GSERM 2020
Regression for Publishing

June 17, 2020 (second session)

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

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

So:

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

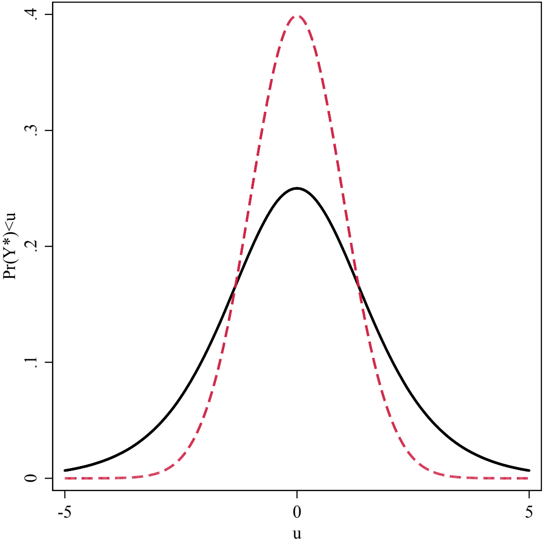
"Standard logistic" PDF:

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

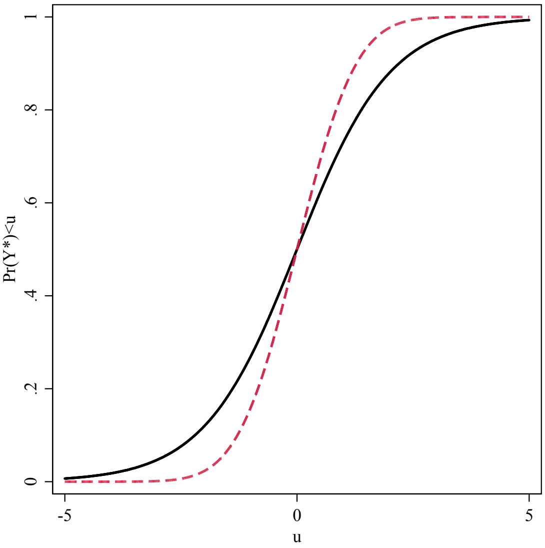
CDF:

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

Standard Normal and Logistic PDFs



Standard Normal and Logistic CDFs



Characteristics

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

Logistic → "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

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

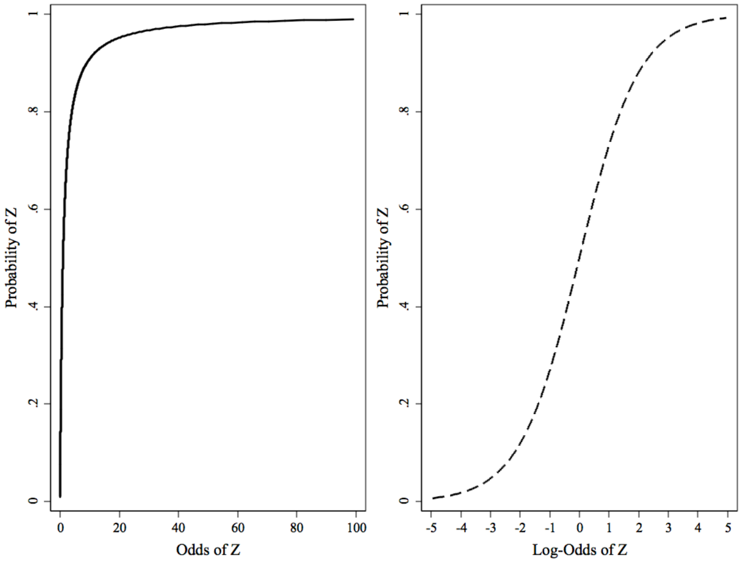
$$\ln[\Omega(Z)] = \ln\left[\frac{\Pr(Z)}{1 - \Pr(Z)}\right]$$

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

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

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

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"

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

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

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

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
```

Table: Probit and Logit Models of the NAFTA Vote

	NAFTA Vote	
	probit Probit	*logistic* Logit
	(1)	(2)
(Constant)	1.08^{***}	1.79^{***}
	(0.15)	(0.28)
Democratic Member	3.03^{***}	6.87^{***}
	(0.74)	(1.55)
Hispanic Percent	0.01^{***}	0.02^{***}
	(0.005)	(0.01)
COPE Score	-0.02^{***}	-0.04^{***}
	(0.004)	(0.01)
Democratic Member x COPE Score	-0.03^{***}	-0.07^{***}
	(0.01)	(0.02)
Observations	434	434
Log Likelihood	-220.53	-218.41
Akaike Inf. Crit.	451.06	446.83
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
------------------------------------------------------------------------------
```

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
```

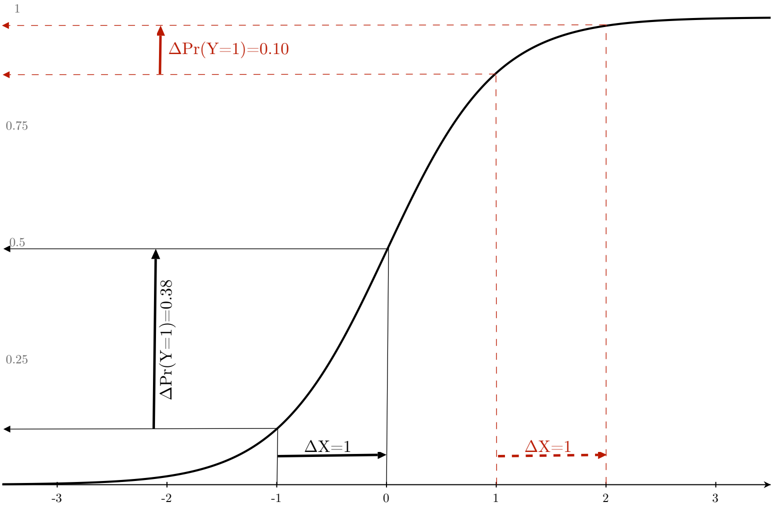
Predicted Probabilities

$$\widehat{\Pr(Y_i = 1)} = F(\mathbf{X}_i\hat{\boldsymbol{\beta}})$$

$$= \frac{\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}})} \text{ for logit,}$$

$$= \Phi(\mathbf{X}_i\hat{\boldsymbol{\beta}}) \text{ for probit.}$$

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,

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

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

or

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

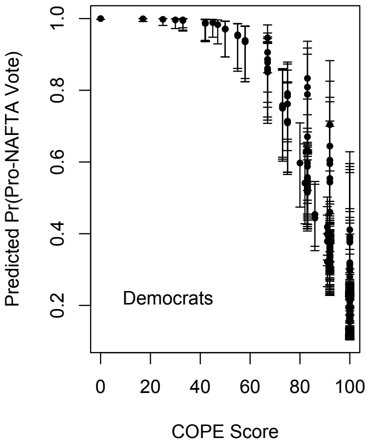
Standard errors obtainable via delta method, bootstrap, etc...

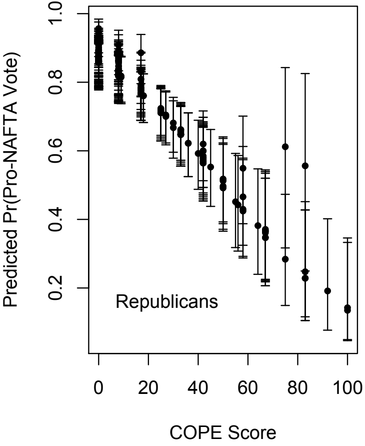
In-Sample Predictions

```
> 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





Out-of-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
```

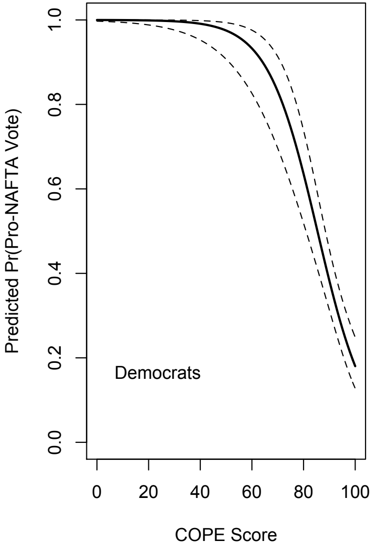
Generate predictions:

```
> 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



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

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

Means:

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

More generally,

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

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

Odds Ratios Implemented

```
> 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."

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

Proportional Reduction in Error

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.

PRE: Example

```
> 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."