POLI 8501
Binary Response Models, III

Heterscedasticity and Binary Response Models

Consider a basic latent-variable probit model with N observations and k independent variables $\mathbf{X_i}$:

$$Y_i^* = \mathbf{X_i}\boldsymbol{\beta} + u_i \tag{1}$$

with

$$u_i \sim N(0, \sigma^2)$$

and

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

We know that

$$\begin{aligned} \Pr(Y_i = 1) =& \Pr(Y_i^* > 0) \\ =& \Pr(u_i > -\mathbf{X_i}\boldsymbol{\beta}) \\ =& \Phi(\mathbf{X_i}\boldsymbol{\beta}) \end{aligned} \tag{2}$$

Now, notice that we can divide both sides of the inequality in (2) by the same number without changing anything, like:

$$\Pr(Y_i = 1) = \Pr\left(\frac{u_i}{\sigma} > -\mathbf{X_i}\frac{\boldsymbol{\beta}}{\sigma}\right)$$

where σ is the standard deviation of the error distribution (that is, $\sqrt{Var(u_i)}$ in (1)). This then gives:

$$\Pr(Y_i = 1) = \Phi\left(-\mathbf{X_i}\frac{\boldsymbol{\beta}}{\sigma}\right) \tag{3}$$

Now, when we estimate the probit model, call what we actually estimate $\hat{\boldsymbol{\beta}}'$, where $\hat{\boldsymbol{\beta}}' = \frac{\hat{\boldsymbol{\beta}}}{\sigma}$. Again, we "get rid off" the σs by assuming that they are equal to 1 (since we assume, generally, that $\sigma^2 = 1$ in the distribution of the u_is.

In a nutshell, this is just illustrating that, because the underlying variable Y^* is latent, we can't estimate its "scale" on the basis of the data we have on Y.

Now suppose you have two groups in your data. Specifically, imagine that you're modeling some choice by individuals, and you have two groups: professors and graduate students. Consider the model:

Professors (p):

$$\begin{aligned} Y_{ip}^* &= \mathbf{X}_{ip}\boldsymbol{\beta} + u_{ip} \\ Y_{ip} &= 1 \text{ if } Y_{ip}^* > 0 \\ Y_{ip} &= 0 \text{ otherwise.} \end{aligned}$$

Graduate students (g):

$$\begin{aligned} Y_{ig}^* &= \mathbf{X}_{ig}\boldsymbol{\beta} + u_{ig} \\ Y_{ig} &= 1 \text{ if } Y_{ig}^* > 0 \\ Y_{ig} &= 0 \text{ otherwise.} \end{aligned}$$

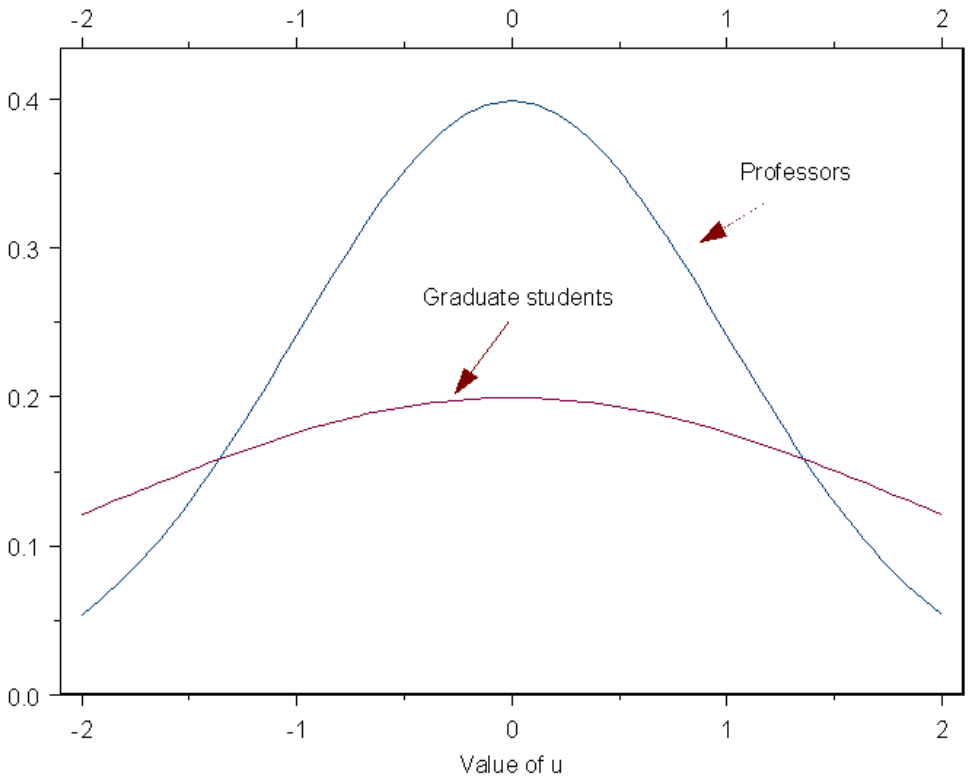
- Both groups give the same "weight" to the explanatory factors in making their choices; i.e., both models have the same coefficients ($\boldsymbol{\beta}$s), BUT
- Professors, being smart, always know exactly what they want. They make rational calculations, on the basis of known variables. As a result, the variance in their error terms is small.
- In contrast, graduate students can't find their ass with both hands. They are unsure, and rarely make decisions in a systematic way. There is a large-variance stochastic component to their decision-making.

Formally:

$$\begin{aligned} u_{ip} &\sim N(0, \sigma_p^2) \\ u_{ig} &\sim N(0, \sigma_g^2) \\ \sigma_g^2 &> \sigma_p^2 \end{aligned}$$

That is, for these two groups, the distributions of errors u look like:

Figure 1: Latent error distributions for professors and graduate students.



Now, what happens if we "pool" the data, and run a single model on both grad students and professors?

You might think that, because the coefficients are the same, we'd get "good" estimates of them, right? WRONG...

Instead, we'd estimate:

$$\begin{aligned} \hat{\boldsymbol{\beta}}'_p &= \frac{\boldsymbol{\beta}_p}{\sigma_p} \\ \hat{\boldsymbol{\beta}}'_g &= \frac{\boldsymbol{\beta}_g}{\sigma_g} \end{aligned} \tag{4}$$

That is, the coefficients are different for each group! Moreover, they're related:

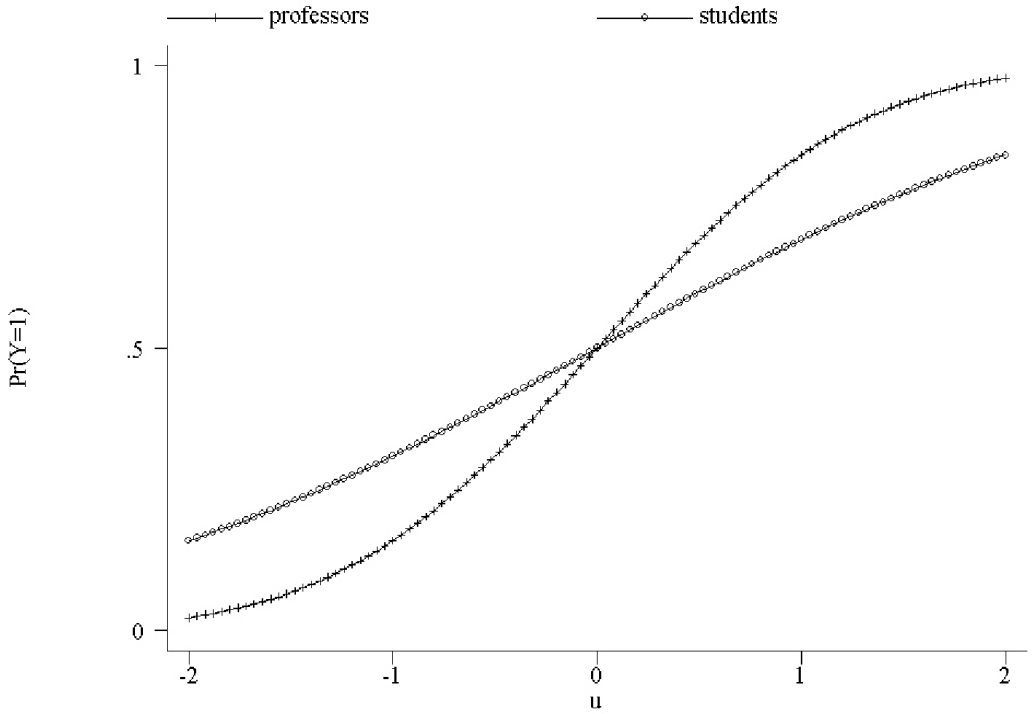
$$\begin{aligned}\frac{\hat{\boldsymbol{\beta}}'_p}{\hat{\boldsymbol{\beta}}'_g} &= \frac{\frac{\boldsymbol{\beta}_p}{\sigma_p}}{\frac{\boldsymbol{\beta}_g}{\sigma_g}} \\ &= \frac{\sigma_g}{\sigma_p}\end{aligned}$$

- Where we'd have expected the ratio of the two coefficients to be 1.0 (since they are, by construction, the same), instead we find that they aren't...

  BUT: What happens if $\sigma_g = \sigma_p$? (A: We're back to the standard probit model).

We can see this if we consider the cumulative densities (that is, $\Pr(Y^*_{ip,g} > u)$, or $\int_{-\infty}^{u} \phi_{p,g}(u) \equiv \Phi_{p,g}(u)$) for the two groups:

Figure 2: CDFs for professors and graduate students.



Think a little about this result.

- In the probit model, we observe only $Y_i^* > 0$ or $Y_i^* \leq 0$.
- If the variance in the error term increases, then so does the variance of the probability of interest $\Pr(u_i > -\mathbf{X}_i\boldsymbol{\beta})$.
- As a result, our predictions of $\Pr(Y_i = 1)$ must move towards 0.5, which in turn means that...
- The coefficients estimated for $\hat{\boldsymbol{\beta}}$ must move towards zero. Or, put differently,
- As long as all of the "underlying" Y^*s have the same "scale", we can assume whatever we want for σ^2 without any problems.

BUT...

- If some subset of observations has a greater variance in their error term, the coefficients for that group will be different from the rest...
- They'll have to be, since if the underlying scale is, say, larger, the coefficient will have to be larger to get the same marginal impact on the (observed) probability of $Y_i = 1$...

Now, suppose a graduate student ignores the differences in the scales between professors and graduate students and estimates $\hat{\boldsymbol{\beta}}'$ on the "pooled" data together...

We'd get an estimates of:

$$\hat{\boldsymbol{\beta}}' = \text{some sort of average of } \hat{\boldsymbol{\beta}}'_p \text{ and } \hat{\boldsymbol{\beta}}'_g.$$

The more grad students there are in the mix, the closer the estimates of $\hat{\boldsymbol{\beta}}'$ will be to $\hat{\boldsymbol{\beta}}'_g$ (and vice-versa).

The problem remains:

- The parameter coefficients are both *biased* and *inconsistent*.
- That is, the expected value of the "pooled" estimates will not be equal to the "true" values,
- Nor will increasing the number of observations remedy this problem.
- Plus, the standard error estimates will also be all wrong...

SO...

Heterodscedasticity in probits is a bad thing...

How Do We Deal With It?

- In the above example, we'd like to allow the variance for the two groups to be different.
- If we did this, we could estimate the parameters consistently, since we'd then be "dividing" (or "rescaling") the coefficients differently, depending on whether we were talking about professors or grad students.

Generally, this is the approach we take.

- We allow the variance of the unobserved variable to vary according to some function of one or more dependent variables...
- Because variance is always positive, we want the function to always be positive as well...

So we might write:

$$\text{Var}(u_i) \equiv \sigma_i^2 = \exp(\mathbf{Z}_i\gamma)^2 \tag{5}$$

where $\mathbf{Z}_i$ is a vector of covariates that define groups with different error variances in the underlying (latent) variable. This means that:

$$\text{s.d.}(u_i) \equiv \sigma_i = \exp(\mathbf{Z}_i\gamma) \tag{6}$$

In this model we assume that the latent errors are distributed $N(0, [\exp(\mathbf{Z}_i\gamma)]^2)$. The probability function for a particular observation then equals:

$$\Pr(Y_i = 1) = \Phi\left[\frac{\mathbf{X}_i\boldsymbol{\beta}}{\exp(\mathbf{Z}_i\gamma)}\right] \tag{7}$$

This compares to that for the standard probit model, where we assume that the variance is the same (and equal to one) for all observations:

$$\Pr(Y_i = 1) = \Phi(\mathbf{X}_i\boldsymbol{\beta}) \tag{8}$$

This alternative, *heteroscedastic probit* model gives us a pretty easy log-likelihood:

$$\ln L(\boldsymbol{\beta}, \gamma|\mathbf{X}_i, \mathbf{Z}_i) = \sum_{i=1}^{N}\left\{Y_i \ln\Phi\left[\frac{\mathbf{X}_i\boldsymbol{\beta}}{\exp(\mathbf{Z}_i\gamma)}\right] + (1 - Y_i)\ln\left[1 - \Phi\left(\frac{\mathbf{X}_i\boldsymbol{\beta}}{\exp(\mathbf{Z}_i\gamma)}\right)\right]\right\} \tag{9}$$

- Maximizing this with respect to $\boldsymbol{\beta}$ and γ gives us our parameter estimates.
- Taking the negative of the inverse of the information matrix yields an estimate of the variance-covariance matrix, from which we get standard errors.

Before you freak out looking at (9), stop and think for a minute...

- What happens if all the estimated parameters in $\hat{\gamma}$ equal zero? (A: The variance equals $\exp(0) = 1$, and we get the standard probit model).
- What happens if a variable in $\mathbf{Z}$ has a positive coefficient?
    - As that variable increases, the variance of Y^* also increases, and the "impact" of the independent variables on $\Pr(Y_i = 1)$ decrease, all else equal.
    - This ought to make some sense: As the variability of the underlying variable gets greater, it takes a bigger change in the independent variable to push the binary variable across the "threshold."

This can be illustrated by looking at the marginal effects (i.e., the partial derivatives). For the standard probit model, we have:

$$\frac{\partial \Pr(Y_i = 1)}{\partial X_k} = \phi(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})\hat{\beta}_k$$

where the $\tilde{\mathbf{X}}$ are the independent variables selected at some predetermined values. For the heteroscedastic probit model, though, we have to consider the variance:

$$\frac{\partial \Pr(Y_i = 1)}{\partial X_k} = \phi\left(\frac{\tilde{\mathbf{X}}\hat{\beta}}{\exp(\tilde{\mathbf{Z}}\hat{\gamma})}\right)\left(\frac{\hat{\beta}_k}{\exp(\tilde{\mathbf{Z}}\hat{\gamma})}\right)$$

where, again, we have to select values for the $\mathbf{Z}$s as well as the $\mathbf{X}$s.

Interpretation

The first thing to remember is the formula for the probability that a particular observation equals one, from above:

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

With this in mind, you can calculate predicted probabilities for different values of $\mathbf{X}$ and $\mathbf{Z}$.

- This is really the best way to interpret these models
- Normal t-tests, etc. can be misleading...

A Familiar Example

Consider the NAFTA vote data again...

```
. probit vote democrat pcthispc cope93

Probit estimates                          Number of obs   =        434
                                          LR chi2(3)      =     144.48
                                          Prob > chi2     =     0.0000
Log likelihood = -227.25328               Pseudo R2       =     0.2412
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    democrat |    1.094652   .2712302     4.04   0.000     .5630503    1.626253
    pcthispc |    .0115188   .0047351     2.43   0.015     .0022381    .0207995
       cope93 |   -.0318482   .0036632    -8.69   0.000    -.0390279   -.0246686
        _cons |    1.330634   .1456266     9.14   0.000     1.045211    1.616057
------------------------------------------------------------------------------
```

Now, run a heteroscedastic probit...

```
. hetprob vote democrat pcthispc cope93, het(democrat pcthispc cope93)

Heteroskedastic probit model              Number of obs     =        434
                                          Zero outcomes     =        200
                                          Nonzero outcomes  =        234
                                          Wald chi2(3)      =      16.89
Log likelihood = -198.0597                Prob > chi2       =     0.0007
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
vote         |
    democrat |    .1530367    .052778     2.90   0.004     .0495937    .2564798
    pcthispc |    .0008621   .0003657     2.36   0.018     .0001454    .0015788
       cope93 |    -.007149   .0018226    -3.92   0.000    -.0107212   -.0035768
        _cons |    .4842633   .1160275     4.17   0.000     .2568536     .711673
-------------+----------------------------------------------------------------
lnsigma2     |
    democrat |   -.2452263   .2145403    -1.14   0.253    -.6657176    .1752649
    pcthispc |   -.0234672   .0066703    -3.52   0.000    -.0365407   -.0103937
       cope93 |   -.0264798    .004055    -6.53   0.000    -.0344275    -.018532
------------------------------------------------------------------------------
Likelihood ratio test of lnsigma2=0: chi2(3)=58.39 Prob > chi2 = 0.000
```

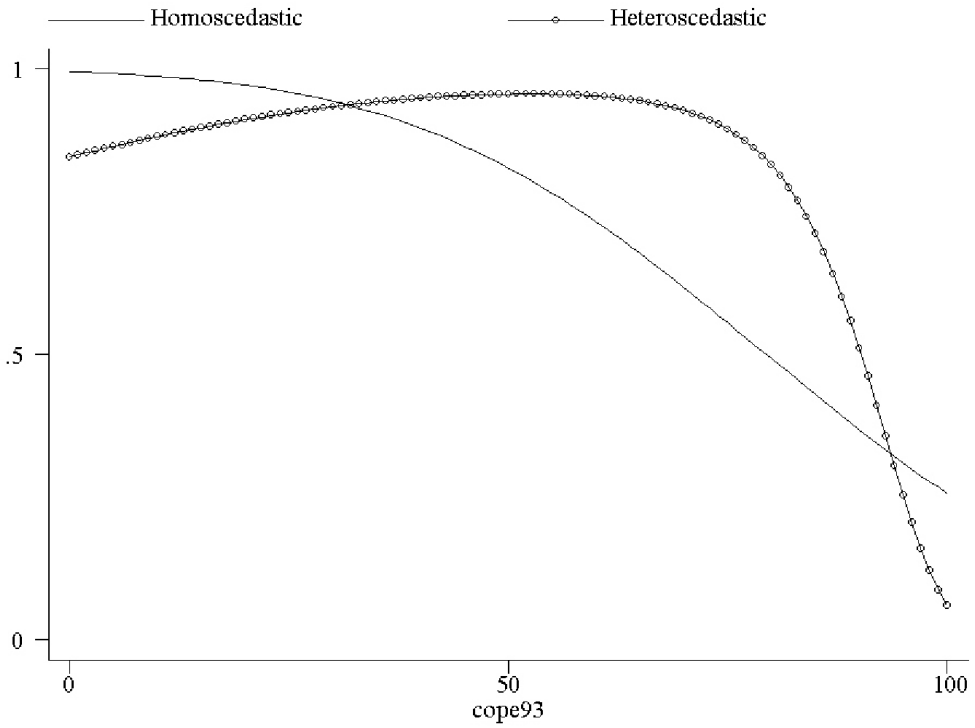
Note the differences – the variables still have the expected effects on the mean, but

- As the percentage Hispanic in a member's district increases, their variability decreases.

- The same is true for more pro-union members (suggesting that union membership / support was a clear, separating signal on the NAFTA vote).
- In contrast, Democrats were neither more nor less "cross-pressured" than were Republicans.

We can generate predicted probabilities by (e.g.) `cope93` score, holding other variables constant at their means/medians, using the `-predict-` command (just as we did for a standard probit/logit model...):

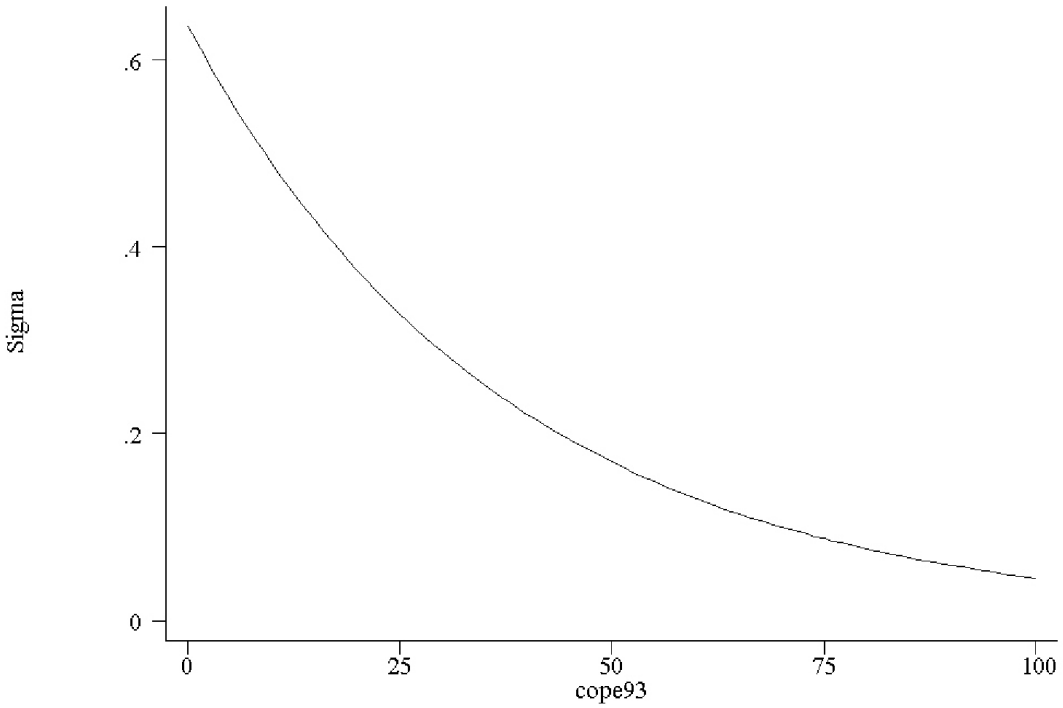
Figure 3: Predicted Probabilities by `cope93`, Homoscedastic and Heteroscedastic Probit Models (with other variables held constant at means/medians)



Note that the predictions for the heteroscedastic model are non-monotonic – this is because the `cope93` variable is in both the numerator and the denominator (that is, both $\mathbf{X}$ and $\mathbf{Z}$) of the model.

If you're directly, substantively interested in the variance as well, one can plot that as a function of a covariate too. Stata will predict the variance (standard deviation) of an observation, either in- or out-of-sample, and you can plot this against values of a covariate:

Figure 4: Predicted $\hat{\sigma}_i$s, by `cope93` (with other variables held constant at their means/medians).



Other Considerations

Bear a few things in mind...

- When calculating predicted probabilities, marginal effects, etc. for the heteroscedastic probit model, we have to select values for the variables in **Z** as well as for those in **X**...
- Likewise, you need to be sure to vary the values of variable that appear in both **X** and **Z** in both parts of the equation.

Estimation/maximization issues

- This function can be a pain to maximize, particularly if there is quite a bit of collinearity among the variables, and/or if the variables in both parts of the model are identical or nearly so.
- May take LOTS of iterations to converge.

- Even then, may reach a *local maximum*.

How to detect a local maximum

- **Log-Likelihood**: Since the *ln*L for a heteroscedastic probit ought to be at least as large as that for a homoskedastic one, you know that there's a problem if it isn't...
- **Parameter values**: If the standard errors are HUGE, and/or the coefficients are funny-looking, you may have a local maximum.

What to do about local maxima...

- Change the *starting values* (if possible).
- Try a different *maximization algorithm* (if possible).
- Tinker with variable *scaling*, or (just a bit) with model specification (but this should be a last resort).

Applications

When might one want to use this type of model?

- Alvarez/Brehm: Ambivalence about abortion policy.
    - "Elaboration Likelihood Model": Where people have conflicting core beliefs on certain issues, the variance of their responses to survey items ought to be greater than those people or issues without such conflicts.
    - Test this on the issue of abortion (GSS data).
    - They find that when a respondent's core values conflict, they are more ambivalent (variable) in their responses than when no such conflict is present.
    - See also Alvarez and Brehm 1999; Sanders 2001.
- More generally...
    - Any time you might expect changes/differences in the variance of the observations' values on the (binary) dependent variable...
    - E.g., *socialization effects*: If people learn a task (e.g. voting) or become socialized and/or "set in their ways", their variance on items relating to that socialization will probably decrease with time, age, etc.
        - Could mean voters/survey items, or
        - Members of some collegial decision-making body (a legislature, a court, etc.), and voting on bills, cases, etc., or

        - Could even mean nations or other large aggregate entities.
    - Da *democratic peace*: What kind of nations might have more variance in their propensity to engage in international conflict (possibly autocracies?...)
    - *Diffusion studies* (e.g. models of policy innovation in states or nations), where some states (again, possibly autocracies) may have different variance in their propensity to adopt some policy innovation.

The point: This is a (potentially very) useful tool, and can reveal a lot about one's data and offer the possibility of testing some interesting hypotheses.

Bivariate Probit

Consider two related decisions by the same actor, or by different ones...

- Greene talks about the decision to vote for or against property taxes, and the decision to send a child to public or private school... or
- A voter voting for House and Senate candidates on the same ballot, or
- A state adopting two different, but related, policy initiatives (e.g. "three strikes" laws and mandatory sentencing legislation).

The idea is that the two decisions are interrelated...

Consider:

$$\begin{aligned} Y_{1i}^* &= \mathbf{X}_{1i}\boldsymbol{\beta}_1 + u_{1i} \\ Y_{1i} &= 1 \text{ if } Y_{1i}^* > 0 \\ Y_{1i} &= 0 \text{ otherwise} \end{aligned} \tag{10}$$

$$\begin{aligned} Y_{2i}^* &= \mathbf{X}_{2i}\boldsymbol{\beta}_2 + u_{2i} \\ Y_{2i} &= 1 \text{ if } Y_{2i}^* > 0 \\ Y_{2i} &= 0 \text{ otherwise} \end{aligned} \tag{11}$$

In other words, a standard setup for two probit or logit models for Y_1 and Y_2.

Normally, for a probit model we assume that the errors are distributed N(0,1)...

- Implicit in this is that the two models' errors are independent of one another.
- That is, $\text{Cov}(u_{1i}, u_{2i}) = 0$.

- If this is the case, and all the other assumptions hold, we can just estimate the two equations separately, with no problems.

But, *what if the errors in the two equations are related?*

For example, consider what happens if we have:

$$\begin{aligned} u_{1i} &= \eta_i + \epsilon_{1i} \\ u_{2i} &= \eta_i + \epsilon_{2i} \end{aligned}$$

In other words, the errors in each model consist of a part (ϵ_i) that is unique to that model, and a second part (η_i) that is common to both.

- We might assume that all three types of errors are normally distributed...
- If this is true, then the u_is will also be normal, but they will also be *dependent*.
- That is, each u_i now depends, in part, on the value of η_i, and this in turn means that u_{1i} and u_{2i} will be related to one another.

Why should we care?

Because it makes a difference...

We're interested in two things:

$$\begin{aligned} \Pr(Y_{1i} = 1) &= \Pr(u_{1i} > -\mathbf{X}_{1i}\boldsymbol{\beta}_1) \\ &= \Pr(\epsilon_{1i} + \eta_i > -\mathbf{X}_{1i}\boldsymbol{\beta}_1) \end{aligned}$$

and

$$\begin{aligned} \Pr(Y_{2i} = 1) &= \Pr(u_{2i} > -\mathbf{X}_{2i}\boldsymbol{\beta}_2) \\ &= \Pr(\epsilon_{2i} + \eta_i > -\mathbf{X}_{2i}\boldsymbol{\beta}_2) \end{aligned}$$

That is, we're interested in the *joint probability* of Y_1 and Y_2...

- We know that if two random variables are independent, their joint probability is just the product of their marginal probabilities.

- So, if Y_1 and Y_2 are independent, we can say that:

$$\begin{aligned}
\Pr(Y_1 = 1, Y_2 = 1) &= F(Y_1) \times F(Y_2) \\
\Pr(Y_1 = 1, Y_2 = 0) &= F(Y_1) \times [1 - F(Y_2)] \\
\Pr(Y_0 = 1, Y_2 = 1) &= [1 - F(Y_1)] \times F(Y_2) \\
\Pr(Y_0 = 1, Y_2 = 0) &= [1 - F(Y_1)] \times [1 - F(Y_2)]
\end{aligned}$$

- These probabilities will sum to one...

But here, the two probabilities are not independent, since they both depend on the (common) value of η_i...

How do we calculate joint probabilities of nonindependent events?

Remember that:

$$\begin{aligned}
\Pr(A \text{ and } B) &= \Pr(A|B) \times \Pr(B) \\
&= \Pr(A) \times \Pr(B|A)
\end{aligned}$$

This means that in the case we have here, if we want to know:

$$\begin{aligned}
\Pr(Y_1 = 1, Y_2 = 1) &= \Pr(Y_1 = 1|Y_2 = 1) \times \Pr(Y_2 = 1) \\
&\text{or } \Pr(Y_1 = 1) \times \Pr(Y_2 = 1|Y_1 = 1)
\end{aligned}$$

We get at this by assuming a *joint distribution* for the Y's...

- In the independent case, this is easy since $\Pr(Y_1 = 1|Y_2 = 1)$ is just $\Pr(Y_1 = 1)$.
- In the case of dependence, we pick some *bivariate joint distribution.*

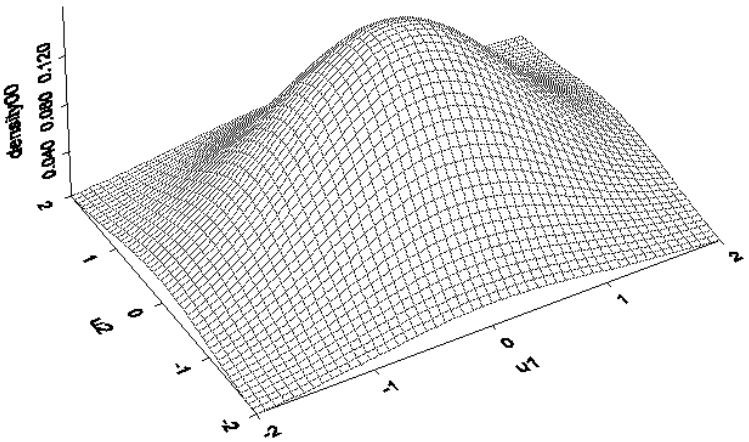
Typically, we use a *bivariate normal distribution*: for two standard-normally distributed us, their joint density will be:

$$\phi(u_1, u_2) = \frac{1}{2\pi\sigma_{u_1}\sigma_{u_2}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\left(\frac{u_1^2 + u_2^2 - 2\rho u_1 u_2}{1-\rho^2}\right)\right] \tag{12}$$

where ρ is a "correlation parameter" denoting the extent to which the two us covary.

Its useful to look at several of these. Figure 5 shows a bivariate normal distribution (often denoted ϕ_2), with $u_1, u_2 \sim \phi_2(0, 0, 1, 1, 0)$ (that is, with *no correlation* between the two us):

Figure 5: Bivariate standard normal density, with $\rho = 0$.



As we increase the correlation between u_1 and u_2, the graph becomes increasingly "ridge-shaped." So, Figure 6 plots the density for $\rho = 0.50$, and Figure 7, for $\rho = 0.90$.

If we integrate with respect to the two variables – that is, if we consider

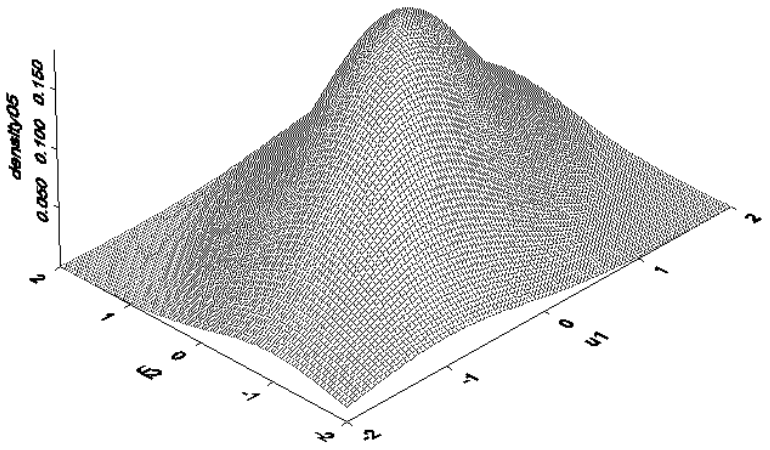
$$\int_{u_1} \int_{u_2} \phi_2(u_1, u_2, \rho) du_1 du_2$$

we get the bivariate normal CDF (sometimes denoted Φ_2). Here, I've plotted it for the case where $\rho = 0.5$:

How to think about ϕ_2

- If $\rho = 0$:
  - The two variables (or errors) are independent, and the Φ_2 reduces to two separate standard normal distributions.
  - The "axes" of the two standard normal distributions, when plotted, will be orthogonal to one another, and the "hill" formed by the density will be "round" or "oval" along the two axes.

Figure 6: Bivariate standard normal density, with $\rho = 0.5$.



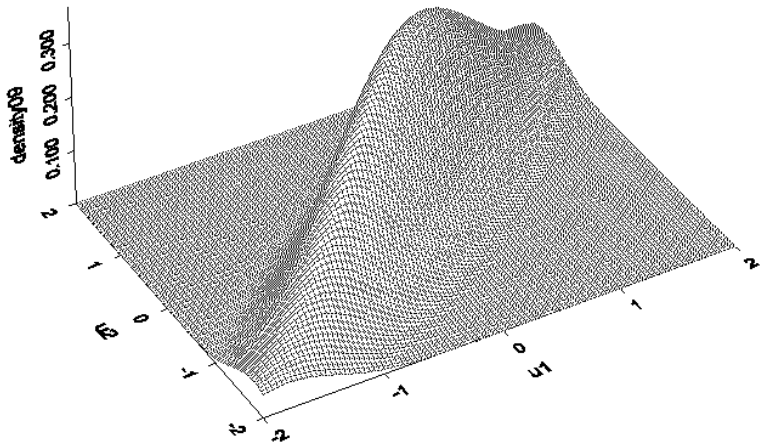
- If $\rho \neq 0$:
  - The two variables/errors will be *correlated*; the probability of one will be dependent on the value/probability of the other.
  - The "axes" of the two distributions will not be perpendicular/orthogonal, but will be at some angle more or less than 90 ... (that is, the "hill" will look more like a "ridge").
- In the extreme case, when $\rho = 1$, the two variables are essentially (actually, exactly) the same.
- Likewise, when $\rho = -1$, the two are exactly negatively correlated (i.e. their scales are reversed).

We use the Φ_2 distribution to estimate bivariate probit models. Typically, we assume:

$$\{u_{1i}, u_{2i}\} \sim \phi_2(0, 0, 1, 1, \rho)$$

- This then gives us the probability statements we need about $\Pr(Y_k = 1)$. Specifically,

Figure 7: Bivariate standard normal density, with $\rho = 0.9$.



$$\begin{aligned} \Pr(Y_{1i} = 1, Y_{2i} = 1) &= \int_{-\infty}^{u_{1i}} \int_{-\infty}^{u_{2i}} \phi_2(\mathbf{X}_{i1}\boldsymbol{\beta}_1, \mathbf{X}_{2i}\boldsymbol{\beta}_2, \rho) du_{1i} du_{2i} \\ &= \Phi_2(\mathbf{X}_{1i}\boldsymbol{\beta}_1, \mathbf{X}_{2i}\boldsymbol{\beta}_2, \rho) \end{aligned} \tag{13}$$

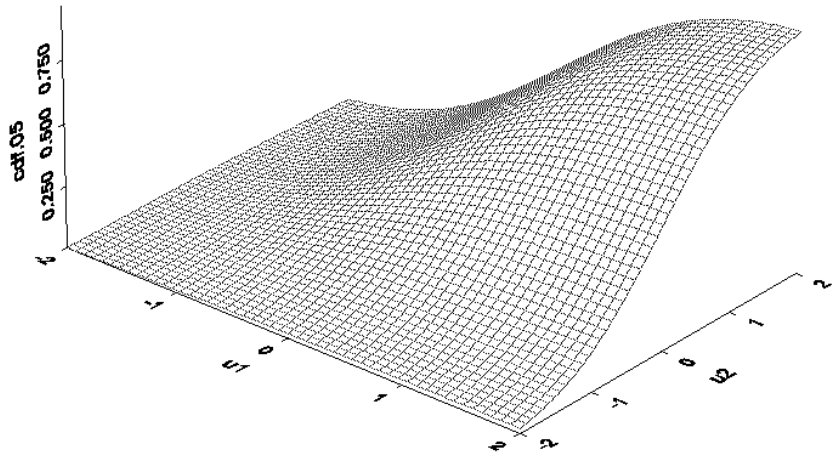
- As in the standard probit model, observations contribute some combination of $\Pr(Y_k = 1)$ for $k \in \{1, 2\}$, depending on their specific values on those variables. The (log-)likelihood is then just a sum across the four possible transition probabilities (that is, the four possible combinations of Y_1 and Y_2) times their associated probabilities; see Greene (2003, 710-11) for a nice description of this.

- Suffice it to say that we estimate the parameters of the model $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \rho)$ via maximum likelihood, usually using the BHHH algorithm.

- The Stata command for this model is `-biprob-`; the general syntax is:

```
. biprob <y1var> <y2var> <xvars> ...
```

Figure 8: Bivariate standard normal CDF, with $\rho = 0.5$.



Model Interpretation

Model Fit

- Separate, independent, binary probit models are "nested" in the bivariate probit model (they occur when $\rho = 0$).
- This means that we can test the hypothesis that the bivariate probit model fits the data better than separate models, using a simple likelihood ratio test...
  - For the separate probit models, the joint likelihood is just the product of the two separate (marginal) likelihoods.
  - This means that the joint log-likelihood is just the sum of the two log-likelihoods.
  - We can compare the joint log-likelihood of the separate models to that for the bivariate probit model using a standard LR test.
- We can also calculate predicted probabilities, and compare them to actual outcomes (a la the PRE for a standard binary probit) – see below for how to do this...

Variable Effects

- We can look at coefficients and s.e.s to gauge direction and statistical significance of individual variable effects, but
- We can't simply apply the usual probit formulas for marginal effects, first derivatives, etc.
- Greene gives formulas for the marginal effects for the bivariate probit estimator...

- These can be useful for computing conditional expected values (aka conditional means).
- E.g.: $E(Y_1|Y_2 = 1, \mathbf{X}_1)$.
- We won't go into these a lot here...

Predicted Probabilities

- The probability of both dependent variables equaling one is:

$$\Pr(Y_{1i} = 1, Y_{2i} = 1) = \Phi_2(\mathbf{X}_{1i}\hat{\boldsymbol{\beta}}_1, \mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_2, \hat{\rho})$$

  where Φ_2 denotes the bivariate normal cumulative distribution function given above.

- The joint probabilities for the other three possible outcomes are:

$$\begin{aligned}
\Pr(Y_{1i} = 1, Y_{2i} = 0) &= \Phi(\mathbf{X}_{1i}\hat{\boldsymbol{\beta}}_1) - \Phi_2(\mathbf{X}_{1i}\hat{\boldsymbol{\beta}}_1, \mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_2, \hat{\rho}) \\
\Pr(Y_{1i} = 0, Y_{2i} = 1) &= \Phi(\mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_2) - \Phi_2(\mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_1, \mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_2, \hat{\rho}) \\
\Pr(Y_{1i} = 0, Y_{2i} = 0) &= 1 - \Phi(\mathbf{X}_{1i}\hat{\boldsymbol{\beta}}_1) - \Phi(\mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_2) - \Phi_2(\mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_1, \mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_2, \hat{\rho})
\end{aligned}$$

- We can use these formulas to construct predicted probabilities (and can also "build" confidence intervals around these predictions if we want).
- We can predict actual predicted values in the data ("in-sample"), or
- we can "simulate" predictions across a range of values for independent variables ("out-of-sample").
- We can then create a table of the changes in probability associated with discrete changes in the independent variables, or
- graph them over a range of values, as in the single-equation case...

An Example

Let's look at some NES data ($N = 1598$), and consider presidential and congressional voting in the 1988 general election. We model two variables:

- `prezgop` is 1 if the respondent voted for Bush, 0 if for Dukakis.
- `housegop` is 1 if the respondent voted for the GOP House candidate, 0 if they voted for the Democrat.

Covariates are few:

- `partyid` is the standard seven-point scale, with 1 = strong Democrat and 7 = strong GOP.
- `ideology` is a seven-point ideological self-placement (1 = "extremely liberal" to 7 = "extremely conservative").
- `demtherm` and `goptherm` are 100-point Democratic and GOP "feeling thermometers", respectively.

The data look like this:

```
. su

  Variable |     Obs        Mean    Std. Dev.       Min        Max
-----------+--------------------------------------------------------
        id |    1598    1005.977    574.1227         4       2039
   prezgop |    1598    .5519399    .4974506         0          1
  housegop |    1598    .4455569    .4971827         0          1
   partyid |    1598    4.122653    2.240958         1          7
  ideology |    1598    4.444305    1.360168         1          7
  demtherm |    1598    59.61452    24.81216         0        100
  goptherm |    1598    61.69712    25.24645         0        100
```

We'll start by running a couple independent probits:

```
. probit prezgop partyid ideology demtherm goptherm if housegop~=.

Probit estimates                          Number of obs   =       1598
                                          LR chi2(4)      =    1363.49
                                          Prob > chi2     =     0.0000
Log likelihood = -417.26564               Pseudo R2       =     0.6203
------------------------------------------------------------------------------
     prezgop |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     partyid |      .2279   .0324578     7.02   0.000     .1642839    .2915161
    ideology |   .2184815    .044875     4.87   0.000     .1305281    .3064349
    demtherm |  -.0374455   .0034766   -10.77   0.000    -.0442595   -.0306314
    goptherm |   .0361379   .0033119    10.91   0.000     .0296466    .0426291
       _cons |   -1.61985   .3231177    -5.01   0.000    -2.253149   -.9865514
------------------------------------------------------------------------------

. probit housegop partyid ideology demtherm goptherm if prezgop~=.

Probit estimates                          Number of obs   =       1598
```

```
                                                LR chi2(4)      =     559.36
                                                Prob > chi2     =     0.0000
Log likelihood = -818.47855                     Pseudo R2       =     0.2547
------------------------------------------------------------------------------
 housegop |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
  partyid |       .24128   .025261     9.55  0.000     .1917694    .2907907
 ideology |     .162891   .0317007     5.14  0.000     .1007588    .2250232
 demtherm | -.0069493   .0019742    -3.52  0.000    -.0108186   -.0030799
 goptherm |    .0022347   .0019352     1.15  0.248    -.0015582    .0060277
    _cons | -1.637397   .2412785    -6.79  0.000    -2.110294   -1.164499
------------------------------------------------------------------------------
```

Everything's pretty much as we'd expect it to be. Next, estimate a bivariate probit:

```
. biprobit prezgop housegop partyid ideology demtherm goptherm

Bivariate probit regression                     Number of obs     =       1598
                                                Model chi2(8)     =    1492.56
                                                Prob > chi2       =     0.0000
Log Likelihood =  -1225.6098862                 Pseudo R2         =     0.3785
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prezgop      |
  partyid    |    .2312522    .032292     7.16  0.000     .1679612    .2945433
 ideology    |    .2297631   .0447584     5.13  0.000     .1420383    .3174879
 demtherm    | -.0360512   .0034408   -10.48  0.000    -.0427951   -.0293073
 goptherm    |    .0348495   .0032825    10.62  0.000     .0284158    .0412831
    _cons    | -1.696476   .3226331    -5.26  0.000    -2.328825   -1.064127
-------------+----------------------------------------------------------------
housegop     |
  partyid    |  .2404114   .0252185     9.53  0.000      .190984    .2898388
 ideology    |  .1629892   .0317416     5.13  0.000     .1007768    .2252016
 demtherm    | -.0070487   .0019761    -3.57  0.000    -.0109217   -.0031757
 goptherm    |  .0022101    .001937     1.14  0.254    -.0015864    .0060065
    _cons    | -1.626407   .2412969    -6.74  0.000    -2.099341   -1.153474
-------------+----------------------------------------------------------------
athrho       |
    _cons    |  .2925434   .060354     4.85  0.000      .1742517    .4108351
-------------+----------------------------------------------------------------
     rho     |    0.28447    0.05547                     0.17251     0.38918
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chi2(1) = 20.2686  Prob > chi2 = 0.00
```

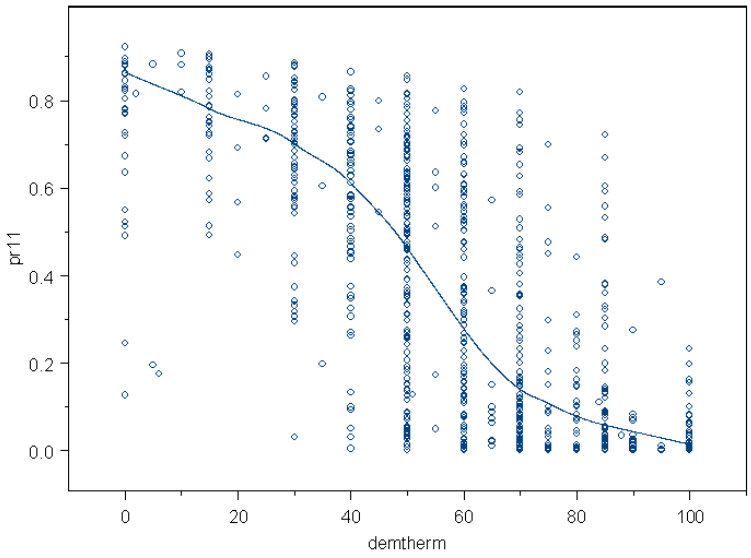
Note that the LR test statistic for the hypothesis that the two equations are independent (which Stata reports for us) is just $-2((-417.27 - 818.48) - (-1225.61)) = -2(-1235.75 + 1225.61) = 20.27$. This is distributed χ^2_1, and is off-the-chart significant here (the highly significant value of $\hat{\rho}$ could also have told us this...).

Once we have our estimates, we can do some predicted probabilities – Stata is good at helping us get those...

```
. predict pr11, p11
```

There are lots of other quantities of interest we can `-predict-`, including standard errors of the linear predictions (which we can use to generate confidence intervals); see the Stata entry for `-biprobit-` for details. Once we have these, we can do cool things like plotting them (see Figure 9, below)...

Figure 9: $\Pr(Y_{1i} = 1, Y_{2i} = 1)$ by `demtherm`, with lowess smooth line.



A Few Final Thoughts

1. Note that, whenever $\hat{\rho} \neq 0$, variables in one part of the model have an indirect influence on the other "part."

2. Relatedly, it can be rough getting this model to converge – the example above converged easily, but that's under very positive circumstances (i.e., lots of data and very few, more-or-less continuous covariates).

3. Finally, model specification is a very important aspect in these models. Think of it this way: if the "true" model for $\Pr(Y_{1i} = 1, Y_{2i} = 1)$ depends on $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_3$, and you omit the latter, your "error" terms are now:

$$\begin{aligned} e_{1i} &= u_{1i} + \beta_3 \mathbf{X}_{3i} \\ e_{2i} &= u_{2i} + \beta_3 \mathbf{X}_{3i} \end{aligned}$$

which, by construction, are correlated. This means that an estimate $\hat{\rho} \neq 0$ may be either due to "actual" correlation between the two processes, or simple specification error.

Models for Rare Events

Sometimes, we study stuff that just doesn't happen all that often...

- Wars,
- Senate confirmation denials,
- etc.

The Problem(s)

Data Collection Issues

When our events are "rare," we often find ourselves gathering data on lots of observations, merely so we can have a sufficient number of "1s" to say something about the phenomenon of interest. (Data collection in international relations is especially prone to do this). This is expensive, and time-consuming, and means we can't spend our time and money collecting better covariates.

Prediction Bias

In addition to wasting lots of data-collection resources, its also the case that regular-old (e.g.) logit models do a pretty crappy job of predicting rare events. In particular, standard models

- underestimate the probability of an event, and
- do so in an increasingly dysfunctional way as the event gets rarer...

The intuition of this is in the King and Zeng articles (e.g., the *PA* paper, pp. 146-47); the basic idea is that $\Pr(\widehat{Y_i = 1}|\mathbf{X})$ will be systematically underestimated in samples in which the $Y = 1$ outcomes are (relatively) rare. This is essentially an issue with classification error – our ability to accurately gauge a "cutting point" for distinguishing $Y = 1|\mathbf{X}$ from $Y = 0|\mathbf{X}$ is "biased in the direction of favoring zeros at the expense of ones." In particular, the bias affects the constant term $\hat{\beta}_0$ directly, and the predictions indirectly.

Formally, if π_i is $\Pr(Y_i = 1)$, and we have a simple one-variable model with $\beta_1 = 1$ (such that $\Pr(Y_i = 1) = \frac{\exp(\beta_0 + X_i)}{1+\exp(\beta_0+X_i)}$), then the bias is about

$$E(\hat{\beta}_0 - \beta_0) \approx \frac{\bar{\pi} - 0.5}{N\bar{\pi}(1 - \bar{\pi})}$$

where $\bar{\pi}$ is just the mean of π_i. When events are rare, so that $\pi_i < 0.5$, this means that

- the bias is always negative (i.e., we'll underestimate the constant term, and therefore consistently underestimate $\Pr(\widehat{Y_i = 1}|X)$),
- for N fixed, as $\bar{\pi} \to 0$ (that is, as the average probability of an event gets smaller), the bias gets worse, but
- as $N \to \infty$ the bias goes away (that is, the estimator is still consistent).

An Alternative Approach

The papers by King and Zeng (in *PA* and *IO*) outline an alternative approach when we're faced with "rare events" data. The data collection approach they suggest involves collecting data on all possible occurrences of "1s" in the data, as well as a random sample of "0s". This is called *choice-based* sampling in econometrics, or *case-control* sampling in statistics and biostatistics. The basic idea is:

1. Figure out the proportion τ of the population which have "1s" (i.e., experience the event of interest).
2. Collect data on all the "1s" in the population.
3. Collect data on a simple random sample of the "0s" as well.
4. Run a logit[1] analysis on the data.
5. "Correct" the coefficients and standard errors after the fact.

There are (at least) two ways of correcting the estimates...

[1] And it *does* have to be a logit; see below.

Sampling

For the sake of simplicity, let's assume (as we often do) that we know the fraction of "1s" in the population of the data;[2] call this proportion τ. Taking data on all the "1s" and a fraction of the "0s" gives us a sample with a mean proportion of "1s" in the data – call this number $\bar{Y}$. So, for IR data, we might observe wars in only 0.1 percent of the cases ($\tau = 0.001$) but actually select a sample with (say) all 1000 instances of conflict, as well as 2000 randomly selected peaceful dyad/years (so that $N = 3000$ and $\bar{Y} = 0.333$).

King and Zeng give some suggestions for how many "0s", relative to the number of "1s", one ought to collect under varying circumstances. Their general suggestion is to collect around 2-5 times as many 0s as 1s, but a ratio of as little as 1/1 (that is, so that $\bar{Y} = 0.5$) can also be fine for some circumstances.

Once we've collected the data in this way, we have to deal with the fact that the data are no longer a purely random sample from the population. There are (at least) a couple ways of doing this.

Weighting

Not surprisingly, econometricians (e.g. Manski and Lerman 1977; also Greene 2002, 673) typically solve the problem of choice-based sampling by weighting the observations. Intuitively, what we want to do is up-weight the 0's and down-weight the 1s, in proportion to their frequency in the population.

Formally, we can think of $w_1 = \frac{\tau}{\bar{Y}}$ as the weights for the 1s, and $w_0 = \frac{1-\tau}{1-\bar{Y}}$ as the weight for the 0s. To the extent that $\tau < \bar{Y}$ (which will always be the case for rare events data), this will have the desired effect. The log-likelihood for this weighted logit then becomes:

$$\ln L(\beta|Y) = \sum_{i=1}^{N} w_1 Y_i \ln\Lambda(\mathbf{X}_i\boldsymbol{\beta}) + w_0(1 - Y_i)\ln[1 - \Lambda(\mathbf{X}_i\boldsymbol{\beta})] \tag{14}$$

Weighting is a good option especially when there is a good chance that the model is misspecified (i.e., almost always, in political science). King and Zeng note that, while intuitively reasonable, there are reasons that the weighting approach may not be optimal:

- Weighting is be less efficient than the prior correction strategy outlined below – making it less useful in small samples,
- Standard errors from the weighted regression are *way* off, and

[2]This would be the case, for example, if we used EuGENE or some similar comprehensive database as our population, and then sampled from it. If we don't know τ, then we have to consider an alternative approach, such as that outlined in King and Zeng (2002 *Statistics in Medicine*).

- before K&Z, finite sample corrections weren't an option.

There's also...

Prior Correction

With case-control data, its pretty straightforward to show[3] that all the usual estimates for $\beta_1 ... \beta_k$ are consistent; the only problem is the intercept, which is biased. Thus, a correction for the standard MLE of the intercept $\hat{\beta}_0$ is:

$$\hat{\beta}_{0pc} = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{Y}}{1-\bar{Y}}\right)\right] \tag{15}$$

Note that, for population data, $\tau = \bar{Y}$ and so the correction doesn't do anything.

Bias Correction

King and Zeng note that with case-control data, we still have bias in our usual estimates of $\boldsymbol{\beta}$. King and Zeng show that the bias is

$$\text{bias}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\xi$$

where ξ is a (rather complicated) combination of the weights w_i discussed above, the predicted probabilities $\hat{\pi}_i$ and the $\mathbf{X}$ matrix. Correcting our usual logit estimates [by considering $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \text{bias}(\hat{\boldsymbol{\beta}})$] gives us the "right" ("corrected") coefficient estimates.

Probability Calculations

Once we've calculated the "corrected" parameters, it seems like we can just plug that number into our standard logit equation to generate predictions:

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

So, it seems like all we need to do is correct our $\hat{\boldsymbol{\beta}}$, substitute them in, and we're off to the races, no?

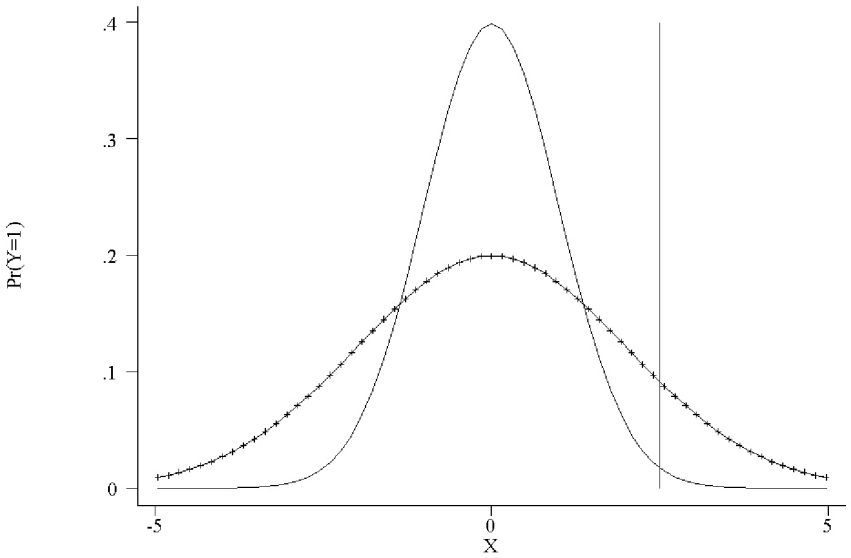
Um, no.

In fact, doing this ignores the variability (i.e., uncertainty) in the estimates $\tilde{\boldsymbol{\beta}}$ – that is, the variability due to the fact that $\tilde{\boldsymbol{\beta}}$ is itself a random variable. This is more than just an innocuous effect, too, since (as King & Zeng note), estimates which have too small a variance

[3] K&Z attribute this to various people, though McCullagh and Nelder (1989) cite Armitage (1971). A recent, brief, technical exposition is Abram Kagan, 2001, "A Note on the Logistic Link Function," *Biometrika* 88:599-601.

will (again) result in underprediction of $\widehat{\Pr(Y_i = 1)}$ (see their figure, p. 149 in *PA*, or Figure 10).

Figure 10: The Effect of Uncertainty on Probability Predictions. Areas to the right of the line are $\Pr(Y = 1)$.



K&Z offer a way of including this variability back into one's predictions. The basic idea is to use the corrected estimates $\tilde{\beta}$ to get predicted probabilities $\tilde{\pi}_i$, and then in turn correct these predictions to

$$\Pr(Y_i = 1) \approx \tilde{\pi}_i + C_i$$

where

$$C_i = (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)\mathbf{X}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})\mathbf{X}_i'$$

Notice a couple of things about this correction:

- The uncertainty is captured in the $\mathbf{V}(\tilde{\boldsymbol{\beta}})$ term, which reflects the extent to which there is intrinsic variability in $\tilde{\boldsymbol{\beta}}$.
- The fact that, in rare-event circumstances, $\tilde{\pi}_i < 0.5$ means that C_i will be positive for those instances – that is, it will add to the probability (which is what we want, cf. Figure 10).

An Example

If all this seems a bit complicated, well, it is. Happily, K&Z have put together a Stata routine, called `-relogit-`, to implement the corrections they outline in their article (and even to calculate predicted probabilities and the like, a la Clarify).

Consider some data from Oneal and Russett (1997) (also Beck, Katz and Tucker (1998) and others) on international disputes (specifically, politically-relevant dyad-years, 1950-1985). We have $N = 20448$, with 405 dyad-years of MIDs. We'll run a basic logit model with six variables: *democracy*, *growth*, *alliance*, *contiguity*, *capability ratio* and *trade*.

First, the basic logit model, on all the data (and some predictions):

```
. tab1 dispute

-> tabulation of dispute

    dispute |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      20043       98.02       98.02
          1 |        405        1.98      100.00
------------+-----------------------------------
      Total |      20448      100.00

. logit dispute dembkt grobkt allies contig capbkt trade

Logit estimates                           Number of obs   =      20448
                                          LR chi2(6)      =     284.79
                                          Prob > chi2     =     0.0000
Log likelihood = -1846.8783               Pseudo R2       =     0.0716
------------------------------------------------------------------------------
 dispute |   Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  dembkt |-.4011976   .1006345  -3.99   0.000    -.5984376   -.2039576
  grobkt | -3.42753   1.251813  -2.74   0.006    -5.881038   -.9740222
  allies |-.4796863   .1127463  -4.25   0.000    -.7006649   -.2587076
  contig | 1.353576   .1209052  11.20   0.000     1.116606    1.590546
  capbkt |-.1961988   .0501055  -3.92   0.000    -.2944037   -.0979938
   trade |-21.07611   11.30396  -1.86   0.062    -43.23146    1.079242
   _cons |-4.326677   .1145089 -37.78   0.000     -4.55111   -4.102243
------------------------------------------------------------------------------

. predict dumbprobs
(option p assumed; Pr(dispute))
```

Next, we'll select out all the "1s" (disputes) as well as a 5 percent sample of the "0s" (non-disputes)...

```
. gen random=.
(20448 missing values generated)

. replace random=invnorm(uniform()) if dispute==0
(20043 real changes made)

. gen select=0

. replace select=1 if dispute==1
(405 real changes made)

. su random, detail
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -2.348678      -3.794222
 5%    -1.649082      -3.786709
10%    -1.274943      -3.714839       Obs                20043
25%    -.6673297      -3.579602       Sum of Wgt.        20043

50%    -.0052992                      Mean           -.0004152
                        Largest       Std. Dev.       .9976583
75%     .6688103       3.449271
90%     1.267455       3.629463       Variance         .995322
95%     1.650757       3.700645       Skewness         .005595
99%     2.378313       3.915892       Kurtosis        3.006424

. replace select=1 if random>1.65 & random~=.
(1004 real changes made)

. tab1 dispute if select==1

-> tabulation of dispute if select==1

    dispute |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |       1004       71.26       71.26
          1 |        405       28.74      100.00
------------+-----------------------------------
      Total |       1409      100.00
```

Now, check out what happens when we run a standard logit on the "selected" data only:

```
. logit dispute dembkt grobkt allies contig capbkt
  trade if select==1

Logit estimates                           Number of obs   =       1409
                                          LR chi2(6)      =     191.85
                                          Prob > chi2     =     0.0000
Log likelihood = -749.25422               Pseudo R2       =     0.1135
------------------------------------------------------------------------------
 dispute |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  dembkt | -.4765852   .1211715    -3.93   0.000    -.7140771   -.2390934
  grobkt | -2.708746   1.599052    -1.69   0.090    -5.842829    .4253377
  allies | -.2583556    .142826    -1.81   0.070    -.5382893    .0215781
  contig |  1.279885   .1436829     8.91   0.000     .9982717    1.561498
  capbkt | -.1742606   .0508522    -3.43   0.001     -.273929   -.0745922
   trade | -16.92824   11.84484    -1.43   0.153    -40.1437    6.287222
   _cons | -1.446798   .1297272   -11.15   0.000    -1.701059   -1.192537
------------------------------------------------------------------------------
```

The results aren't all that different, except for the constant (which, unsurprisingly, is a *lot* larger in the latter analysis).

Next, we use the `-relogit-` command to correct these estimates for the biases K&Z talk about. There are two ways of going about this, one which uses the weighted regression approach in (14), the other which uses the prior correction method in (15). We'll do both here:

```
. relogit dispute dembkt grobkt allies contig capbkt
  trade if select==1, pc(.02)

Corrected logit estimates                    Number of obs =     1409
------------------------------------------------------------------------------
         |               Robust
 dispute |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  dembkt | -.475866   .1199331    -3.97   0.000    -.7109306   -.2408014
  grobkt |-2.695515   1.567849    -1.72   0.086    -5.768443    .3774127
  allies |-.2559159   .1464495    -1.75   0.081    -.5429517    .0311199
  contig | 1.274234   .1422197     8.96   0.000     .9954883    1.552979
  capbkt |-.1654231    .058356    -2.83   0.005    -.2797987   -.0510474
   trade |-14.16053   12.70195    -1.11   0.265     -39.0559    10.73484
   _cons |-4.432804   .1286873   -34.45   0.000    -4.685027   -4.180582
```

```
------------------------------------------------------------------------------

. relogit dispute dembkt grobkt allies contig capbkt
  trade if select==1, wc(.02)

Corrected logit estimates                       Number of obs =     1409
------------------------------------------------------------------------------
             |               Robust
     dispute |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dembkt |  -.4558843   .1220236    -3.74   0.000    -.6950462   -.2167224
      grobkt |  -3.853615   1.674077    -2.30   0.021    -7.134746   -.5724839
      allies |  -.3319833   .1502555    -2.21   0.027    -.6264787   -.0374879
      contig |   1.293922    .149677     8.64   0.000     1.000561    1.587284
      capbkt |  -.1744629   .0607157    -2.87   0.004    -.2934635   -.0554622
       trade |  -14.18243    13.9973    -1.01   0.311    -41.61663    13.25176
       _cons |  -4.392938   .1332107   -32.98   0.000    -4.654027    -4.13185
------------------------------------------------------------------------------
```

Not surprisingly, the two don't look all that different.

Finally, `-relogit-` makes it easy for us to generate "quantities of interest" (mainly predictions), using the same `-setx-` command we used with ℭlarify) to set the values of the independent variables, and then using a corollary command (`-relogitq-`) to create the probabilities. Check the documentation on that...

Wrap-Up

∃ several important things to remember about this rare-events stuff...

- You're better off knowing this stuff in advance – the whole point of all this is to make it possible to do case-control studies (which, both implicitly and explicitly, K&Z advocate). Its a good approach for a lot of things we might study (not just wars...).
- In the example above, the corrections to the predictions don't really make that much of a difference in the actual outcomes. Which is to say: *Some things are just rare, and (therefore) hard to predict.* No statistical technique can make up for that fact.