PLSC 503 – Spring 2022
Binary Response Models, II

April 6, 2022

Running Example: House Vote on NAFTA (1993)

Response / Outcome

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

Predictors

- PropHisp – The proportion 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 ($=0$).
- COPE – The 1993 AFL-CIO (COPE) voting score of the member in question; the original variable ranges from 0 to 100, with higher scores indicating more pro-labor positions. Rescaled to range from 0 to 1.
- DemXCOPE – The multiplicative interaction of Democrat and COPE.

$$\begin{aligned}\Pr(\text{vote}_i = 1) \quad = \quad & f[\beta_0 + \beta_2(\texttt{PropHisp}_i) + \beta_e(\texttt{Democrat}_i) + \\ & \beta_3(\texttt{COPE}_i) + \beta_4(\texttt{Democrat}_i \times \texttt{COPE}_i) + u_i]\end{aligned}$$

```
> summary(NAFTA)

      Vote          PropHisp        Democrat          COPE          DemXCOPE
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.010   1st Qu.:0.000   1st Qu.:0.170   1st Qu.:0.000
 Median :1.000   Median :0.030   Median :1.000   Median :0.810   Median :0.750
 Mean   :0.539   Mean   :0.088   Mean   :0.585   Mean   :0.602   Mean   :0.516
 3rd Qu.:1.000   3rd Qu.:0.100   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000
 Max.   :1.000   Max.   :0.830   Max.   :1.000   Max.   :1.000   Max.   :1.000
```

Logit:

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

or probit:

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

```
> NAFTA.probit<-glm(Vote~PropHisp+Democrat+COPE+DemXCOPE,
                    NAFTA,family=binomial(link="probit"))
> summary(NAFTA.probit)

Call:
glm(formula = Vote ~ PropHisp + Democrat + COPE + DemXCOPE,
    family = binomial(link = "probit"), data = NAFTA)

Deviance Residuals:
   Min     1Q  Median     3Q    Max
-3.173 -0.677   0.362  0.764  1.817

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.078      0.153    7.03  2.1e-12 ***
PropHisp       1.279      0.467    2.74   0.0062 **
Democrat       3.034      0.739    4.11  4.0e-05 ***
COPE          -2.201      0.440   -5.00  5.8e-07 ***
DemXCOPE      -2.888      0.903   -3.20   0.0014 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

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

Number of Fisher Scoring iterations: 8
```

```
> NAFTA.fit<-glm(Vote~PropHisp+Democrat+COPE+DemXCOPE,
                 NAFTA,family=binomial)
> summary(NAFTA.fit)

Call:
glm(formula = Vote ~ PropHisp + Democrat + COPE + DemXCOPE,
    family = binomial, data = NAFTA)

Deviance Residuals:
   Min     1Q   Median      3Q     Max
-3.264  -0.650   0.310   0.728   1.818

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.792      0.275    6.50  7.8e-11 ***
PropHisp       2.091      0.794    2.63  0.00846 **
Democrat       6.866      1.547    4.44  9.1e-06 ***
COPE          -3.650      0.760   -4.80  1.6e-06 ***
DemXCOPE      -6.705      1.820   -3.68  0.00023 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 598.99  on 433  degrees of freedom
Residual deviance: 436.83  on 429  degrees of freedom
AIC: 446.8

Number of Fisher Scoring iterations: 5
```
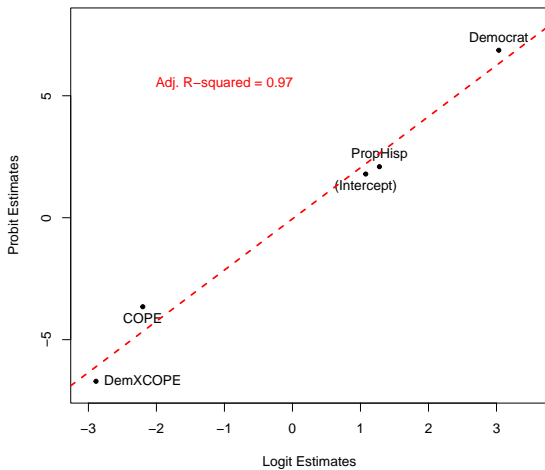
$\hat{\beta}_{\text{probit}}$ vs. $\hat{\beta}_{\text{logit}}$

Log-Likelihoods, "Deviance," etc.

- R / glm 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:

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

Example:

```
> LLR<-NAFTA.fit$null.deviance - NAFTA.fit$deviance

> LLR
[1] 162

> pchisq(LLR,4,lower.tail=FALSE)
[1] 5.04e-34
```

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:

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

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

```
> NAFTA.fit$coeff[4]+NAFTA.fit$coeff[5]

 COPE
-10.4

> # z-statistic:
>
> (NAFTA.fit$coeff[4]+NAFTA.fit$coeff[5]) /
+    (sqrt(vcov(NAFTA.fit)[4,4] +
+    (1)^2*vcov(NAFTA.fit)[5,5] +
+    2*1*vcov(NAFTA.fit)[4,5]))
 COPE
-6.25

> # Square that, and it's a chi-square statistic:
>
> ((NAFTA.fit$coeff[4]+NAFTA.fit$coeff[5]) /
+    (sqrt(vcov(NAFTA.fit)[4,4] +
+             (1)^2*vcov(NAFTA.fit)[5,5] +
+             2*1*vcov(NAFTA.fit)[4,5])))^2
COPE
  39
```

```
> library(car)
> linearHypothesis(NAFTA.fit,"COPE+DemXCOPE=0")
Linear hypothesis test

Hypothesis:
COPE  + DemXCOPE = 0

Model 1: restricted model
Model 2: Vote ~ Democrat + PropHisp + COPE + DemXCOPE

  Res.Df Df Chisq   Pr(>Chisq)
1    430
2    429  1    39 0.00000000042 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```
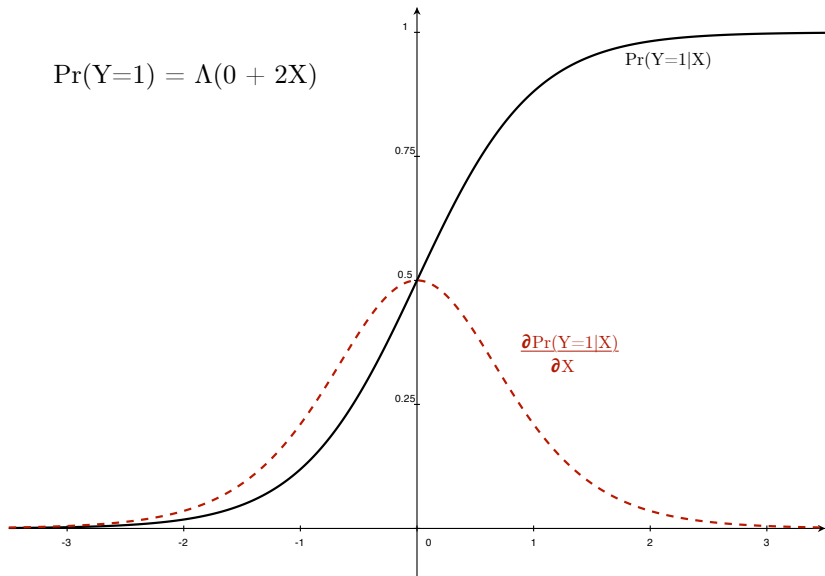
The *marginal effect* is:

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

Note that these depend on $\mathbf{X}\hat{\boldsymbol{\beta}}$, which means we either have to:

1. ...hold $\mathbf{X}\hat{\boldsymbol{\beta}}$ constant at some value(s), or
2. ...average over the actual values of $\mathbf{X}_i\hat{\boldsymbol{\beta}}$ observed in the data.

$$\Pr(Y=1) = \Lambda(0 + 2X)$$
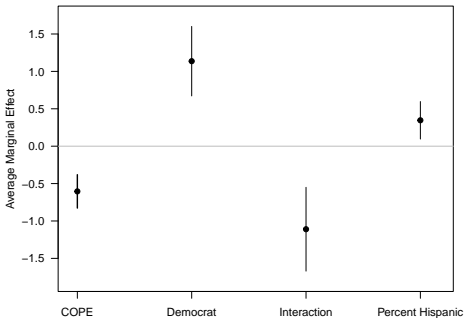
$\Pr(Y=1|X)$

$\dfrac{\partial \Pr(Y=1|X)}{\partial X}$

```
> summary(margins(NAFTA.fit))
  factor      AME     SE       z      p    lower   upper
    COPE -0.6043 0.1139 -5.3048 0.0000 -0.8275 -0.3810
 Democrat  1.1366 0.2370  4.7953 0.0000  0.6720  1.6011
 DemXCOPE -1.1101 0.2858 -3.8836 0.0001 -1.6703 -0.5498
 PropHisp  0.3462 0.1280  2.7054 0.0068  0.0954  0.5970
```

Plotted:

Log-Odds of $Y = 1$ are linear in $\mathbf{X}$:

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

That implies that:

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

OR for a one-unit change in X_k:

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

OR for a δ-unit change in X_k:

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

Also:

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

```
> P<-qnorm(0.975)

> lreg.or <- function(model)
+ {
+   coeffs <- coef(summary(model))
+   lowerCI <- exp(coeffs[ ,1] - P * coeffs[ ,2])
+   OR <- exp(coeffs[ ,1])
+   upperCI <- exp(coeffs[ ,1] + P * coeffs[ ,2])
+   lreg.or <- cbind(OR,lowerCI,upperCI)
+   lreg.or
+ }

> lreg.or(NAFTA.fit)
                  OR    lowerCI    upperCI
(Intercept)   5.99928   3.4965990   10.2933
PropHisp      8.09352   1.7068838   38.3770
Democrat    958.67832  46.1969511 19894.4757
COPE          0.02599   0.0058625    0.1152
DemXCOPE      0.00122   0.0000345    0.0434

> Or via -confint-...

> exp(cbind(OR=coef(NAFTA.fit),confint.default(NAFTA.fit)))
                  OR       2.5 %      97.5 %
(Intercept)   5.99928   3.4965990   10.2933
PropHisp      8.09352   1.7068838   38.3770
Democrat    958.67832  46.1969511 19894.4757
COPE          0.02599   0.0058625    0.1152
DemXCOPE      0.00122   0.0000345    0.0434
```

```
> NAFTA.fit

Coefficients:
(Intercept)      PropHisp      Democrat          COPE      DemXCOPE
       1.79          2.09          6.87         -3.65         -6.71

> exp(cbind(OR=coef(NAFTA.fit),confint.default(NAFTA.fit)))
                   OR        2.5 %       97.5 %
(Intercept)   5.99928    3.4965990      10.2933
PropHisp      8.09352    1.7068838      38.3770
Democrat    958.67832   46.1969511   19894.4757
COPE          0.02599    0.0058625       0.1152
DemXCOPE      0.00122    0.0000345       0.0434
```

Consider PropHispc:

- A one-unit change (from 0 percent to 100 percent) in PropHisp corresponds to a $[(8.094 - 1) \times 100] =$ **709 percent expected increase in the odds** of a member of Congress voting in favor of NAFTA.

- A change of 0.10 (that is, a ten percentage-point increase) in the proportion of a member's district who is Hispanic corresponds to an **odds ratio** of:

$$\exp(2.09 \times 0.10) = \exp(0.209)$$
$$= 1.232$$

- This means that an increase of 0.10 in PropHisp corresponds to a $[(1.232 - 1) \times 100] =$ **23.2 percent expected increase in the odds** that a member of Congress would have voted in favor of NAFTA.

- For an increase of 0.20 (that is, 20 percentage points), the corresponding odds ratio and percent increase are 1.519 and 51.9 percent, respectively.
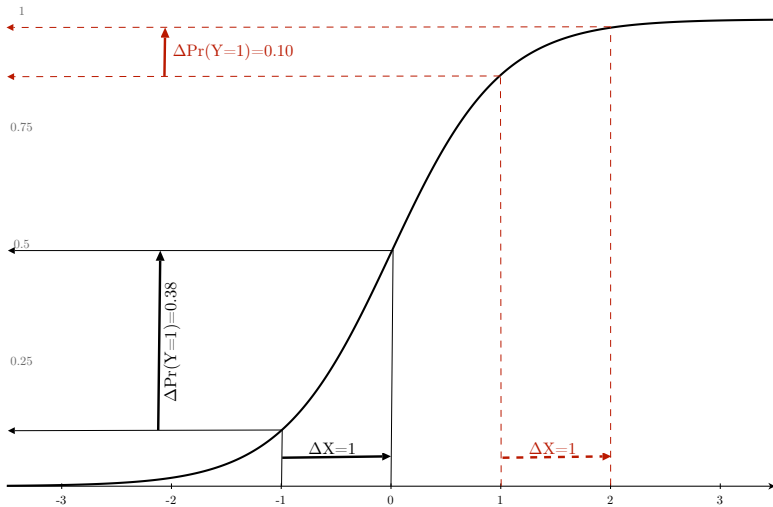
Percentage Change in Odds, by ΔPropHisp
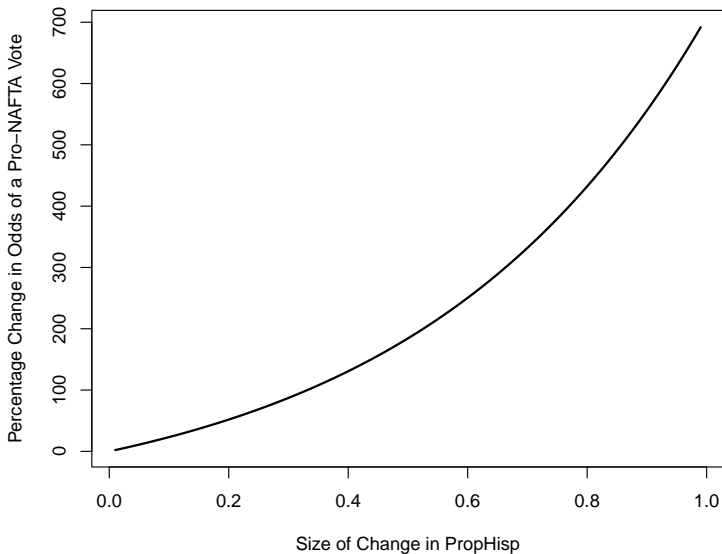
Predicted probabilities:

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

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

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

Predicted Probabilities Illustrated

Predicted Probabilities: Standard Errors

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

$$= [f(\mathbf{X}_i\hat{\boldsymbol{\beta}})]^2 \mathbf{X}_i' \hat{\mathbf{V}} \mathbf{X}_i$$

So,

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

Changes in $\widehat{\Pr(Y = 1)}$:

$$
\begin{aligned}
\Delta\widehat{\Pr(Y = 1)}_{\mathbf{x}_A \to \mathbf{x}_B} &= \frac{\exp(\mathbf{X}_B\hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{X}_B\hat{\boldsymbol{\beta}})} - \frac{\exp(\mathbf{X}_A\hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{X}_A\hat{\boldsymbol{\beta}})} \\
&\qquad \text{or} \\
&= \Phi(\mathbf{X}_B\hat{\boldsymbol{\beta}}) - \Phi(\mathbf{X}_A\hat{\boldsymbol{\beta}})
\end{aligned}
$$

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

```
> preds<-NAFTA.fit$fitted.values

> hats<-predict(NAFTA.fit,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))
```
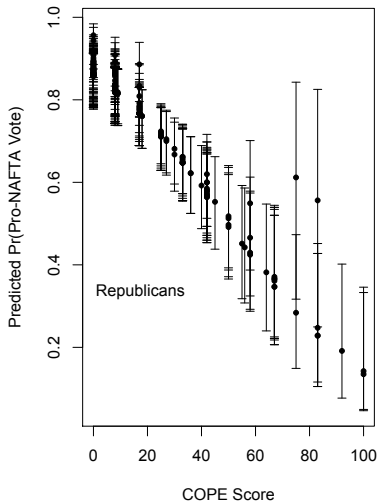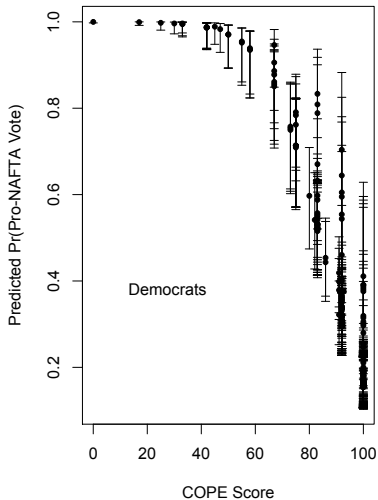
```
...
> par(mfrow=c(1,2))
> library(plotrix)

> with(NAFTA,
+ plotCI(COPE[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)"))

> with(NAFTA,
+ plotCI(COPE[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)"))
```

Out-of-Sample Predictions

"Fake" data:

```
> sim.data<-data.frame(PropHisp=mean(NAFTA$PropHisp),Democrat=rep(0:1,101),
                       COPE=seq(from=0,to=1,length.out=101))
> sim.data$DemXCOPE<-sim.data$Democrat*sim.data$COPE
```

Generate predictions:

```
> OutHats<-predict(NAFTA.fit,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))
```

Plot:
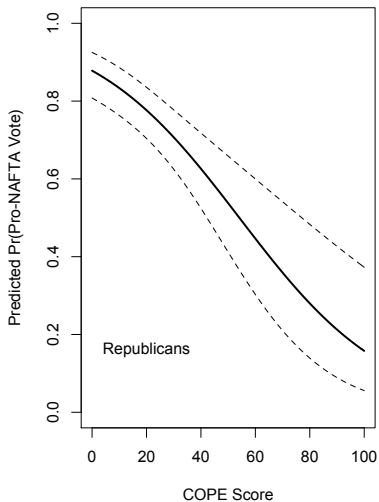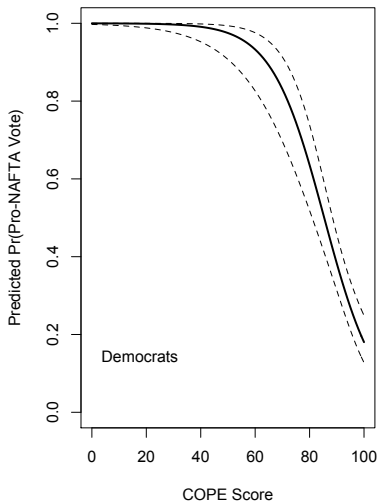
```
> both<-cbind(sim.data,OutHats)
> both<-both[order(both$COPE,both$Democrat),]
> bothD<-both[both$Democrat==1,]
> bothR<-both[both$Democrat==0,]

> par(mfrow=c(1,2))

> plot(bothD$COPE,bothD$fit,t="l",lwd=2,ylim=c(0,1),
        xlab="COPE Score",ylab="Predicted Pr(Pro-NAFTA Vote)")
> lines(bothD$COPE,bothD$OutHatsUB,lty=2)
> lines(bothD$COPE,bothD$OutHatsLB,lty=2)
> text(0.3,0.2,label="Democrats")

> plot(bothR$COPE,bothR$fit,t="l",lwd=2,ylim=c(0,1),
        xlab="COPE Score",ylab="Predicted Pr(Pro-NAFTA Vote)")
> lines(bothR$COPE,bothR$OutHatsUB,lty=2)
> lines(bothR$COPE,bothR$OutHatsLB,lty=2)
> text(0.7,0.9,label="Republicans")
```

- Pseudo-R^2 (no!)

- Proportional reduction in error (PRE) – a/k/a "accuracy"

- ROC curves.

Suppose we assign:

$$\hat{Y}_i = 0 \quad \text{if} \quad \widehat{\Pr(Y_i = 1)} \leq \tau$$
$$\hat{Y}_i = 1 \quad \text{if} \quad \widehat{\Pr(Y_i = 1)} > \tau$$

This would then give us a "confusion matrix":

	Predicted $\hat{Y}_i$	
Actual Y_i	$\hat{Y}_i = 0$	$\hat{Y}_i = 1$
$Y_i = 0$	True Negative ("TN")	False Positive ("FP")
$Y_i = 1$	False Negative ("FN")	True Positive ("TP")

This means we have:

- Total actual negatives = TN + FP
- Total actual positives = TP + FN
- Number correctly predicted = TP + TN

Proportional Reduction in Error (PRE):

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

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

PRE tells us how much (proportionally) better our model does at predicting Y in-sample than would a model that only contained an intercept.

```
> Assume tau = 0.5...
>
> table(NAFTA.fit$fitted.values>0.5,nafta$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.fit$fitted.values>0.5,NAFTA$Vote==1)

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

data:  NAFTA.fit$fitted.values > 0.5 and NAFTA$Vote == 1
X-squared = 120, df = 1, p-value <2e-16
```

Concepts:

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

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

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

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.

```
> # 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*)

- 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
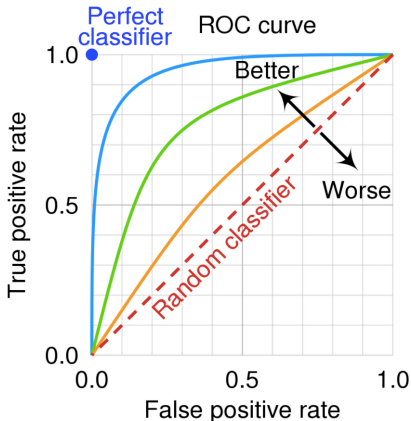
(Source)

"AUROC": Area under the ROC curve
→ assessment of model fit

```
> 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)
```
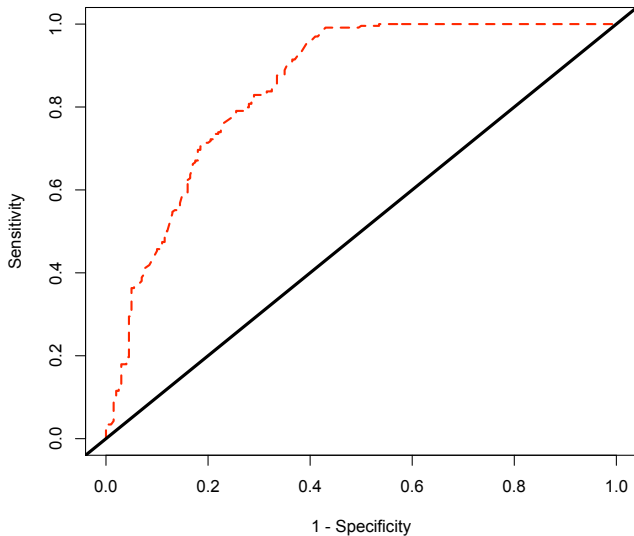
Interpreting AUROC Curves

- Area under ROC = 0.90-1.00 → Excellent (A)

- Area under ROC = 0.80-0.90 → Good (B)

- Area under ROC = 0.70-0.80 → Fair (C)

- Area under ROC = 0.60-0.70 → Poor (D)

- Area under ROC = 0.50-0.60 → Total Failure (F)

ROC Curve: A Poorly-Fitting Model

Model is:

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

```
> 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!
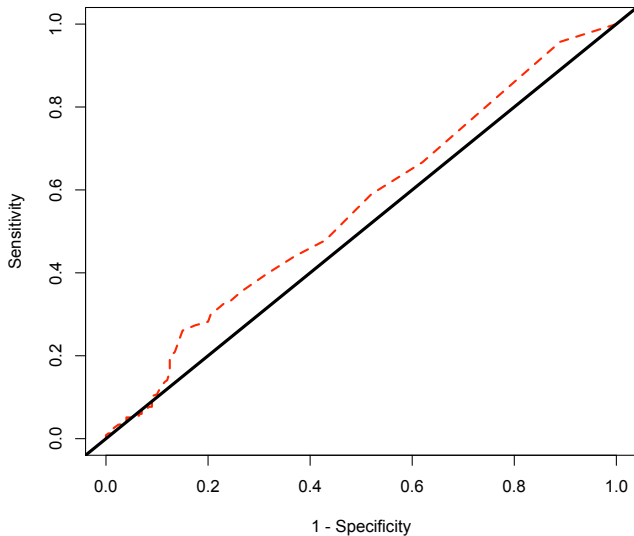
Comparing ROCs

```
> 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
```
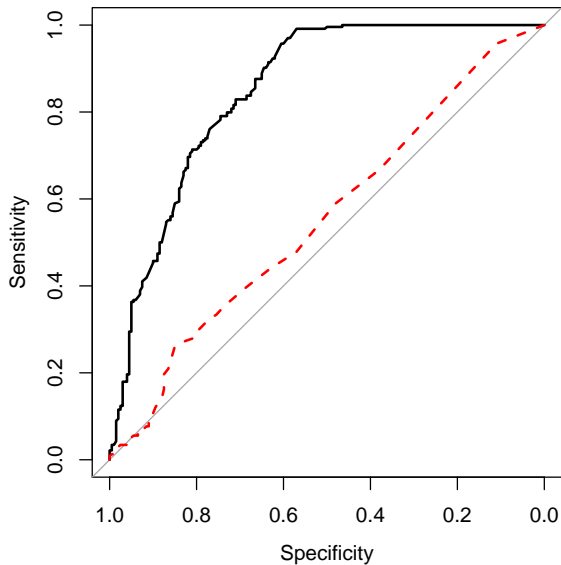
Things we'll probably talk about later:

- Rare events and "separation"
- Binary responses in panel / longitudinal data
- Multilevel / hierarchical models for binary responses
- Models with (binary) sample selection
- Measurement models for binary outcomes (e.g., item response models)
- Semi- and non-parametric models (see, e.g., Horowitz and Savin 2001)
- "Heteroscedastic" models (where $\sigma_i^2 \neq \sigma^2 \, \forall \, i$) (see, e.g., Alvarez and Brehm 1995, 1997; Tutz 2018)
- "Bivariate" probit models, where

$$\{Y_{1i}, Y_{2i}\} \sim BVN(0, 0, 1, 1, \rho)$$