PLSC 503 – Spring 2020
Binary Response Models, II

April 9, 2020

Example: House Voting on NAFTA (1993)

Response / Outcome

- `vote` – Whether (=1) or not (=0) the House member in question voted in favor of NAFTA.

Predictors

- `pcthispc` – The percentage of the House member's district who are of Latino/hispanic origin.
- `democrat` – Whether the House member in question is a Democrat (=1) or a Republican (=0).
- `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`.

Model & Data

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

```
> 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
```

Basic Model(s)

Logit:

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

or probit:

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

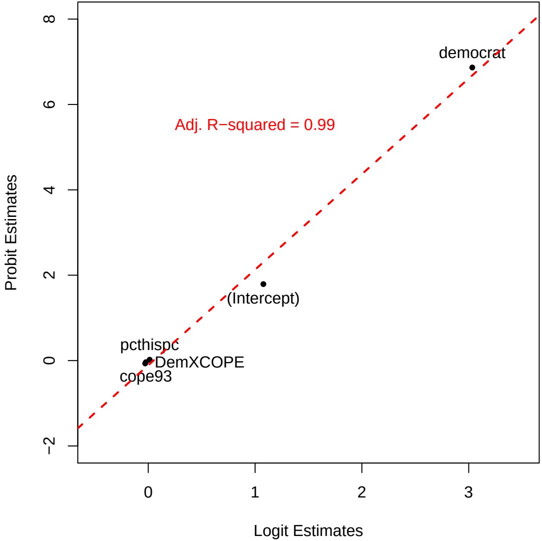
Probit Estimates

```
> NAFTA.GLM.probit<-glm(vote~democrat+pcthispc+cope93+DemXCOPE,
  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,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
```

$\hat{\beta}_{\text{probit}}$ vs. $\hat{\beta}_{\text{logit}}$



Log-Likelihoods, "Deviance," etc.

- `R` / `lm` 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
```

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:

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

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

$\hat{\phi}_{\texttt{cope93}}$ 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
```

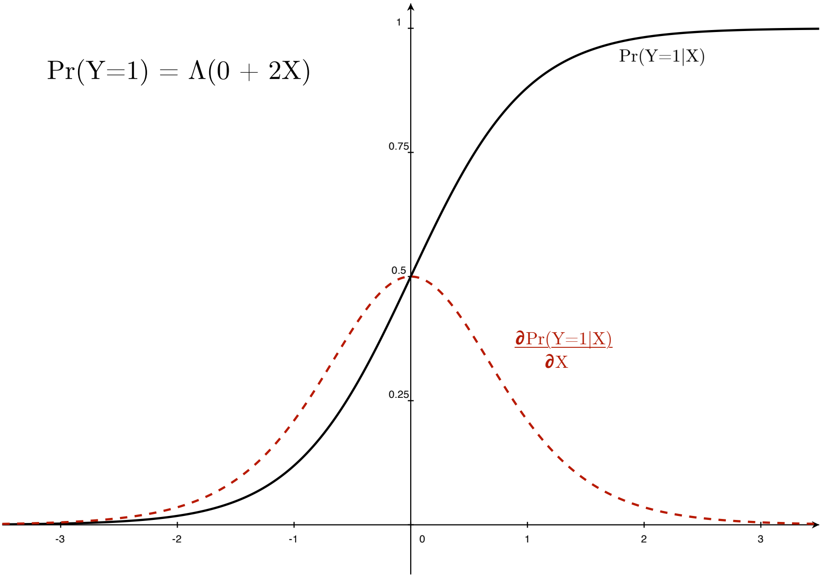
(Or use car...)

```
> 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
```

Marginal Effects

$$\begin{aligned}
\frac{\partial \text{Pr}(\hat{Y}_i = 1)}{\partial X_k} &= \frac{\partial F(\mathbf{X}_i \hat{\boldsymbol{\beta}})}{\partial X_k} \\
&= f(\mathbf{X}_i \hat{\boldsymbol{\beta}}) \hat{\beta}_k \\
\\
&= \Lambda(\mathbf{X}_i \hat{\boldsymbol{\beta}})[1 - \Lambda(\mathbf{X}_i \hat{\boldsymbol{\beta}})] \hat{\beta}_k \quad \text{(logit) or} \\
\\
&= \phi(\mathbf{X}_i \hat{\boldsymbol{\beta}}) \hat{\beta}_k \quad \text{(probit)}
\end{aligned}$$

Marginal Effects Illustrated



Odds Ratios

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

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

Means:

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

More generally,

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

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

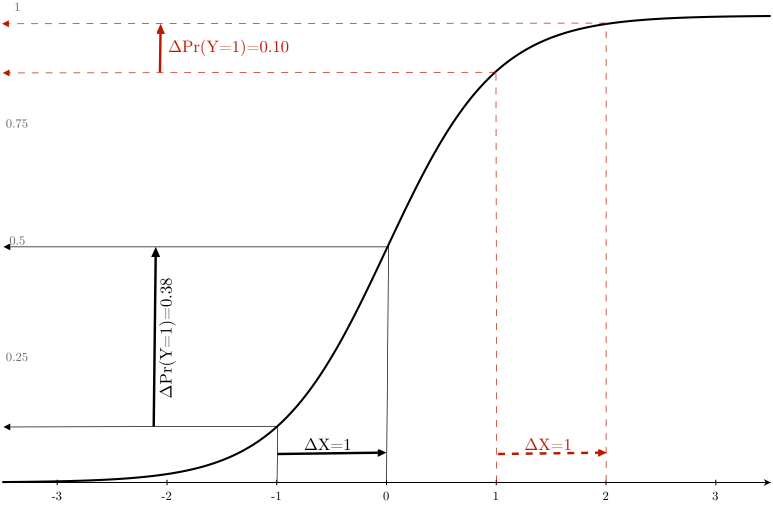
Odds Ratios Implemented

```
> 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
```

Predicted Probabilities

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

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,

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

Probability Changes

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

or

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

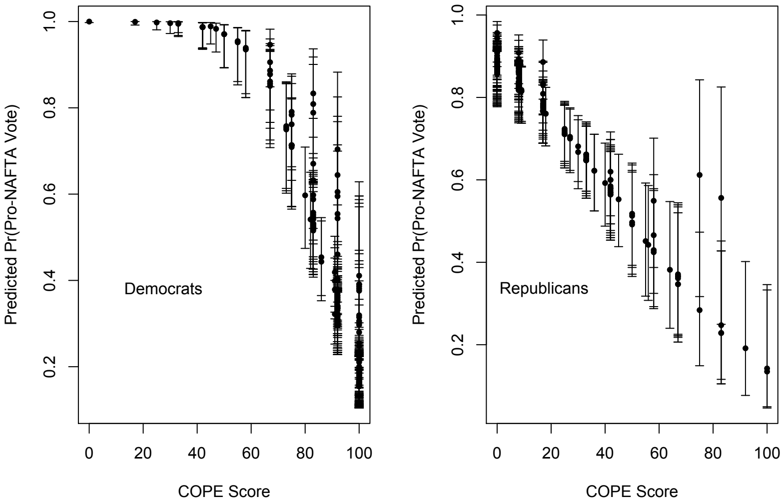
Standard errors obtainable via delta method, bootstrap, etc...

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



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
```

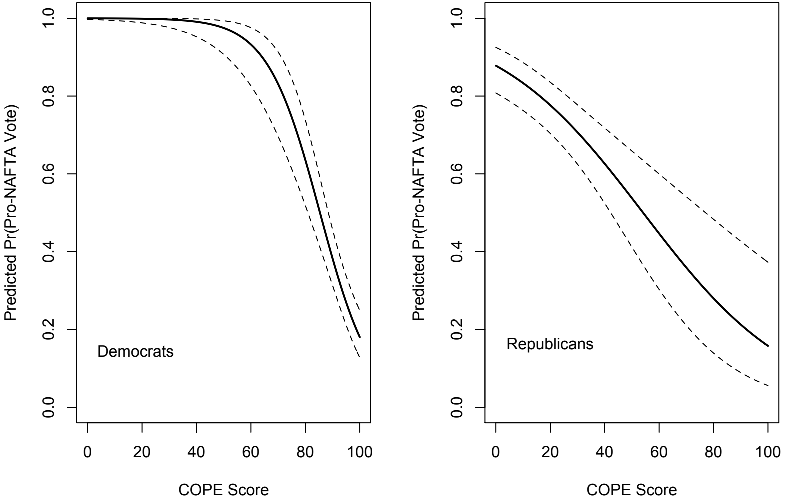
Generate predictions:

```
> 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



Goodness-of-Fit

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

Model Fit: 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.GLM.logit$fitted.values>0.5,nafta$vote==1)

       FALSE TRUE
  FALSE   148   49
  TRUE     52  185
```

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

Chi-Square test:

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

Pearson's Chi-squared test with Yates' continuity correction

data:  NAFTA.GLM.logit$fitted.values > 0.5 and nafta$vote == 1
X-squared = 120.3453, df = 1, p-value < 2.2e-16
```

Related Ideas

- *Sensitivity*
  - $\Pr\widehat{(Y=1)}|Y=1$
  - "true positives"

- *Specificity*
  - $\Pr\widehat{(Y=0)}|Y=0$
  - "true negatives"

- $1-$*Specificity* $=$ "false positives"
- $1-$*Sensitivity* $=$ "false negatives"

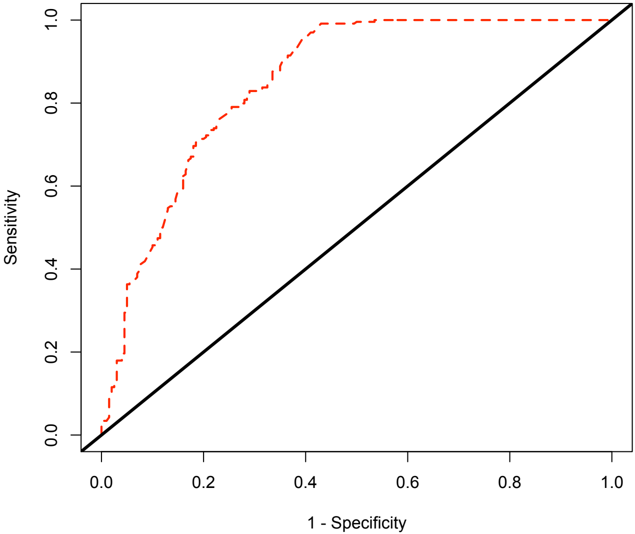
"Receiver Operating Characteristic" (ROC) Curves

- Plot: true positive rate vs. false positive rate (i.e., specificity vs. 1 - sensitivity)
- "aROC": Area under the curve
- $\rightarrow$ assessment of model fit

ROC Curves Implemented

```
> library(ROCR)

> NAFTA.GLM.logithats<-predict(NAFTA.GLM.logit,
+   type="response")

> preds<-prediction(NAFTA.GLM.logithats,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



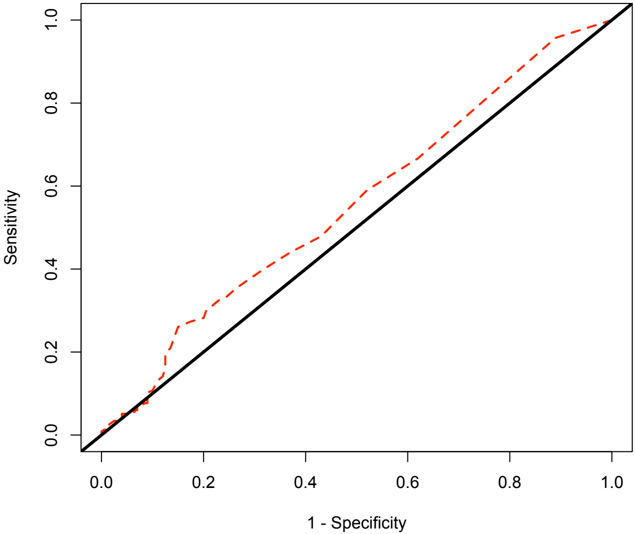
Interpreting ROC Curves

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

ROC Curve: A Poorly-Fitting Model

```
> NAFTA.bad<-glm(vote~pcthispc,family=binomial(link="logit"))
> NAFTA.bad.hats<-predict(NAFTA.bad,type="response")
> bad.preds<-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!



Comparing ROCs

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

> GoodROC<-roc(nafta$vote,NAFTA.GLM.logithats,ci=TRUE)
> GoodAUC<-auc(GoodROC)
> BadROC<-roc(nafta$vote,NAFTA.bad.hats)
> BadAUC<-auc(BadROC)
> GoodAUC
Area under the curve: 0.85
> BadAUC
Area under the curve: 0.556
```

Combined Plot