# Gov 2001 Section 6: Dichotomous Outcomes

Konstantin Kashin[1]

March 6, 2013

[1] Thanks to Jen Pan, Brandon Stewart, Iain Osgood, and Patrick Lam for contributing to this material.

# Outline

# Replication Paper

- You should be in the process of replicating one (or more) articles!
- Replication and preliminary write-up due Wednesday, March 27 at 7pm
- Submit using a link (to a Dropbox or other means of sharing)

# Problem Set

- Problem set 5 released today
- Due next Wednesday, March 13 at 7pm
- One assessment problem for which you will not be allowed to collaborate / talk to other people

# Section Next Week

Section next week will be on Tuesday, March 12 from 7-8pm and 8-9pm in CGIS K354.

# Outline

# Generalized Linear Models

All of the models we've talked about so far (and for most of the rest of the class) belong to a class called **generalized linear models (GLM)**.

Three elements of a GLM (equivalent to Gary's stochastic and systematic component):

- A distribution for $Y$ (stochastic component)
- A linear predictor $X\beta$ (systematic component)
- A link function that relates the linear predictor to the mean of the distribution. (systematic component)

# 1. Specify a distribution for $Y$

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson
- Duration: Exponential
- Ordered Categories: Normal with observation mechanism
- Unordered Categories: Multinomial

These distributions are all members of the exponential family.

# 2. Specify a linear predictor

We are interested in allowing some parameter of the distribution $\theta$ to vary as a (linear) function of covariates. So we specify a linear predictor.

$$X\beta = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k$$

# 3. Specify a link function

The link function is a one-to-one continuous differentiable function that relates the linear predictor to some parameter $\theta$ of the distribution for $Y$ (almost always the mean).

Let $g(\cdot)$ be the link function and let $E(Y) = \theta$ be the mean of distribution for $Y$.

$$\begin{aligned} g(\theta) &= X\beta \\ \theta &= g^{-1}(X\beta) \end{aligned}$$

Note that we usually use the **inverse link function** $g^{-1}(X\beta)$ rather than the link function.[2]

This is the systematic component that we've been talking about all along.

[2] In some texts, $g^{-1}(\cdot)$ is referred to as the link function

# Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$
- Inverse Link: $\pi = \frac{1}{1+e^{-X\beta}}$

Probit:

- Link: $\Phi^{-1}(\pi) = X\beta$
- Inverse Link: $\pi = \Phi(X\beta)$

Log:

- Link: $\ln(\lambda) = X\beta$
- Inverse Link: $\lambda = \exp(X\beta)$

# 4. Estimate Parameters via ML

Use `optim` to estimate the parameters just like we have all along.

# 5. Quantities of Interest

1. Simulate parameters from multivariate normal.
2. Run $X\beta$ through inverse link function to get expected values.
3. Draw from distribution of $Y$ for predicted values.

Note: this can also be calculated analytically via the Delta Method. We aren't going to cover this in section though...

# Outline

# Binary Dependent Variable

Let's do the Probit as an example.

Let our dependent variable be a binary random variable that can take on values of either 0 or 1.

1. Specify a distribution for $Y$

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ p(\mathrm{y}|\boldsymbol{\pi}) &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \end{aligned}$$

2. Specify a linear predictor: $X\beta$

3. Specify a link (or inverse link) function.
   - Logit: $\pi_i = \frac{1}{1+e^{-x_i\beta}}$
   - Probit: $\pi_i = \Phi(x_i\beta)$
   - Complementary Log-log (cloglog): $\pi_i = 1 - \exp(-\exp(X\beta))$
   - Scobit: $\pi_i = (1 + e^{-x_i\beta})^{-\alpha}$

4. Estimate parameters via ML.

$$\begin{aligned} \ln L(\beta|y) &\propto \sum_{i=1}^{n} y_i \ln(\pi_i) + (1 - y_i)\ln(1 - \pi_i) \\ &= \sum_{i=1}^{n} y_i \ln(\Phi(X_i\beta)) + (1 - y_i)\ln(1 - \Phi(X_i\beta)) \end{aligned}$$

5. Simulate Quantities of Interest

# Outline

# The Data: Political Assassinations

Taken from Olken and Jones (2009), "Hit or Miss? The Effect of Assassinations on Institutions and War", *American Economic Journal: Macroeconomics*.

Dataframe is called `as` and contains information on assassination attempts, success or failure, and various covariates.

```
> as[as$country == "United States" & as$year == "1975",]
      country year leadername age tenure attempt
United States 1975       Ford  62    510    TRUE
      survived result dem_score civil_war war
             1     24        10         0   0
           pop  energy solo weapon
        215973 2208506    1    gun
```

Observations are country-year-leaders, so some country-years have multiple observations.

# The Probit Model

Let's try to predict assassination attempts with some of our covariates.

"The question of which distribution to use is a natural one... There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds ...[A]s a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference."
-*Econometric Analysis*, Greene. pg. 774.

# The Probit Model

Let's try to predict assassination attempts with some of our covariates.

The model:

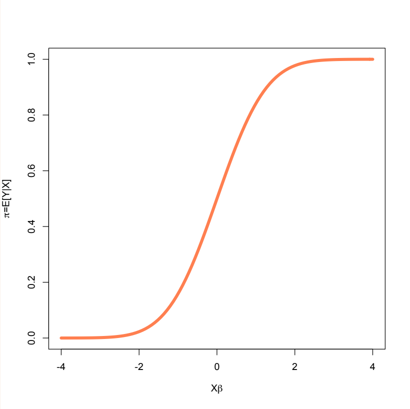
1. $Y_i \sim f_{\text{bern}}(y_i|\pi_i)$.
2. $\pi_i = \Phi(X_i\beta)$ where $\Phi$ is the CDF of the standard normal distribution.
3. $Y_i$ and $Y_j$ are independent for all $i \neq j$.

Like all CDF's, $\Phi$ has range 0 to 1, so it puts our $\pi_i$ on the right scale.

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(\frac{z^2}{2}) dz.$$

# Visualizing the Probit Link

$X_i\beta \longrightarrow$




$\longrightarrow \pi_i = \mathrm{E}[Y_i]$

# The Probit Model

We can then derive the log-likelihood for $\beta$:

$$\begin{aligned} L(\beta|y) &\propto \prod_{i=1}^{n} f_{\text{bern}}(y_i|\pi_i) \\ &= \prod_{i=1}^{n} (\pi_i)^{y_i}(1-\pi_i)^{(1-y_i)} \end{aligned}$$

Therefore:

$$\begin{aligned} \ln L(\beta|y) &\propto \sum_{i=1}^{n} y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i) \\ &= \sum_{i=1}^{n} y_i \ln(\Phi(X_i\beta)) + (1-y_i)\ln(1-\Phi(X_i\beta)) \end{aligned}$$

# Implementing the Likelihood

Here's one approach:

```
ll.probit <- function(beta, y=y, X=X){
  if(sum(X[,1]) != nrow(X)) X <- cbind(1,X)
  phi <- pnorm(X%*%beta, log = TRUE)
  opp.phi <- pnorm(X%*%beta, log = TRUE, lower.tail = FALSE)
  logl <- sum(y*phi + (1-y)*opp.phi)
  return(logl)
}
```

Notes:

1. The STN CDF is evaluated with pnorm. R's pre-programmed log of the CDF has greater range than `log(pnorm())` (try Z=-50).
2. If `lower.tail = FALSE` gives $Pr(Z \geq z)$.
3. Uses a logical test to check that an intercept column has been added

# What's in our model?

We wish to predict assassination attempts for country-year-leaders.

- `tenure`: number of days in office
- `age`: age of leader, in years
- `dem_score`: polity score, -10 to 10
- `civil_war`: is there currently a civil war?
- `war`: is country in an international conflict?
- `pop`: the country's population, in thousands
- `energy`: energy usage

# Maximization

```
y <- as$attempt
X <- as[,c("tenure","age","dem_score","civil_war",
           "war","pop","energy")]

yX <- na.omit(cbind(y,X))
y <- yX[,1]; X <- yX[,-1]

X$tenure <- (X$tenure-mean(X$tenure))/sd(X$tenure)
X$pop <- (X$pop-mean(X$pop))/sd(X$pop)
X$energy <- (X$energy-mean(X$energy))/sd(X$energy)

opt <- optim(par = rep(0,8), fn = ll.probit, y=y, X=X,
       method = "BFGS", control = list(fnscale = -1,
       maxit = 1000), hessian = TRUE)
```

# The Fit Model

| | Coef | SE | Z-stat |
|---|---|---|---|
| tenure | 0.0291 | 0.0294 | 0.9922 |
| age | -0.0058 | 0.0025 | -2.2960 |
| dem score | -0.0124 | 0.0045 | -2.7921 |
| civil war | 0.0663 | 0.0890 | 0.7445 |
| war | 0.3175 | 0.1014 | 3.1298 |
| pop | 0.0409 | 0.0237 | 1.7264 |
| energy | 0.0269 | 0.0243 | 1.1034 |
| intercept | -1.8362 | 0.1399 | -13.1262 |

# Outline

# Quantities of Interest

General considerations:

1. Incorporating estimation uncertainty.
2. Incorporating fundamental uncertainty when making predictions.
3. Establishing appropriate baseline values for QOI, and considering plausible changes in those values.

# Expected Values

- Expected values: for this model we will be interested in estimating the predicted probability of an assassination attempt at some level for the covariate values. In general, $E[y|X]$.
- Let's identify the covariates under which an assassination attempt occured and call it `highrisk` ($X_{HR}$)

```
attempt.vec <- y == 1   # y is attempt without NAs
highrisk <- apply(X = X[attempt.vec,],
                  MARGIN = 2, FUN = median)
```

| **Predictor**: | tenure | age | dem score | civil war | war | pop. | energy |
|---|---|---|---|---|---|---|---|
| **Value**: | 1392 | 54 | -3 | 0 | 0 | 12980000 | 3828 |

Table: Median values for year-states with an assassination attempt.

# Expected Values

What's the estimated probability of an assassination at $X_{HR}$?

```
library(MASS)
beta.draws <- mvrnorm(10000, mu = opt$par, Sigma
              = solve(-opt$hessian))

p.ests <- c()
for(i in 1:10000){
  p.ass.att <- pnorm(highrisk%*%beta.draws[i,])
  outcomes <- rbinom(10000, 1, p.ass.att)
  p.ests[i] <- mean(outcomes)
}

> mean(p.ests)
[1] 0.0166266
> quantile(p.ests, .025); quantile(p.ests, .975)
 2.5%    97.5%
0.0134  0.0201
```

# Expected Values

What are the steps that I just took?

1. simulate from the estimated sampling distribution of $\hat{\beta}$ to incorporate estimation uncertainty.
2. start our for-loop which will do steps 3-5 each time.
3. combine one $\tilde{\beta}$ draw with $X_{HR}$ as $X_{HR}\tilde{\beta}$, then plug into $\Phi()$ to get probability of attempt for that $\tilde{\beta}$ draw.
4. draw a bunch of outcomes from the $Bernoulli(\Phi(X_{HR}\tilde{\beta}))$.
5. average over those draws to get one simulated $E[y|X_{HR}]$.
6. return to step 3.