t=FALSE,prop.chisq=FALSE)

             | Hats08
  NAFTA$Vote |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |       178 |        22 |       200 |
-------------|-----------|-----------|-----------|
           1 |       123 |       111 |       234 |
-------------|-----------|-----------|-----------|
Column Total |       301 |       133 |       434 |
-------------|-----------|-----------|-----------|

TPR = 111/234 = 0.474
FPR = 178/200 = 0.890
```

“Receiver Operating Characteristic” (ROC) Curves

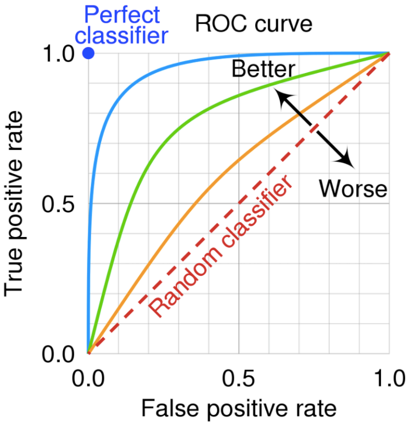
Now, imagine:

1. Fit a model
2. Choose a value of τ very near 0
3. Generate $\hat{Y}_i$s
4. Calculate and save the TPR and FPR for that value of τ
5. Increase τ by a very small amount
6. Go to (3), and repeat until τ is very close to 1.0

We could then plot the true positive rate vs. false positive rate (i.e., *Specificity* vs. 1 - *Sensitivity*)

ROC Curves (continued)

- If the model fits perfectly, it will have a 1.0 true positive rate, and a 0.0 false negative rate
- If the model fits no better than random chance, the curve defined by those points will be a diagonal line.
- (Intuition: If each prediction is no better than a (weighted) coing flip, the rate of true positives and false positives will increase together.)
- In between these extremes, better-fitting models will have curves that are closer to the upper-left corner



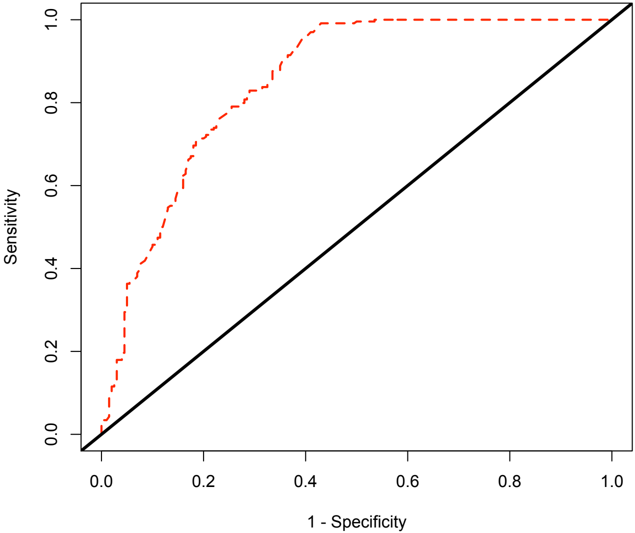
(Source)

"AUROC": Area under the ROC curve
$\rightarrow$ assessment of model fit

ROC Curves Implemented

```
> library(ROCR)

> NAFTA.hats<-predict(NAFTA.fit,type="response")

> preds<-ROCR::prediction(NAFTA.hats,NAFTA$Vote)

> plot(performance(preds,"tpr","fpr"),lwd=2,lty=2,
       col="red",xlab="1 - Specificity",ylab="Sensitivity")
> abline(a=0,b=1,lwd=3)
```

ROC Curve: Example



Interpreting AUROC Curves

- Area under ROC $= 0.90\text{-}1.00 \rightarrow$ Excellent (A)
- Area under ROC $= 0.80\text{-}0.90 \rightarrow$ Good (B)
- Area under ROC $= 0.70\text{-}0.80 \rightarrow$ Fair (C)
- Area under ROC $= 0.60\text{-}0.70 \rightarrow$ Poor (D)
- Area under ROC $= 0.50\text{-}0.60 \rightarrow$ Total Failure (F)

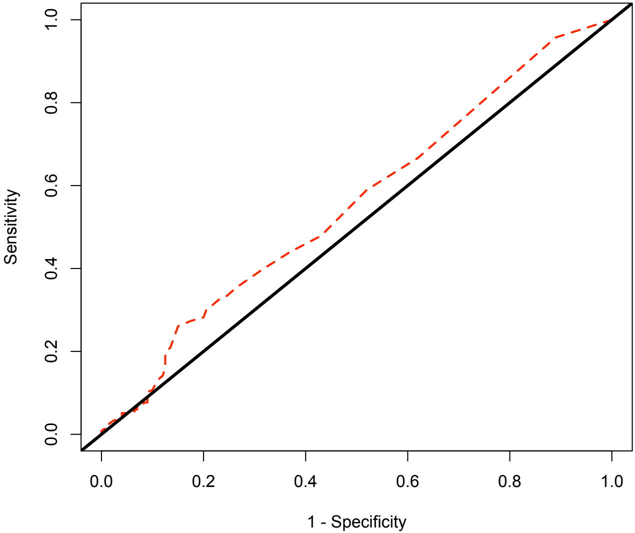
ROC Curve: A Poorly-Fitting Model

Model is:

$$\Pr(\texttt{vote}_i = 1) \quad = \quad f[\beta_0 + \beta_1(\texttt{PropHisp}_i) + u_i]$$

```
> NAFTA.bad<-with(NAFTA,
                  glm(Vote~PropHisp,family=binomial(link="logit")))
> NAFTA.bad.hats<-predict(NAFTA.bad,type="response")
> bad.preds<-ROCR::prediction(NAFTA.bad.hats,NAFTA$Vote)

> plot(performance(bad.preds,"tpr","fpr"),lwd=2,lty=2,
       col="red",xlab="1 - Specificity",ylab="Sensitivity")
> abline(a=0,b=1,lwd=3)
```

Bad ROC!



```
> install.packages("pROC")
> library(pROC)

> GoodROC<-roc(NAFTA$Vote,NAFTA.hats,ci=TRUE)
> GoodAUC<-auc(GoodROC)
> BadROC<-roc(NAFTA$Vote,NAFTA.bad.hats,ci=TRUE)
> BadAUC<-auc(BadROC)

> GoodAUC
Area under the curve: 0.85

> BadAUC
Area under the curve: 0.556
```

Combined Plot