Binary Response Models

1 Introduction

There are many binary social outcomes that occur naturally:

- A citizen votes or not
- A cabinet forms or not
- A pre-electoral coalition forms or not
- A war is fought or not

1.1 Linear Probability Model

You might think to run OLS in this situation - the linear probability model (LPM). In other words, you'd model the expected value of Y as a linear function of some independent variables X.

$$E[y] = x_i\beta \tag{1}$$

In the binary world, the expected value of Y is:

$$\begin{aligned} E[y] &= 1 \times \Pr(y = 1) + 0 \times \Pr(y = 0) \\ &= \Pr(y = 1) \end{aligned} \tag{2}$$

From this, we can see that we have:

$$\Pr(y = 1) = x_i\beta \tag{3}$$

This model is obviously very easy to estimate and you can interpret the coefficients as you would from a normal OLS model i.e. the coefficients indicate how a one-unit change in X affects $\Pr(y = 1)$.

There are some problems with the LPM, though.

1. **Unbounded Predicted Values:** $x_i\beta$ can take on values greater than 1 and less than 0.
2. **Conditional Heteroskedasticity:** The variance of the residual is related to the value of x. Specifically,

$$\begin{aligned} \text{var}(y) &= E[y](1 - E[y]) \\ &= x_i\beta(1 - x_i\beta) \end{aligned} \tag{4}$$

As this illustrates, the variance of y depends on the values of X and β and is, therefore, heteroskedastic by construction.

3. **Non-Normal Errors:** The errors can only take on two values, $1 - x_i\beta$ or $-x_i\beta$. As a result, the errors can never be normally distributed, therefore causing problems for hypothesis testing.
4. **Functional Form:** Given the nature of probabilities, we'd expect that the marginal impact of an independent variable would exhibit diminishing returns; that is, as the value of the independent variable increases, its impact on y should decrease. The LPM does not allow for this possibility.

It should be pointed out that none of these problems actually cause a problem with the point estimates; that is, they do not cause bias. The OLS point estimates remain unbiased estimates of the true parameter values of β.

The bottom line, though, is that we should come up with a better approach to dealing with binary dependent variables.

2 Logit and Probit

There are three ways of coming up with a better approach than the LPM:

1. Pure Probability Approach
2. Latent Variable Approach
3. Random Utility Approach

They all lead us to the logit and probit models.

2.1 Pure Probability Approach: The Bernoulli Distribution

We have already encountered a distribution for outcomes which take on only two values - the Bernouilli distribution

$$\begin{aligned} f(1) &= \pi \\ f(0) &= 1 - \pi \end{aligned} \tag{5}$$

where the event occurs with probability π and fails to occur with probability $1 - \pi$. Recall that the likelihood of the Bernoulli distribution is

$$\mathcal{L} = \prod_{i=1}^{N} \pi^{y_i}(1-\pi)^{1-y_i} \tag{6}$$

and that the log-likelihood is

$$\ln\mathcal{L} = \sum_{i=1}^{N}[y_i \ln(\pi) + (1 - y_i)\ln(1 - \pi)] \tag{7}$$

The Bernoulli distribution would be an appropriate model of dichotomous choice if each event had the same chance of occurring. However, it is not a good model for widely variable outcomes. For example, we would not think that all voters have the same probability of voting or that all countries have the same probability of going to war. Thus, the Bernoulli distribution is too restrictive as it stands. Instead we would like to let π vary across cases i.e. π_i. This keeps the Bernoulli form but allows us to capture variation across cases in the probability. However, we can't just let each observation have its own π_i since the model would not be identified. This is why we write $\pi_i = g(X, \beta)$ in order to both reduce the number of parameters and to add substantive explanatory variables. In effect, we have

$$Y_i \overset{bern}{\sim} (\pi_i) \tag{8}$$

where $\pi_i = g(x_i, \beta)$. All we need now is to find a function $g(\cdot)$ of the xs and the βs to substitute into the Bernoulli likelihood function i.e.

$$\mathcal{L} = \prod_{i=1}^{N} g(x_i, \beta)^{y_i}(1 - g(x_i, \beta))^{1-y_i} \tag{9}$$

or, alternatively, the log-likelihood function i.e.

$$\ln\mathcal{L} = \sum_{i=1}^{N}\{y_i \ln[g(x_i, \beta)] + (1 - y_i)\ln[(1 - g(x_i, \beta)]\} \tag{10}$$

2.2 What functional form should we choose for $g(\cdot)$?

We know that π_i represents a probability. For simplicity, let's assume that $f(x_i, \beta) = g(x_i\beta)$. Thus, we want a function that satisfies the following properties

1. $g : \Re \rightarrow (0, 1)$
2. $\lim_{x_i\beta \rightarrow \text{inf}} = 1$
3. $\lim_{x_i\beta \rightarrow -\text{inf}} = 0$

There are an infinite number of possible functions that fit these criteria, but political scientists have settled on a number of 'S-shaped' curves for g. The most popular are two cumulative distribution functions:

- The cumulative standard normal distribution, $\Phi(x_i\beta)$

- The cumulative standard logistic distribution, $\Lambda(x_i\beta)$

2.2.1 Logit

The cumulative standard logistic is

$$\Pr(y_i = 1|x_i) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} = \frac{1}{1 + e^{-x_i\beta}} = \Lambda(x_i\beta) \tag{11}$$

where $x_i\beta$ is just a linear function of some sort. Substituting this in for $g(\cdot)$ gives the following likelihood function

$$\mathcal{L} = \prod_{i=1}^{N} [\Lambda(x_i\beta)]^{y_i} [1 - \Lambda(x_i\beta)]^{1-y_i} \tag{12}$$

and the following log-likelihood function

$$\begin{aligned} \ln\mathcal{L} &= \sum_{i=1}^{N} \{y_i \ln[\Lambda(x_i\beta)] + (1 - y_i)\ln[1 - \Lambda(x_i\beta)]\} \\ &= \sum_{i=1}^{N} \left\{ y_i \ln\left[\frac{1}{1 + e^{-x_i\beta}}\right] + (1 - y_i)\ln\left[1 - \frac{1}{1 + e^{-x_i\beta}}\right] \right\} \end{aligned} \tag{13}$$

Since the log-likelihood function is non-linear, there is no closed-form solution for β. However, numerical maximization is easy since the log-likelihood is globally concave as we'll see.

The Gradient vector is

$$G = \frac{\partial \ln\mathcal{L}}{\partial\beta} = \sum_{i=1}^{n} (y_i - \Lambda(x_i\beta))x_i \tag{14}$$

The Hessian matrix is

$$H = \frac{\partial^2 \ln\mathcal{L}}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n} \Lambda(x_i\beta)[1 - \Lambda(x_i\beta)]x_i x_i' \tag{15}$$

Note that $\Lambda(x_i\beta)$ and $1-\Lambda(x_i\beta)$ are always between 0 and 1. $x_i x_i'$ is the $k \times k$ matrix of the squares and cross-products of the k independent variables. This will be positive definite. As a result, the Hessian will be negative definite everywhere i.e. the log-likelihood function is globally concave and so maximization will be easy.

2.2.2 Probit

The cumulative standard normal is

$$\Pr(y_i = 1|x_i) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i\beta)^2}{2\sigma^2}} dx = \Phi(x_i\beta) \tag{16}$$

where $x_i\beta$ is just a linear function of some sort. The integral doesn't have a closed form solution, which is why we normally abbreviate it as $\Phi(x_i\beta)$. Substituting this in for $g(\cdot)$ gives the following likelihood function

$$\mathcal{L} = \prod_{i=1}^{N} [\Phi(x_i\beta)]^{y_i} [1 - \Phi(x_i\beta)]^{1-y_i} \tag{17}$$

and the following log-likelihood function

$$\ln\mathcal{L} \quad = \quad \sum_{i=1}^{N} \{y_i \ln[\Phi(x_i\beta)] + (1 - y_i)\ln[1 - \Phi(x_i\beta)]\} \tag{18}$$

Because of the symmetry of the normal density, we can express $1 - \Phi(x_i\beta)$ as $\Phi(-x_i\beta)$. This means that we can express the log-likelihood function as

$$\ln\mathcal{L} \quad = \quad \sum_{i=1}^{N} \{y_i \ln[\Phi(x_i\beta)] + (1 - y_i)\ln[\Phi(-x_i\beta)]\} \tag{19}$$

Since the log-likelihood function is non-linear, there is no closed-form solution for β. However, numerical maximization is easy since the log-likelihood is globally concave.

2.3 A Latent Variables Approach

The latent variable approach essentially treats dichotomous variables as a problem of measurement. In effect, there exists a continuous underlying or latent variable but we just haven't measured it. Instead, we have only a dichotomous indicator of the latent variable. To see how this works, suppose that there is some unobserved or unmeasured (latent) variable such that we had the following regression

$$y_i^* = x_i\beta + \epsilon_i \tag{20}$$

where we assume that ϵ has mean 0 and has either a standard logistic distribution with (known) variance $\frac{\pi^2}{3}$ or a standard normal distribution with (known) variance 1.

We do not observe the latent variable, y_i^*, itself. All we observe is:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > t \\ 0 & \text{if } y_i^* \leq t \end{cases}$$

where t is some threshold. For convenience, we assume $t = 0$ (As it turns out, we'll have to make an assumption of this sort if we want to identify the constant term). Thus, we have

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

This means that probit or logit is just regression with less information; that is, we know all about the covariates but only the sign of the dependent varible - y_i indicates whether the latent variable y_i^* is positive or negative.

Brief Aside on Assumptions:

1. **Known variance of ϵ:** This is an innocent assumption. For example, say that the variance of ϵ is scaled by an unrestricted parameter σ^2. The latent regression in this case would be $y^* = x\beta + \sigma\epsilon$. However, this can be rewritten as $\frac{y^*}{\sigma} = x\frac{\beta}{\sigma} + \epsilon$. Note that this is the same model with the same data. The observed data will be unchanged: y is still 0 or 1, depending only on the sign of y^* and not on its scale. Why make these particular assumptions about the variance of ϵ? Essentially, because they make the distributions easier to handle.

2. **Threshold is 0:** This is an innocent assumption so long as the model contains a constant term. Let a be the supposed non-zero threshold and α be an unknown constant term. Let x not include the constant term for now. If this is the case, then

$$\Pr(y^* > a) = \Pr(\alpha + x\beta + \epsilon > a) = \Pr((\alpha - a) + x\beta + \epsilon > 0) \tag{21}$$

   Since α is unknown, the difference $(\alpha - a)$ remains an unknown parameter. As a result, we can arbitrarily (but helpfully) set the threshold as zero (this will just affect the constant term).

With these assumptions, we have

$$\begin{aligned} \Pr(y = 1) &= \Pr(y^* > 0) \\ &= \Pr(x\beta + \epsilon > 0) \\ &= \Pr(\epsilon > -x\beta) \\ &= 1 - \Pr(\epsilon < -x\beta) \\ &= 1 - F(-x\beta) \end{aligned} \tag{22}$$

where F is the cumulative distribution of ϵ i.e. either the standard logistic or the standard normal. If F is symmetric about 0 (as is the case with logit and probit), we have

$$\begin{aligned} \Pr(y = 1) &= 1 - F(-x\beta) \\ &= F(x\beta) \end{aligned} \tag{23}$$

Thus, if we choose a standard normal distribution, we have a probit model

$$\Pr(y = 1) = F(x\beta) = \Phi(x\beta) \tag{24}$$

and if we have a standard logistic model,we have a logit model

$$\Pr(y = 1) = F(x\beta) = \Lambda(x\beta) \tag{25}$$

2.4 Random Utility Approach

The random utility approach to modeling dichotomous dependent variables has its origins in microeconomic theory. Imagine a situation where an individual is choosing between a set of alternatives or outcomes. The random utility model assumes that a decision maker attaches a utility to each alternative that he must choose between and that this utility is based on both a systematic and stochastic component. It is assumed that the individual will choose the alternative or outcome that maximizes the utility gained from that choice.

$$U_{im} = V_{im} + \epsilon_{im} \tag{26}$$

where U_{im} is individual i's utility for alternative m, V_{im} is the systematic component of utility for individual i associated with choice m, and ϵ_{im} is the stochastic component of utility for individual i associated with choice m. Notice that e_{im} is subscripted by both i and m. In other words, there is one disturbance per individual per choice. We assume that the systematic component of the utility is just a linear function of some variables i.e.

$$V_{im} = X_{im}\beta_m \tag{27}$$

So, in effect, the utility of individual i for choice m is:

$$U_{im} = X_{im}\beta_m + \epsilon_{im} \tag{28}$$

In this setup, we assume that individual i chooses choice m iff:

$$U_{im} > U_{ij} \quad \forall\ j \neq m \tag{29}$$

The random utility setup is very flexible and allows us to look at 2 or more choices.[1] As you can see, random utility models (RUM) are framed in terms of a particular model of behavior based on utility maximization that you are probably all familiar with. Now, let's look at how the RUM framework is used in the specific case of logit and probit.

Say individual i must choose between two alternatives. The probability of choosing alternative 1 is just the probability that the utility from alternative 1 exceeds the utility from alternative 2 i.e.

$$\begin{aligned} P(y_i = 1) &= P(U_{i1} > U_{i2}) \\ &= P(V_{i1} + \epsilon_{i1} > V_{i2} + \epsilon_{i2}) \\ &= P(\epsilon_{i2} - \epsilon_{i1} < V_{i1} - V_{i2}) \end{aligned} \tag{30}$$

Obviously, to calculate this probability, $P(U_{i1} > U_{i2})$, we need to make some distributional as-

[1] In a few weeks time, we will apply the random utility model setup to the situation where we have more than two choices.

sumptions about ϵ_{im}. More specifically, we need to make some distributional assumptions about the difference between the disturbances i.e. $\epsilon_{i1} - \epsilon_{i2}$. It turns out that if the disturbances (ϵ_{im}) are identically and independently distributed according to the Gumbel distribution (otherwise known as Type 1 Extreme Value distribution), then their difference ($\epsilon_{im} - \epsilon_{ij}$) is distributed according to the logistic distribution. This takes us back to the logit model.[2] To get the probit model, we assume that the disturbance terms are mean-zero normally distributed; it turns out that the difference between two normally distributed variables is also normally distributed.

3 Comparing Logit and Probit

3.1 Which should you use?

As you can see, the setup of the logit and probit models (whichever derivation you prefer) is essentially the same. Empirically we can't really tell which model fits the data best. Since logit and probit are not nested, we cannot use tests like the LR test or the Wald test to distinguish between them. Logit is generally used these days because it is numerically simpler - but computers have made this distinction meaningless in a practical sense. For historical reasons, some things (ordered probit) are always done in a probit context, whereas others (multi-choice unordered logit) are always done in a logit context. There is no reason why we cannot substitute logit or probit in these contexts.

3.2 Coefficients and Predicted Probabilities

The coefficients from a probit model will be different to those from a logit model since the transformation from the coefficient to a probability in probit is different from the equivalent transformation in logit. However, logit and probit will produce similar predicted probabilities, $\hat{\pi}$. Why? Since MLE is choosing the parameters so that we get π_i as close to 1 when $y_i = 1$ and π_i as close to 0 when $y_i = 0$, both logit and probit should generate similar $\hat{\pi}$s.

[2] Thus, to get the logit model in our random utility model setup, we assume that the cumulative distribution function of the disturbances is the Gumbel or Type 1 Extreme Value distribution:

$$F(\epsilon_{im}) = \exp(-\exp^{-\epsilon_{im}}) \tag{31}$$

and that the probability distribution function of the disturbances is:

$$f(\epsilon_{im}) = \exp^{-\epsilon_{im}} \exp(-\exp^{-\epsilon_{im}}) \tag{32}$$

3.3 Variance and Coefficients

You will recall that we made assumptions about σ^2 in both logit and probit. The basic reason for this is that there is not enough information in a binary y to identify both β and σ^2. In the probit model, we generally assume that $\sigma^2 = 1$ and in the logit model we generally assume that $\sigma^2 = \frac{\pi^2}{3}$. These assumptions are entirely arbitrary in the sense that they cannot be disconfirmed by the data. For example, you could use $\sigma^2 = 1$ in a logit model; if you did, the results from the logit model would be almost identical to those from the probit model. This illustrates that the reason why the coefficients from logit and probit differ so much is that they are measured on different scales. You can put the coefficients on roughly the same scale by multiplying the probit coefficients by $\frac{\pi^2}{3} = 1.81$. If you did this, the coefficients from probit and logit would be very similar (though not exactly the same).

4 Interpretation - Quantities of Interest

There are various quantities of interest that one might want to calculate. We will focus on (i) predicted probabilities, (ii) marginal effects, and (iii) first differences. We will focus on the following basic model.

$$\text{PEC}_i^* = \beta_0 + \beta_1 \textsc{Parties} + \beta_2 \textsc{Threshold} + \epsilon \tag{33}$$

where PEC* is the latent variable that is assumed to be less than or equal to zero when we do not observe a pre-electoral coalition and greater than zero when we do. PEC* captures the latent propensity of two parties to form a pre-electoral coalition. Parties is the effective number of parties in the system and Threshold is the electoral threshold.

4.1 Predicted Probabilities

Recall that

$$\Pr(y_i = 1|x_i) = F(x_i\beta) \tag{34}$$

where F is either the cumulative standard normal Φ or the cumulative standard logistic Λ. Thus, to calculate the predicted probability that $y = 1$, you just need to set the independent variables to substantively relevant values and use the equation shown above. Say we wanted to calculate the predicted probability that a pre-electoral coalition forms when the effective number of parties is 2 and when the electoral threshold is 8.[3]

[3] Oftentimes, we will set variables to their means (continuous variables) or their modes (dichotomous variables). However, you should decide the values that are most substantively relevant for your particular theoretical question.

4.1.1 Computing Predicted Probabilities

There are at least three ways of calculating this predicted probability:

1. **By Hand (in STATA):**
   For probit, we might have:

   ```
   probit pec enep threshold

   display normal(_b[_cons] + _b[enep]*2 +_b[threshold]*8)
   ```

   For logit, we might have:

   ```
   logit pec enep threshold

   display 1/(1+exp(-(_b[_cons] + _b[enep]*2 +_b[threshold]*8)))
   ```

   Note that these commands will not give you any measure of uncertainty around your quantity of interest. You should always have a measure of uncertainty around a quantity of interest. We have already seen how to calculate this via simulation. So you might just want to type

   ```
   probit pec enep threshold

   preserve

   drawnorm MG_b1-MG_b3, n(1000) means(e(b)) cov(e(V)) clear

   save simulated_betas, replace

   restore

   merge using simulated_betas

   summarize _merge

   drop _merge

   summarize

   scalar  h_enep = 2

   scalar h_threshold = 8

   scalar h_constant = 1
   ```

```
generate x_betahat = MG_b1*h_enep + MG_b2*h_threshold + MG_b3*h_constant

generate prob_hat = normal(x_betahat)

sum prob_hat

centile prob_hat, centile(2.5 97.5)
```

2. **STATA command: prvalue**

   You may want to look at the following book: Long, J. Scott & Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using STATA* (Second Edition). STATA Press. This book describes a package called SPOST that can be very helpful in calculating quantities of interest. To install SPOST, open STATA and type the following:

   ```
   net from http://www.indiana.edu/~jslsoc/stata/

   net install spost9_ado

   net get spost9_do
   ```

   You can then use various commands to calculate quantities of interest. To calculate the predicted probability (and confidence intervals) that we want, you would type

   ```
   probit pec enep threshold

   prvalue, x(enep=2 threshold=8)

   logit pec enep threshold

   prvalue, x(enep=2 threshold=8)
   ```

   The only disadvantage with prvalue is that you do not know the underlying code used to get the predicted probability and confidence intervals.

3. **CLARIFY**

   You can also calculate the predicted probability using CLARIFY. If you have not already installed CLARIFY, you need to type the following:

   ```
   net from http://gking.harvard.edu/clarify

   net install clarify

   net get clarify
   ```

   Once installed, you would type:

```
estsimp probit pec enep threshold

setx enep 2 threshold 8

simqi
```

You should note that the predicted probability and confidence intervals from these three different methods will not be identical - why?

4.1.2 Presenting Predicted Probabilities

There are various ways of presenting predicted probabilities - tables and graphs etc. Whichever method you use, make sure to show the appropriate measure of uncertainty as well.

While it is relatively easy to get predicted probabilities and we often see articles reporting or graphing predicted probabilities, you should ask yourself how often this is the relevant quantity of interest. After all, when we do OLS, how often is it that we report the predicted values of the dependent variable. Instead, the quantity of interest is often how a change in x affects y i.e. the marginal effect of x on y.

4.2 Marginal Effects

We like linear regression because it is easy to compute the impact of a change in an independent variable on y. Some people say that coefficients from probit or logit are difficult to interpret or even uninterpretable. This is nonsense.

If you are interested in understanding how the independent variables affect the unobserved latent variable y^*, then the probit and logit coefficients can be interpreted in exactly the same way as OLS coefficients i.e. the coefficients tell you how much y^* changes with a one unit increase in the independent variables. Of course, it is almost never the case that you will be interested in y^*. Instead, you want to know the effect of your independent variables on y i.e. the probability of getting a 1 or a 0.

However, even if you are interested in y and not y^*, the logit and probit coefficients will still tell you the (i) direction and (ii) statistical significance associated with the effect of increasing an independent variable just like OLS coefficients. Thus, a positive coefficient β_k tells you that an independent variable X_k increases the probability that $y = 1$. If β_k is significant, then we can say that this positive effect will be statistically significant. BUT what you cannot say by looking at the probit and logit coefficients is how much the probability that $y = 1$ will change when you change X_k i.e. the coefficients tell you nothing about the magnitude of the effect of a change in X_k on $\Pr(y = 1)$. To find the magnitude of the effect, you will need to do a little work - but it is not hard.

4.2.1 Marginal Effects in OLS

In a linear model, we might have

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon \tag{35}$$

You learned that β_1 tells you the effect of a one unit increase in X_1 on y. However, you should have learned that β_1 is just the marginal effect of y with respect to X_1 i.e.

$$\frac{\partial y}{\partial X_k} = \beta_k \tag{36}$$

As you can see, the 'effect' of x on y is really a derivative (it is very helpful to always think in terms of marginal effects and derivatives). It is the linearity of this model which makes the derivative trivial and hence makes the discussion of the marginal effect of X_k on y simply β_k. This holds regardless of what values of X_k or y we are considering and regardless of what values other variables in the model take on.

4.2.2 Marginal Effects in Logit and Probit

In logit and probit the marginal effects are not constant (i.e. they are not some β_k) because the relationship between x and y is not linear. This is where people start to get confused for some reason. However, you should just go back to recognizing that the effect of x on y in any model is just a derivative. In our binary response model, we have parameterized π_i as $F(X_i\beta)$ where F is the cumulative standard logistic or the cumulative standard normal. Thus, the marginal effect is just:

$$\frac{\partial y_i}{\partial x_k} = \frac{\partial F(x_i\beta)}{\partial x_k} = \frac{\partial F(x_i\beta)}{\partial x_i\beta} \cdot \frac{\partial x_i\beta}{\partial x_k} = F'(x_i\beta)\beta_k = f(x_i\beta)\beta_k \tag{37}$$

by application of the chain rule (derivative of the outside function multiplied by the derivative of the inside function). Note that the effect of any variable x_k will be affected by the values of all the other x variables through $f(x_i\beta)$.

Specifically, in the case of probit we have:

$$\frac{\partial y_i}{\partial x_k} = \frac{\partial \Phi(x_i\beta)}{\partial x_k} = \phi(x_i\beta)\beta_k \tag{38}$$

where ϕ is the pdf of the standard normal cdf.[4] Thus, the marginal effect of increasing x_k results in a change in y of magnitude $\phi(x_i\beta)\beta_k$.

[4] Recall that the derivative of a cdf is the pdf.

In the case of logit we have:

$$\frac{\partial y_i}{\partial x_k} = \frac{\partial \Lambda(x_i\beta)}{\partial x_k} = \lambda(x_i\beta)\beta_k \tag{39}$$

where λ is the pdf of the standard logistic cdf. It is sometimes useful to rewrite this marginal effect as:

$$\begin{aligned}
\frac{\partial y_i}{\partial x_k} = \frac{\partial \Lambda(x_i\beta)}{\partial x_k} &= \lambda(x_i\beta)\beta_k \\
&= \frac{e^{x_i\beta}}{[1+e^{x_i\beta}]^2}\beta_k \\
&= \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \cdot \frac{1}{1+e^{x_i\beta}}\beta_k \\
&= \Lambda(x_i\beta)\left(\frac{1+e^{x_i\beta}-e^{x_i\beta}}{1+e^{x_i\beta}}\right)\beta_k \\
&= \Lambda(x_i\beta)\left(1-\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)\beta_k \\
&= \Lambda(x_i\beta)(1-\Lambda(x_i\beta))\beta_k \\
&= \Pr(y=1)\cdot\Pr(y=0)\cdot\beta_k
\end{aligned} \tag{40}$$

This reformulation is useful since it tells us that the marginal effect of increasing x_k is just the probability of success (defined by $x_i\beta$) times the probability of failure times the coefficient on x_k.

4.2.3 Computing Marginal Effects

Say we want to know the marginal effect of the effective number of parties on the probability of pre-electoral coalition formation. There are at least three ways of calculating this marginal effect and confidence interval:

1. **By Hand (in STATA):**

```
probit pec enep threshold

preserve

drawnorm MG_b1-MG_b3, n(1000) means(e(b)) cov(e(V)) clear

save simulated_betas, replace

restore

merge using simulated_betas
```

```
summarize _merge

drop _merge

summarize

scalar  h_enep = 2

scalar h_threshold = 8

scalar h_constant = 1

generate x_betahat = MG_b1*h_enep + MG_b2*h_threshold + MG_b3*h_constant

generate enep_marg = normalden(x_betahat)*MG_b1

sum enep_marg;

centile enep_marg, centile(2.5 97.5)
```

2. **STATA command: prchange**

   You can also use the prchange command by typing

   ```
   probit pec enep threshold

   prchange enep, x(enep=2 threshold=8)
   ```

   One problem is that this command does not calculate a confidence interval around the marginal effect.

3. **STATA command: mfx**

   Alternatively, you can use STATA's canned command

   ```
   probit pec enep threshold

   mfx, at(2 8)
   ```

4.2.4 Presenting Marginal Effects

There are various ways of presenting marginal effects - tables and graphs etc. Whichever method you use, make sure to show the appropriate measure of uncertainty as well.

While it is relatively easy to get marginal effects, you should again ask yourself whether this is the relevant quantity of interest. In a linear model, the marginal effect of x on y is the same as the

effect of a one unit increase in x on y - this is due to the linearity of the model. However, this is not the case in a binary response model due to the non-linearity of the model [use figure to illustrate]. A marginal effect in a binary response model is the effect of a very, very small change in x on the probability of $y = 1$. I would argue that this is rarely what you want to know about. Instead, you are more likely to want to know how the probability of $y = 1$ changes when we increase x by one unit (or some number of units). To know this you have to computer first differences.

4.3 First Differences

A first difference is just the change in the probability that $y = 1$ associated with some unit change in x. Thus, you might calculate the probability that $y = 1$ when $x_k = 1$ and the other xs are set to some values, then calculate the probability that $y = 1$ when $x_k = 2$ and the other xs are set at the same values as before, then calculate the difference in these probabilities.

Say we want to know how the probability that a pre-electoral coalition forms changes when the effective number of parties increases from 2 to 4. There are at least two ways of computing a first difference like this.

1. **By Hand (in STATA):**

```
probit pec enep threshold

preserve

drawnorm MG_b1-MG_b3, n(1000) means(e(b)) cov(e(V)) clear

save simulated_betas, replace

restore

merge using simulated_betas

summarize _merge

drop _merge

summarize

scalar  h_enep = 2

scalar  h_enep1 = 4

scalar h_threshold = 8
```

```
scalar h_constant = 1

generate x_betahat = MG_b1*h_enep + MG_b2*h_threshold + MG_b3*h_constant

generate prob_hat = normal(x_betahat)

generate x_betahat1 = MG_b1*h_enep1 + MG_b2*h_threshold + MG_b3*h_constant

generate prob_hat1 = normal(x_betahat1)

generate diff_prob = prob_hat1-prob_hat

sum prob_hat prob_hat1 diff_prob

centile prob_hat prob_hat1 diff_prob, centile(2.5 97.5)
```

2. **Clarify**

   You can also use Clarify to compute first differences by typing

```
estsimp probit pec enep threshold

setx enep 2 threshold 8

simqi, fd(pr) changex(enep 2 4) level(95)
```

4.4 Odds Ratios and the Logit Model

Odds ratios are an alternative and informative way of interpreting results when we are employing a logit model. The 'odds' of $y = 1$ for a given observation with some values of x are:

$$\begin{aligned} \Omega(x) &= \frac{\Pr(y=1|x)}{\Pr(y=0|x)} \\ &= \frac{\Pr(y=1|x)}{1-\Pr(y=1|x)} \\ &= \frac{\frac{e^{x\beta}}{1+e^{x\beta}}}{1-\frac{e^{x\beta}}{1+e^{x\beta}}} \end{aligned} \tag{41}$$

If $\Pr(y = 1)$, then it is easy to see that the odds are 1 to 1. If the $\Pr(y = 1) = 0.75$, then the odds are 3 to 1.

Instead of the odds, Ω, we can think in terms of the log-odds, $\ln\Omega$. The log-odds is sometimes

known as the logit. The log-odds is given by:

$$\ln\Omega(x) = \ln\left[\frac{\frac{e^{x\beta}}{1+e^{x\beta}}}{1-\frac{e^{x\beta}}{1+e^{x\beta}}}\right] = x\beta \tag{42}$$

As you can see, the log-odds are linear in x in the logit model. This means that the effect of a change in x_k on the log-odds of y is:

$$\frac{\partial \ln\Omega}{\partial x_k} = \beta_k \tag{43}$$

In other words, a unit change in x_k leads to a β_k change in the log-odds or logit. However, most people don't think in terms of log-odds; rather they think in terms of odds. However, from Eq. (43), we can see that the effect of a one-unit change in x_k on the *odds* of $y = 1$ is:

$$\frac{\Omega(x_k + 1)}{\Omega(x_k)} = e^{\hat{\beta}_k} \tag{44}$$

As you can now see, we can interpret the exponentiated coefficients from a logit model as the change in the odds of $y = 1$ associated with a one-unit increase in x_k. If $e^{\beta_k} > 1$, then we say that the odds of $y = 1$ are e^{β_k} times larger. If $e^{\beta_k} < 1$, then we say that the odds of $y = 1$ are e^{β_k} times smaller. Instead of a one-unit increase in x_k, we can generalize to a δ-sized increase in x_k in the following way:

$$\frac{\Omega(x_k + \delta)}{\Omega(x_k)} = e^{\hat{\beta}_k \delta} \tag{45}$$

To get STATA to report exponentiated coefficients, you can type:

```
logit pec enep threshold, or
```

where 'or' stands for odds-ratio.

It is also easy to get a percentage change in the odds associated with a one-unit change in x_k. This is given by:

$$\begin{aligned} \text{Percentage Change} &= 100\frac{\Omega(x_k + \delta) - \Omega(x_k)}{\Omega(x_k)} \\ &= 100[e^{\hat{\beta}_k \delta} - 1] \end{aligned} \tag{46}$$

This quantity can be interpreted as the percentage change in the odds of $y = 1$ for a δ unit change in x_k.

To illustrate how odds-ratios work, consider the following examples from Zorn.

1. Suppose that our logit estimate of $\hat{\beta}_k = 2.3$. Then a one-unit change in x_k:

- corresponds to an increase in the log-odds of $y = 1$ of 2.3
- a change in the odds that $y = 1$ of $e^{2.3} = 9.974$ i.e. the odds are 9.974 times larger.
- a percentage change in the odds that $y = 1$ of $100[e^{2.3} - 1] = 897.4\%$ i.e. the odds are 897.4% greater.

2. Suppose that our logit estimate of $\hat{\beta}_k = -0.22$. Then an eleven-unit change in x_k:

   - corresponds to a $-0.22 \times 11 = -2.42$ decrease in the log-odds of $y = 1$
   - a change in the odds that $y = 1$ of $e^{-0.22 \times 11} = e^{-2.42} = 0.089$
   - a percentage change in the odds that $y = 1$ of $100[e^{-0.22 \times 11} - 1] = 100(0.089 - 1) = -91.1\%$ i.e. the odds are 91.1% smaller.

4.5 Binary Response Models and Interaction Effects

There has been some recent debate about binary response models and interaction effects. Here is my take on it. Let's start with the claim that some scholars have made that it is not necessary to include an interaction term in logit and probit models in order to take account of conditional hypotheses since the non-linearity of these models implicitly force the effect of all the independent variables to depend on each other anyway (Berry & Berry 1991). Consider an additive probit model i.e. no explicit interaction term.

$$\Pr(y = 1) = \Phi(\gamma_0 + \gamma_1 X + \gamma_2 Z) = \Phi(\cdot) \tag{47}$$

The marginal effect of X as we have seen is $\frac{\partial \Phi(\cdot)}{\partial X} = \phi(\cdot)\gamma_1$. It is easy to see that the marginal effect of X depends, or is conditional on, the value of the other independent variables even in this additive model. The reason for this conditionality is basically that probabilities are constrained to run between 0 and 1. The effect of increasing (or decreasing) $x_i\beta$ on the probability that $y = 1$ has to get smaller and smaller at some point otherwise we will go outside the 0-1 bounds. We can think of this as a "compression effect". This compression effect is substantively meaningful and can be interpreted as such.

However, note that this conditionality or compression effect occurs whether the analyst's hypothesis is conditional or not – it is just part and parcel of deciding to use a non-linear model such as probit; it is always there. If you have a specific hypothesis that the effect of some X on the probability that $y = 1$ depends on the value of some other variable Z (above and beyond compression effects that are always there), then it is necessary to include an explicit interaction term as in the model below (Nagler 1991).

$$\Pr(y = 1) = \Phi(\beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ) = \Phi(\cdot) \tag{48}$$

The marginal effect of X is now $\frac{\partial \Phi(\cdot)}{\partial X} = \phi(\cdot)(\beta_1 + \beta_3 Z)$. The key then is to decide whether you have a conditional hypothesis that goes above and beyond simply stating that there is a compression effect.[5] If you do, then you should include an explicit interaction term.

[5]It is my belief that in almost all cases where a scholar has a conditional hypothesis, they mean a conditional

Let's assume that we have an explicit interaction term. How can we interpret the results from such a model. Let's start by being clear on our terminology. Imagine that I have some conditional hypothesis whereby some variable Z modifies the effect of X on Y. One question we might ask is how the value of Z modifies the effect of X on Y i.e. $\frac{\partial^2 Y}{\partial X \partial Z}$. This is what we mean by an "interaction effect". In the OLS world, where our model is

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon \tag{49}$$

we have

$$\frac{\partial y}{\partial X} = \beta_1 + \beta_3 Z \tag{50}$$

and so the interaction effect is

$$\frac{\partial^2 Y}{\partial X \partial Z} = \beta_3 \tag{51}$$

In other words, the coefficient (and standard error) on the interaction term tells us the direction, magnitude, and significance of the "interaction effect".

It turns out that things are not so simple in the probit or logit world? Let's assume that we have a logit model. We have

$$P(y_i = 1) = \frac{1}{1 + e^{-x_i \beta}} = \Lambda(x_i \beta) = \Lambda \tag{52}$$

Thus, we have

$$\frac{\partial P(y_i = 1)}{\partial X} = [\Lambda(1 - \Lambda)][\beta_1 + \beta_3 Z] \tag{53}$$

and the interaction effect is

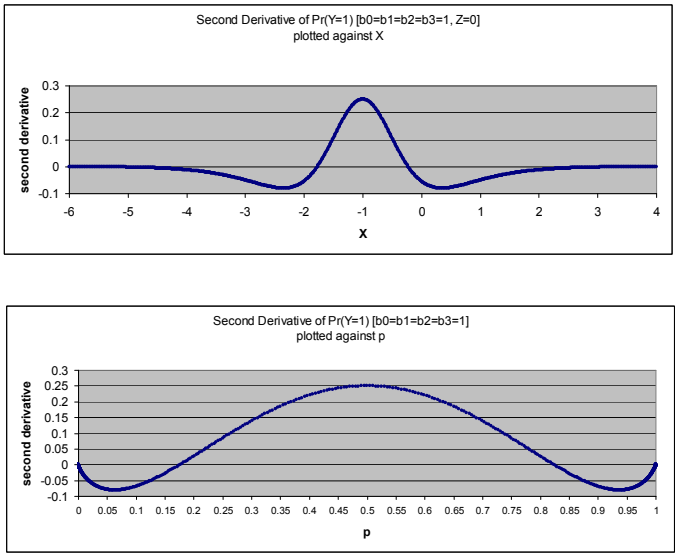
$$\frac{\partial^2 P(y_i = 1)}{\partial X \partial Z} = \beta_3 \Lambda(1 - \Lambda) + (\beta_1 + \beta_3 Z)(\beta_2 + \beta_3 X)\Lambda(1 - \Lambda)(1 - 2\Lambda) \tag{54}$$

The point to take away from Eq. (54) is that the coefficient (and standard error) on the interaction term does NOT tell us the direction, magnitude, or significance of the "interaction effect". Eq. (54) can be positive, negative, or zero (significant or not significant) irrespective of the sign and significance of the coefficient on the interaction term. In other words, you should not draw inferences about the interaction effect from the sign and significance of the coefficient on the interaction term i.e. β_3. You have to actually calculate the interaction effect for the values of the variables that you are interested in.

To illustrate the problem with just looking at the coefficient on the interaction term, look at the

relationship that goes beyond compression effects. This is why the general advice in the methods literature has been to always include an explicit interaction term.

Figure 1: A Simulation from a Logit Model with an Interaction Term



results from the following simulation. We started with a logit model where

$$x_i\beta = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ \tag{55}$$

and we set the values of the parameters to all be 1 i.e.

$$\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1 \tag{56}$$

In Figure 1, we have plotted how the "interaction effect" varies depending on the value of X. We have also plotted how the "interaction effect" varies with the probability that $y = 1$. The point to take away from here is that even though the coefficient on the interaction term β_3 is positive, the "interaction effect" can be negative (or 0) for certain values of X or for certain probabilities that $y = 1$.

5 Assessing Model Fit

Analysts often look for a one-number summary of model fit. In linear regression, we have the R^2 - I hope someone has explained why the R^2 is not particularly useful! In binary response models, we don't even have an R^2. STATA will report a 'pseudo-R^2' but this is even worse than an R^2. If anyone ever makes a big deal of their 'pseudo-R^2', ask them how it is calculated - most people will not know. Even if they did, ask them how they know whether it is big or small - there is no distribution and hence no way of knowing whether one R^2 is significantly different from another R^2.

There are a number of different ways to evaluate model fit in binary response models.

5.1 PCP, PRE, ePCP

The inherent issue is that your model has generated some predicted probabilities ($\hat{p}_i$) and you want to go from this predicted probability to a predicted outcome - either 1 or 0 ($\hat{y}_i$). Once we have the predicted outcome ($\hat{y}_i$), we can then compare this to the actual outcome (y_i). So, how do we do this?

5.1.1 PCP

Analysts often ask what is the percentage of observations that are correctly predicted (PCP or percent correctly predicted). The big issue is how we classify predicted outcomes based on predicted probabilities. The most common way to do this is to say that any observation with a predicted probability greater than or equal to 0.5 should be classified as a 1 and any observation with a predicted probability less than 0.5 should be classified as a 0. Thus, you have the following steps to calculate PCP

1. Estimate $\hat{\beta}$ using probit and for each observation i, calculate $\hat{p}_i$
2. For those observations with $\hat{p}_i \geq 0.5$, set $\hat{y}_i = 1$; otherwise set $\hat{y}_i = 0$.
3. Call each observation i with $y_i = \hat{y}_i$ a correct prediction. PCP is defined as the percentage of observations that are correctly predicted i.e.

$$\text{PCP} = \frac{100 \times (\text{Number of Correct Predictions})}{N} \tag{57}$$

You should note that maximizing the probit log-likelihood function will not necessarily produce a $\hat{\beta}$ that maximizes the PCP.

There are several problems with PCP.[6]

[6] One problem that few have noticed is that PCPs incorporate a notion that is opposed to the meaning of probabilities (Train 2007). PCPs are based on the idea that the decision maker is predicted by the researcher to choose the alternative for which the model gives the highest probability. However, the researcher obviously does not have enough information to predict the decision maker's choice; all he can do is predict the probability that the decision maker will choose each alternative. In stating choice probabilities, the researcher is only saying that if the choice situation were repeated numerous times, then each alternative would be chosen a certain number of times. This is very different from saying that the alternative with the highest probability will be chosen each time. Train (2007, 73) gives the following example. Suppose our model predicts choice probabilities of 0.75 and 0.25 in a two-alternative situation. These probabilities mean that if 100 people faced the same choice situation, the researcher's best prediction would be that 75 people choose one alternative and 25 people choose the other. However, the PCP statistic is based on the notion that the best prediction for each person is the alternative with the highest probability. This notion would predict that one alternative would be chosen by all 100 people. This procedure misses the point of probabilities and gives inaccurate market shares.

- Our estimate of β is measured with uncertainty, thus there should be an uncertainty measure associated with PCP.
- The process of calculating PCP treats an observation with $\hat{p}_i = 0.51$ the same as an observation with $\hat{p}_i = 0.99$ despite the fact that the former value of $\hat{p}_i$ says much less than the latter. Again, this classification procedures overstates the precision of PCP.

5.1.2 PRE

One alternative to PCP is known as the percentage reduction in error (PRE). PRE is based on a comparison of PCP and PMC, where PMC is the percentage of observations in the modal category of the observed data.

$$\text{PMC} = \% \text{ in Modal Category} \tag{58}$$

For example, if a probit data set has 100 observations and $y_i = 1$ for 60 of them, then $PMC = 0.6$.

The percentage reduction in error is calculated in the following way

$$\text{PRE} = \frac{\text{PCP} - \text{PMC}}{1 - \text{PMC}} \tag{59}$$

PRE seeks to compare the information provided by probit fitted categories with the classification errors a researcher would make if she naively assigned all fitted categories to the modal category. If PCP is less than PMC, then the PCP-based classification errors are greater than the classification errors a researcher would generate if she did not run a probit model and simply classified observations based only on the modal category.

Example: Say $PCP = 0.85$ and $PMC = 0.8$. Then

$$\text{PRE} = \frac{0.85 - 0.8}{0.2} = 0.25 \tag{60}$$

Since PRE is just a function of PCP, it still has the precision problems we outlined earlier with PCP.

5.1.3 ePCP

Herron (1999) has proposed an expected percent correctly predicted (ePCP). This statistic essentially provides the expected percentage of correct predictions and helps avoid the problem of treating an observation with $\hat{p}_i = 0.51$ the same as an observation with $\hat{p}_i = 0.99$. ePCP is calculated as

$$\text{ePCP} = \frac{1}{N}\left(\sum_{y_i=1} \hat{p}_i + \sum_{y_i=0} (1 - \hat{p}_i)\right) \tag{61}$$

As an example, suppose we have

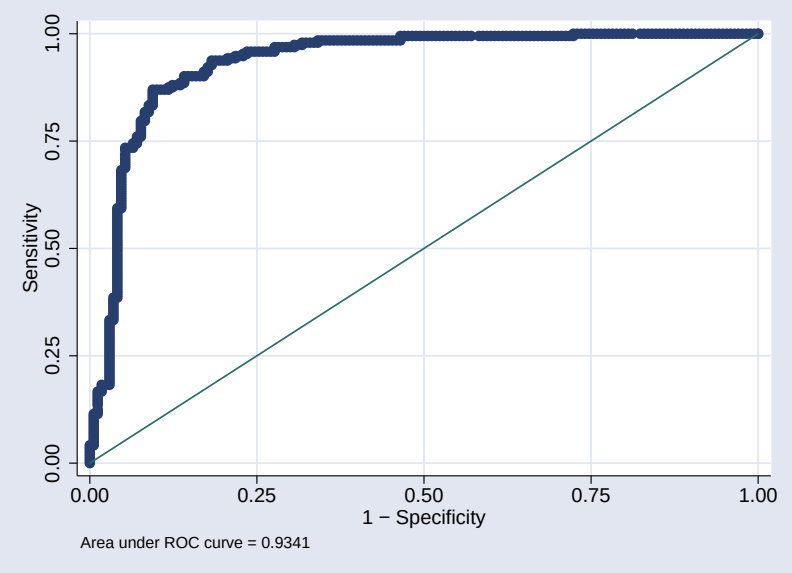
y_i	$\hat{p}_i$
0	0.6
1	0.6
1	0.8

Then ePCP=$\frac{1}{3}(0.4 + 0.6 + 0.8) = 0.6$. Given a particular data set, ePCP may be larger or smaller than PCP. To capture the uncertainty that arises from $\hat{\beta}$, we could use simulation methods to produce confidence intervals around ePCP. If you want to use a number to capture model fit, you should use ePCP instead of PCP. You may want to note ePCP, PCP, PRE and PMC.

5.2 ROC Curves

An alternative measure of model fit is the ROC curve (Receiver Operating Characteristic curve). One issue that remains with numbers such as PCP and ePCP is that they arbitrarily use the classification rule that $\hat{p}_i \geq 0.5 \rightarrow 1$ and $\hat{p}_i < 0.5 \rightarrow 0$. By using this classification rule, you are assuming that Type I and Type II errors are equally bad. Clearly, increasing the cut-point will reduce the chances of making one type of error at the expense of increasing the other type of error. A ROC curve basically tells you exactly how this tradeoff works for all possible cut-points. Consider the following ROC curve.

Figure 2: An Example of a ROC Curve



The y-axis captures Sensitivity which is the probability of correctly predicting a 1. The x-axis is 1-Specificity, where Specificity is the probability of correctly predicting a 0. The 45 degree line

indicates how a model with no covariates makes the tradeoff between Sensitivity and 1-Specificity (Sensitivity). The curved line (ROC curve) comes from the model with covariates. Any point on this line indicates how the probability of correctly predicting a 1 is traded off against the probability of correctly predicting a 0. For example, if Sensitivity=0.75 (probability of correctly predicting a 1 is 0.75), then Specificity=0.92 (probability of correctly predicting a 0 is 0.92). The Specificity number here comes from the fact that when Sensitivity=0.75, then 1-Specificity=0.08, and so Specificity=0.92. It is easy to see that the further the ROC curve is away from the 45 degree line the better the model predicts both 1s and 0s. A single statistic that conveys this information is the area under the ROC curve. When this area is 1, we are not making any tradeoff between predicting 1s and 0s and the model is correctly predicting everything. This statistic falls as the model becomes worse. The area under the ROC curve in this case is 0.9341 - and thus we might infer that the model fits quite well.

There are two commands in STATA to produce a ROC curve: LROC and ROCTAB. To use LROC, type:

```
probit Y X;

lroc, nograph;

lroc;
```

To use roctab, type:

```
probit Y X

predict p if e(sample), p;

roctab clintonvote p
```

The advantage of ROCTAB over LROC is that it gives you a measure of uncertainty around the ROC area.

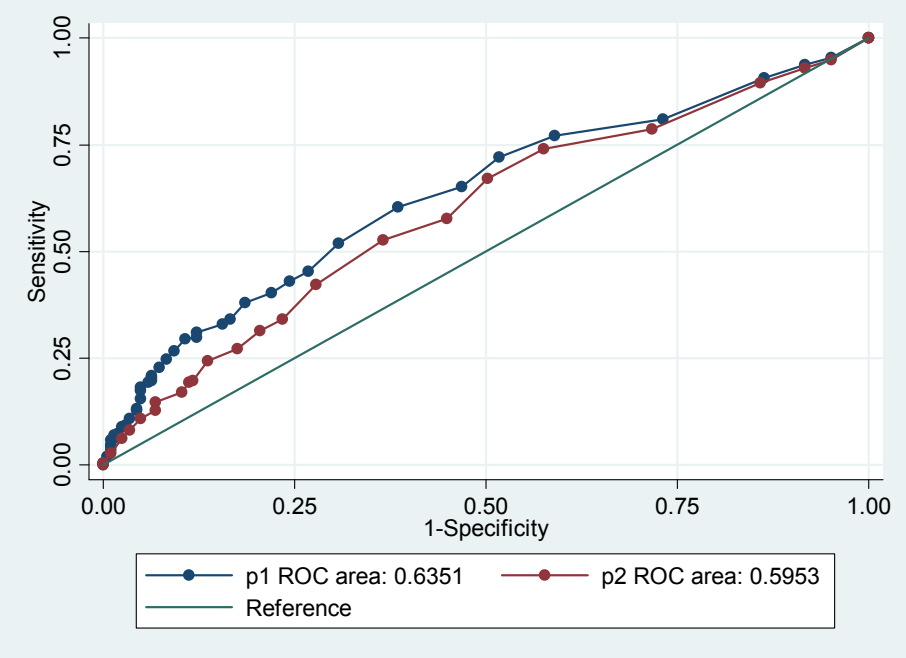
The ROC curve can also be used to test between models. For example, if the ROC curve from one model is always outside the ROC curve of the other model, then the first model is always better. To do this, you will use the ROCCOMP command. To compare two different models when you are using the *same set of observations*, type:

```
probit Y X1 X2 X3;
predict p1 if e(sample), p;

probit Y X1 X2;
predict p2 if e(sample), p;
```

```
roccomp clintonvote p1 p2;
roccomp clintonvote p1 p2, graph;
```

Figure 3: Comparing Two Models with the Same Set of Observations



	Obs	ROC Area	Std. Err.	-Asymptotic Normal-- [95% Conf. Interval]	
p1	463	0.6351	0.0256	0.58487	0.68535
p2	463	0.5953	0.0264	0.54365	0.64697

Ho: area(p1) = area(p2)
chi2(1) = 8.24 Prob>chi2 = 0.0041

Note that the previous set of commands require that you use the same set of observations. However, sometimes you find that you want to compare the same model on different sets of similar observations or compare two or more models estimated on subsets of the same data. You can do this for two subsamples, one where $Z = 1$ and one where $Z = 0$, by typing:

```
gen Y_dummy = 1 if Z==1;
replace Y_dummy=0 if Z==0;

probit Y X if clintonvote_dummy==1;
predict p1 if e(sample), p;
```

```
lroc, nograph;

probit Y X if clintonvote_dummy==0;
predict p2 if e(sample), p;
lroc, nograph;

gen newp=p1 if p1~=.;
replace newp=p2 if p2~=.;

roccomp Y newp, by(Y_dummy);
```

For more information about this, see Cleves (2002). To see an application of ROC curves to distinguish between models, see Beck, King & Zeng (2004).

5.3 Cross-Validation

Another way to evaluate the performance of your empirical model is through its ability to make out-of-sample predictions. This is called cross-validation. Essentially, what you do is iteratively drop one observation from your sample, run the model on the restricted sample, and then use the results of that model to predict the observation that was dropped. The code below performs cross-validation on a probit model, though, in principle, it can be adapted to deal with any cross-sectional model.

```
gen predicted=.;
gen count=_n;
for values i=1 (1) 100 {;
    probit Y X if _n~='i';
    predict fit;
    replaced predicted=fit if count=='i';
    drop fit;
    };
replace predicted=1 if predicted>=0.5;
replace predicted=0 if predicted<0.5;
gen correct=predicted-Y;
sum correct if correct==0;
sum correct if correct==1;
sum correct if correct==-1;
```

The loop should run from observation 1 to the highest numbered observation (in this case 100). The predicted value for observation i is stored in the variable PREDICTED. This is then rounded up to 1 to down to 0 as with PCPs. The number of correctly predicted observations is then calculated by subtracting the dependent variable from the predicted value. This is stored in the CORRECT

variable. If correctly predicted, this variable takes the value 0. A false positive is a 1 and a false negative is -1.

6 Separation Problems

One issue that often arises in binary response models is that of separation (Zorn 2005). Separation refers to a situation where one or more of your covariates perfectly predict the outcome of interest. Formally, separation implies the existence of a subvector $X_s \subseteq X$ by which all N observations can be correctly categorized as either $y_i = 0$ or $y_i = 1$. The simplest example is a 2×2 table of Y and X with an empty cell. Separation results in infinite coefficients and standard errors.

To see the problem that arises with separation, it is useful to think in terms of 'overlap'. Figure 4 shows actual and predicted values from a simulated bivariate logistic regression presented in Zorn (2005, 160). The different panels vary in terms of the degree to which the different values of X are

Figure 4: Actual and Predicted Values from Simulated Logistic Regression

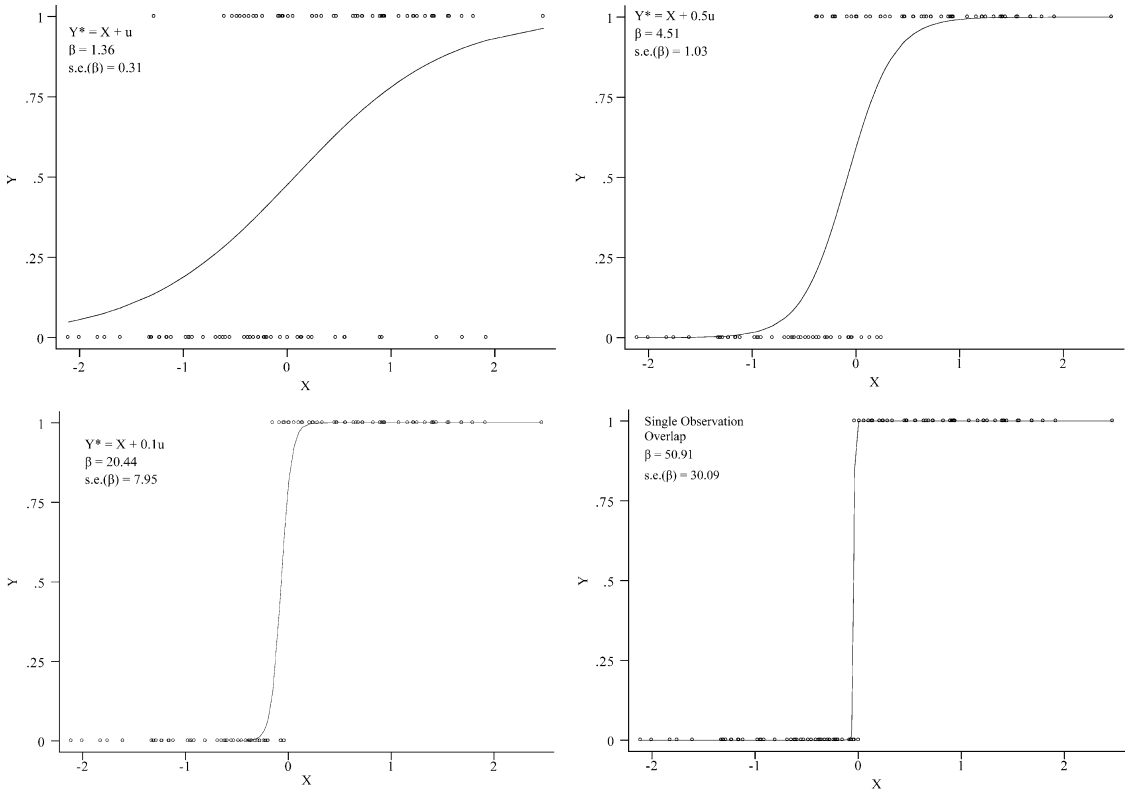


Fig. 1 Actual and predicted values, simulated logistic regressions.

associated with both 0s and 1s. For example, in the top left panel, there is considerable overlap in the sense that many of the values of X are associated with 0 and 1 on the outcome variable. As

you move from this panel to the bottom right panel, the amount of overlap decreases. By the time we are in the bottom right panel, there is only one value of X that has a 1 and a 0 associated with it; all the values of X below this point are associated with 0s only and all the values of X above this point are associated with 1s only. It is instructive to look at the estimated parameters and standard errors as we move across the panels. On the one hand, the coefficient estimates become larger as we are better able to correctly predict Y conditional on the value of X. On the other hand, the estimates of the standard errors also increase. The reason for this latter result is that the likelihood function is almost completely flat in the region of the parameter estimate.

The intuition is similar when we have a binary covariate, X_s. Situations in which X and Y do not overlap are equivalent to empty cells in the implied 2×2 table formed by the two variables. It turns out that we can distinguish between two different types of separation and this is easy to explain in the situation where we ave a binary covariate. Consider Figure 5. *Complete* separation is equivalent to the case in which only the two diagonal cells of the 2×2 table contain data. In this

Figure 5: Complete and Quasi-Complete Separation

Complete Separation

	Y=0	Y=1
X=0	xxxx	-----
X=1	-----	xxxx

Quasi-Complete Separation

	Y=0	Y=1
X=0	xxxx	-----
X=1	xxxx	xxxx

case, Y can be perfectly predicted by X_s for all observations in the data i.e., when $X_s = 0$, $Y = 0$; when $X_s = 1$, $Y = 1$, or vice versa. This results in no variance left to be explained in Y by the other covariates in the model and so the parameter estimates for the remaining covariates will be zero. You will also have infinite standard errors because the likelihood is flat. In contrast, *quasi-complete* separation occurs when only one cell of the 2×2 table contains data. In this situation,

the parameter estimate for the separating variable X_s (and its standard errors) will be infinite but the model's other covariates may remain relatively unaffected. Quasi-complete separation is more common than complete separation in practice. An important thing to note is that most statistical packages will not indicate whether there is separation, complete or quasi-complete.

What can be done about separation? The typical response is to just remove the offending variable or variables. This is obviously problematic because you are, in effect, removing a variable that you know is strongly associated with the dependent variable and, thereby, causing deliberate specification bias. STATA will automatically omit variables and drop observations when there is quasi-complete separation and fail to provide any estimate when there is complete separation. An alternative solution is to use penalized maximum likelihood (Heinze & Schemper 2002). Penalized maximum likelihood was originally developed to deal with small-sample bias issues in maximum likelihood (Firth 1993). Basically, we end up with the following generic penalized likelihood function:

$$\mathcal{L}(\theta|y)^* = \mathcal{L}(\theta|y)|I(\theta)|^{\frac{1}{2}} \tag{62}$$

where $\mathcal{L}(\theta|y)^*$ is our penalized likelihood function, $\mathcal{L}(\theta|y)$ is our standard likelihood function, and $I(\theta)$ is our standard information matrix. The corresponding log-likelihood is:

$$\ln\mathcal{L}(\theta|y)^* = \ln\mathcal{L}(\theta|y) + 0.5\ln|I(\theta)| \tag{63}$$

The corresponding gradient vector is:

$$G(\theta)^* = G(\theta) + 0.5tr\left\{I(\theta)^{-1}\left[\frac{\partial I(\theta)}{\partial\theta}\right]\right\} \tag{64}$$

To provide some intuition, Zorn notes that in the case of a binary logit model with a single dichotomous covariate, the penalized likelihood correction is equivalent to adding 0.5 to each cell of the implied 2×2 table. The bottom line is that the penalized-likelihood approach produces consistent parameter estimates in the presence of complete or quasi-complete separation.

There are various packages in R that can be used to estimate penalized-likelihood models: BRLR and LOGISTF. There is also now an ado file in STATA called FIRTHLOGIT. To obtain the STATA ado file, type: SSC INSTALL FIRTHLOGIT. To learn about how this command works, type: HELP FIRHLOGIT.

7 Scobit (Skewed Logit)

So far we have only examined probit and logit. However, all we really need to deal with dichotomous dependent variables is a probability distribution function that is data admissible (returns values between 0 and 1 etc.). Probit and logit do this. However, by using probit and logit we are automatically assuming that the maximal impact of any variable occurs when $\Pr(y = 1) = 0.5$

or $x_i\beta = 0$ (and that the probability distribution is symmetric).[7] These characteristics may not describe the world we are trying to model. As a result, an alternative has been proposed that allows the point of maximal impact to be determined by the data. This is Scobit and it is based on the Burr-10 distribution (Nagler 1994). As a comparison, we can see that in Probit:

$$\Pr(y_i = 1) = \pi_i = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} = \Phi(x_i\beta) \tag{65}$$

in Logit:

$$\Pr(y_i = 1) = \pi_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} = \frac{1}{1 + e^{-x_i\beta}} = \Lambda(x_i\beta) \tag{66}$$

and in Scobit:

$$\Pr(y_i = 1) = \pi_i = (1 + e^{-x_i\beta})^{-\alpha} \tag{67}$$

It should be obvious that logit is nested in scobit since when $\alpha = 1$ we have the logit model. Thus, when $\alpha = 1$ the maximal impact is when $\Pr(y = 1) = 0.5$. When $\alpha < 1$, the maximal impact is when $\Pr(y = 1) < 0.5$ i.e. the maximal impact is on respondents with low initial probabilities of choosing alternative 1. When $\alpha > 1$, the maximal impact is when $\Pr(y = 1) > 0.5$ i.e. among respondents likely to choose alternative 1. The fact that logit is nested in scobit makes it easy to test whether logit is preferable to scobit - we could just do an LR test or we could do a Wald test that $\alpha = 1$.

The following comments are from Neal Beck, "In practice, I find that the standard error on α is large and with reasonable sized samples it is hard to find α significantly different from 1. But even if Scobit is not helpful, it does sensitize you to the assumption of logit/probit that the maximal impact is for people with $\Pr(y = 1) = 0.5$. This is an assumption, not an empirical conclusion!"

8 Heteroskedastic Probit

Let's briefly look at the setup of the probit model again and see what happens if we have heteroskedasticity.[8] We'll use the latent variable setup. Suppose we have a latent or unobserved continuous variable, y_i^*.

$$y_i^* = x_i\beta + \epsilon_i \tag{68}$$

An observed realization of y_i^*, y_i is related to y_i^* in the following way:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > t \\ 0 & \text{if } y_i^* \leq t \end{cases}$$

[7] When $x_i\beta = 0$, $\Phi(x_i\beta) = \Pr(y = 1) = 0.5$.

[8] This section draws, among others, on Keele & Park (2006) and notes by Zorn.

where t is some threshold. As before, we'll assume that $t = 0$. From this setup, we have:

$$\begin{aligned} \Pr(y_i = 1) &= \Pr(y_i^* > 0) \\ &= \Pr(x_i\beta + \epsilon_i > 0) \\ &= \Pr(\epsilon_i > -x_i\beta) \\ &= \Pr(\epsilon_i \leq x_i\beta) \end{aligned} \tag{69}$$

If we assume that ϵ_i follows some distribution, then we can integrate over the distribution to estimate the probability that ϵ_i is less than or equal to $x_i\beta$. For this example, let's assume that the errors of y_i^* are normally distributed and so we estimate a probit model:

$$\Pr(y_i = 1) = \Phi(x_i\beta) \tag{70}$$

where Φ is the cumulative normal distribution.

As you will recall, we need to make an assumption about the error variance of the unobserved y_i^*. In effect, we assumed that the error term was homoskedastic or constant across all observations. This assumption was incorporated into the probit model by dividing both sides of the inequality in Eq. (69) by σ, the standard deviation of ϵ_i:

$$\Pr(y_i = 1) = \Pr\left(\frac{\epsilon_i}{\sigma} > \frac{-x_i\beta}{\sigma}\right) \tag{71}$$

This eventually gives us:

$$\Pr(y_i = 1) = \Phi\left(\frac{x_i\beta}{\sigma}\right) \tag{72}$$

As you will remember, we assume that σ is a constant that equals 1, so that we have:

$$\Pr(y_i = 1) = \Phi(x_i\beta) \tag{73}$$

This assumption is required to identify the model. We can now estimate $\hat{\beta}$. This the standard probit model and it assumes constant variance or homoskedasticity.

But what happens if there is heteroskedasticity? What if σ is not a constant? In the context of heterogeneous choice, σ is known or expected to vary systematically, such that $\sigma = \sigma_i$ with $i = 1 \ldots N$. For example, consider attitudes towards abortion. One might expect σ_i to be great for respondents who experience value conflict or who are politically ignorant. In contrast, one might expect σ_i to be small for respondents who are political sophisticates or who are strongly committed to one side of the abortion debate. It should be clear, though, that if the errors are nonconstant (heteroskedastic), then $\hat{\beta} = \frac{\hat{\beta}}{\sigma}$ and the parameter estimates will be biased, inconsistent, and inefficient; the standard errors will also be wrong.

8.1 A More Detailed Example to Illustrate the Problem

Zorn gives the following example to nicely illustrate what happens if we have heteroskedasticity. Suppose we are modeling some choice between individuals but that we can think of these individuals as falling into two groups: professors and graduate students. We might have the following two models:

Professors (p):

$$y_{ip}^* = x_{ip}\beta_p + \epsilon_{ip}$$

$$y_{ip} = \begin{cases} 1 & \text{if } y_{ip}^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

Graduate students (g):

$$y_{ig}^* = x_{ig}\beta_g + \epsilon_{ig}$$

$$y_{ig} = \begin{cases} 1 & \text{if } y_{ig}^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

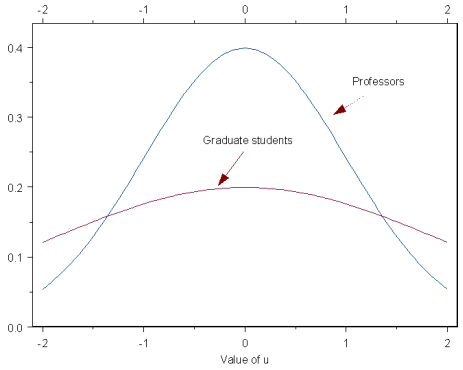
Let's assume that both groups give the same 'weight' to the independent variables when making their choice. In other words, the coefficients are the same in both models, i.e. $\beta_p = \beta_g$. However, we will assume that professors know exactly what they want and so the variance in their error terms is small. In contrast, graduate students are unsure and do not make decisions in a systematic way. As a result, the variance in their error terms is large. Thus, we have:

$$\begin{aligned} \epsilon_{ip} &\sim N(0, \sigma_p^2) \\ \epsilon_{ig} &\sim N(0, \sigma_g^2) \\ \sigma_g^2 &> \sigma_p^2 \end{aligned} \tag{74}$$

The different latent error distributions for professors and graduate students is shown graphically in Figure 6.

What happens if we pool the data from these two groups of individuals? You might think that, because the coefficients are the same, we would get good estimates of them. However, it turns out

Figure 6: Latent Error Distributions for Professors and Graduate Students



that we don't. In a sense, we would estimate different coefficients for each group.[9]

$$\begin{aligned} \hat{\beta}'_p &= \frac{\beta_p}{\sigma_p} \\ \hat{\beta}'_g &= \frac{\beta_g}{\sigma_g} \end{aligned} \tag{76}$$

In effect, pooling the two groups would lead to an estimate $\hat{\beta}'$ that is some sort of average of $\hat{\beta}'_p$ and $\hat{\beta}'_g$. The more graduate students in the sample, the closer the estimates of $\hat{\beta}'$ will be to $\hat{\beta}'_g$. And the more professors in the sample, the closer the estimates of $\hat{\beta}'$ will be to $\hat{\beta}'_p$. The bottom line is that the parameter estimates will be biased, inconsistent, and inefficient; the standard errors will also be wrong. Thus, heteroskedasticity is a big problem.

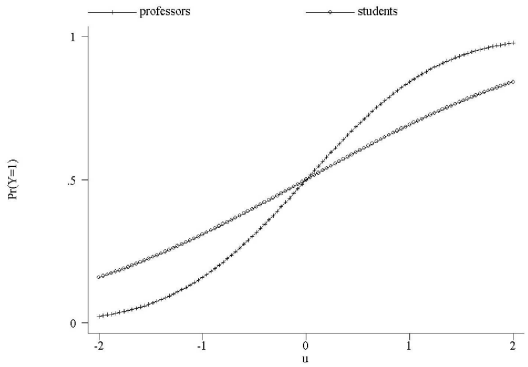
What's the intuition behind what is happening here? Think about it this way. If some subset of observations has a greater variance in their error term, then this essentially means that they have a flatter S-curve for their CDF (the scale is larger). For example, Figure 7 indicates that the CDF for the graduate students is flatter (less steep) than for the professors. The result is that the coefficients for the students will be different (larger) from those for the professors. This is because the underlying scale is larger (flatter S-curve) for the students. The larger underlying scale means that the coefficients for the students will have to be larger in order to get the same marginal impact on the (observed) probability of $y_i = 1$. Another (related) thing that the larger variance in the error term does is that it moves our predictions of the probability of $y_i = 1$ towards 0.5, which means that the estimated coefficients move towards 0.

[9] Moreover, the two sets of coefficients would be related in the following way:

$$\frac{\hat{\beta}'_p}{\hat{\beta}'_g} = \frac{\frac{\beta_p}{\sigma_p}}{\frac{\beta_g}{\sigma_g}} = \frac{\sigma_g}{\sigma_p} \tag{75}$$

If this ratio were 1, then we'd have the standard probit model. However, this ratio will obviously not be 1 if the standard deviations of the errors differ.

Figure 7: CDFs for Professors and Graduate Students



The bottom line is that if the underlying latent variables, y^*s have the same 'scale', then we can assume whatever we want for σ^2 withou any problems. But if some subset of observations has a greater variance in their error term (the scale is larger), then the coefficients for that group will be different from the rest. Pooling observations will lead to an estimated $\hat{\beta}'$ that is some sort of average of $\hat{\beta}'_p$ and $\hat{\beta}'_g$. The more graduate students in the sample, the closer the estimates of $\hat{\beta}'$ will be to $\hat{\beta}'_g$ and vice versa. Coefficients are biased and inconsistent, the standard errors will be wrong, and larger sample size will not help.

8.2 Solutions

One approach would be to treat the heteroskedasticity as a nuisance factor and use robust standard errors the same way we did with OLS and heteroscedasticity. However, an alternative approach would be to estimate a heteroskedastic probit (or logit). We might want to do this if we know what causes the heteroskedasticity and are explicitly interested in knowing how an independent variable affects the variance in the probability of some choice.

Consider the example from Zorn again. In that setup, we know that the variance differs across professors and graduate students. As a result, we would like to allow the variance for the two groups to be different. If we did this, then we could estimate the parameters consistently because we would be 're-scaling' the coefficients differently depending on whether we were considering professors or graduate students. In practice, the way we go about doing this is by allowing the variance of the unobserved variable to vary according to some function of one or more independent variables, z. Given that variance is always positive, we need to make sure that this function is always positive as well. One way to do this is to adopt the following multiplicative functional form for the variance of ϵ_i:

$$\text{var}(\epsilon_i) \equiv \sigma_i^2 = \exp(z_i\gamma)^2 \tag{77}$$

where z_i is a vector of independent variables of the i^{th} observation that define groups with different

error variances in the underlying latent variable, y_i^*. By taking the positive square root of both sides, we have a model for the standard deviation of the error distribution:

$$s.d.(\epsilon_i) \equiv \sigma_i = \exp(z_i\gamma) \tag{78}$$

In this setup, then we assume that the latent errors are distributed $N(0, [\exp(z_i\gamma)]^2)$. The probability for a particular observation is now:

$$\Pr(y_i = 1) = \Phi\left(\frac{x_i\beta}{\exp(z_i\gamma)}\right) \tag{79}$$

To see how this compares to the standard probit model, look back to Eq. (72). From this probability, we can derive the heteroskedastic probit log-likelihood:

$$\ln\mathcal{L} = \sum_{i=1}^{N}\left\{y_i \ln\Phi\left(\frac{x_i\beta}{\exp(z_i\gamma)}\right) + (1-y_i)\ln\left[1-\Phi\left(\frac{x_i\beta}{\exp(z_i\gamma)}\right)\right]\right\} \tag{80}$$

We would then maximize the equation with respect to β and γ to get our parameter estimates.[10] We would get our standard errors through the same procedure as before – find the negative of the inverse of the information matrix.

To estimate the model in STATA, you would use the HETPROB command:

```
hetprob Y X, het(Z)
```

What can we see from the model setup? The first thing to notice is that if all the elements of $\gamma = 0$, then $\exp(0) = 1$ and the model is just our standard probit model. Thus, the probit model is nested in the heteroskedastic probit model. As a result, you can use a likelihood ratio test to determine whether you need to run the heteroskedastic version or not.

What happens if a variable in z has a positive coefficient? As this variable increases, the variance of y_i^* also increases. This leads to a flatter S-shaped curve, which in turn means that the 'impact' of the independent variables on $\Pr(y_i = 1)$ decreases. In other words, as the variability of the underlying variable gets larger, it takes a bigger change in the independent variables to push the binary variable across the 'threshold'. We can see this more clearly by looking at marginal effects. Recall that the marginal effect in a standard probit model with respect to some x_k is:

$$\frac{\partial \Pr(y_i = 1)}{\partial x_k} = \phi(x_i\hat{\beta})\hat{\beta}_k \tag{81}$$

where the xs are set to some predetermined values. In contrast, the marginal effects in the het-

[10] You should note that there is no constant in the variance model; the model is not identified if a constant is included.

eroskedastic probit model with respect to some x_k is:

$$\frac{\partial \Pr(y_i = 1)}{\partial x_k} = \phi\left(\frac{x_i\hat{\beta}}{\exp(z_i\gamma)}\right)\left(\frac{\hat{\beta}_k}{\exp(z_i\gamma)}\right) \tag{82}$$

where we have to select values for the zs as well as the xs.[11] Again, it is easy to see that the marginal effect of some x_k declines if a variable in z has a positive coefficient.

8.3 Interpretation

When it comes to interpreting results from a heteroskedastic probit model, the best thing to do is to calculate predicted probabilities and changes in predicted probabilities.[12] The important thing to remember is that the predicted probability that a particular observation equals one is:

$$\Pr(y_i = 1) = \Phi\left(\frac{x_i\beta}{\exp(z_i\gamma)}\right) \tag{83}$$

You can use this equation to calculate predicted probabilities for different values of x and z.

8.4 Estimation Issues

As Zorn notes, the heteroskedastic probit model can be difficult to maximize. This is particularly the case if there is a lot of collinearity among the variables or if you are using similar variables in both parts of the model. Even if you achieve convergence, you may reach a local, as opposed to a global, maximum. Part of the problem is, as Keele and Park (2006) identify, that the heteroskedastic probit model suffers from fragile identification. What can you do about this? Things to look out for that might indicate a local maximum are (i) a smaller log-likelihood for the heteroskedastic probit than for a standard probit, (ii) huge standard errors, or (iii) strange coefficients. If these types of things occur, Keele and Park suggest trying different starting values, using different maximization algorithms, or trying to get more information about the heterogeneity.

8.5 Heteroskedastic Logit

Instead of the heteroskedastic probit, you could estimate a heteroskedastic logit (although STATA does not have a canned routine for this). The heteroskedastic logit is:

$$\Pr(y_i = 1) = \frac{1}{1 - e^{x_i\beta\theta_i}} \tag{84}$$

[11] Note that the marginal effect is much more complicated if you are looking at an independent variable in z or an independent variable that is in both x and z. You should be careful about this if you are calculating marginal effects.

[12] Don't simply look at z-tests etc. because they can be misleading.

This is exactly the same as a standard logit model except for the new parameter θ_i. The effect of θ_i is to spread out the logistic curve for some observations. For example, people with higher knowledge may be more able to discern their self interest so that their probability of voting for something is high if in their interest but low if not; for those with low knowledge, it takes a bigger movement in the independent variable to induce the same change in probabilities (Gerber & Lupia 1993). Since we cannot estimate a separate θ for each observation, we need to parameterize it as we did in the heteroskedastic probit model. For more information, see Dubin and Zeng (1991).

8.6 Applications

Alvarez & Brehm (1995) used a heteroskedastic probit model to examine ambivalence towards abortion policy. Their argument was that where people have conflicting core beliefs on certain issues, then the variance of their responses to survey items on those issues ought to be greater than those people without such conflicts. Alvarez and Brehm test this argument by looking to see whether respondents' ability to offer multiple and competing arguments for and against abortion increases the variation in abortion preferences. They found that respondents who were able to offer conflicting considerations on abortion displayed increased variance in their choices.

Zorn offers a number of other potential applications. Basically, any time you expect differences in the variance of the observations' values on the dependent variable, you should consider using a heteroskedastic probit model. When might this be? Well, this might occur if you are dealing with socialization effects. If people learn a task such as voting or become socialized, their variance on items relating to that socialization will probably decrease with time or age. In another area, do dictatorships have greater variance in their propensity to engage in conflict or not? Another area might be diffusion studies, where states might have a different variance in their propensity to adopt some policy innovation. Once you start to think about possible applications, a lot come to mind.

9 Rare Events Logit

Sometimes we study stuff that does not happen very often.[13] In effect, we have what are called rare events data – binary dependent variables with dozens to thousands of times fewer ones (events such as wars, coups, pre-electoral coalition dyads etc.) than zeros (non-events).

9.1 Problems

Political scientists have not been great at explaining or predicting rare events. King and Zeng (2001*a*, 2001*b*) argue that there are at least two reasons for this:

1. Data collection issues

[13] Notes are based on Zorn.

2. Prediction bias

Let's look at these problems in more detail.

9.1.1 Data Collection Issues

Given fixed resources, a tradeoff always exists between collecting more observations and including better or additional independent variables. In rare events data, the fear of collecting data sets with no events or few events has led scholars to collect very large numbers of observations with few, and in most cases poorly measured, independent variables. In effect, the decision to spend time getting large numbers of observations so that we can have some 1s means less time spent trying to get good independent variables. Thus, we have a very inefficient data collection strategy.

9.1.2 Prediction Bias

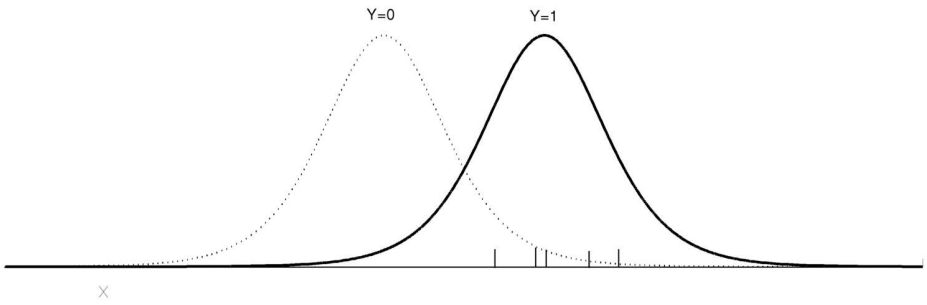
Regular models such as logit (and probit) do not do a good job at predicting rare events. As we have already seen, logit coefficients are biased in small samples. However, as King and Zeng (2001*a*, 2001*b*) note, logit coefficients can lead us to underestimate the probability of an event even with sample sizes in the thousands when we have rare events data. The bottom line is that logit (i) underestimates the probability of an event and (ii) does so in an increasingly dysfunctional way as the event gets rarer. Why is this the case?

The basic idea is that the $\Pr(y_i = 1)$ will be systematically underestimated in samples in which the $y_i = 1$ outcomes are relatively rare. The reason has to do with classification errors – our ability to accurately gauge a 'cutting point' for distinguishing $y_i = 1|x$ from $y_i = 0|x$ is biased in the direction of favoring zeros at the expense of ones when we have rare events data. In effect, we have a better idea about the distribution of 0s than 1s and so do a better job of classifying 0s than 1s. The 'classification' bias affects the constant term $\hat{\beta}_0$ in the model and through this, the predictions that we ultimately calculate.

To see how rare events bias logit coefficients graphically, see Figure 8. A large number of 0s allows us to estimate the dotted density essentially without error. But the density for $Y = 1$ will be poorly estimated because it's based on just 5 observations. Now consider finding a cutting point that maximally distinguishes between 0s and 1s by making the fewest mistakes. This cut point is related to the MLE of β and would probably be placed just to the left of the vertical line furthest to the left. With many more 0s than 1s, the area in the right tail of $\Pr(X|Y = 0)$ will be well estimated but the area in the left tail of $\Pr(X|Y = 1)$ will not be. Since the cutting point is a function of these tails, it will be biased in the direction of favoring 0s rather than 1s.

Think about this a little more formally with the help of a simple example from King and Zeng. Let π_i be $Pr(y_i = 1)$ and assume that we have a model with just one independent variable with coefficient $\beta_1 = 1$. From this we can see that $Pr(y_i = 1) = \frac{\exp(\beta_0 + x_i)}{1+\exp(\beta_0 + x_i)}$. In this model, King and

Figure 8: How Rare Events Bias Logit Coefficients



NOTES: Observations are arrayed horizontally according to the value of X, where $\beta_1 > 0$. The few $Y = 1$ observations appear as short vertical lines, along with the (solid) line for the density from which they were drawn. The many $Y = 0$ observations do not appear but their density appears as a dotted line. Taken from King and Zeng (2001*a*, 146).

Zeng show that the bias is about:

$$E[\hat{\beta}_0 - \beta_0] \approx \frac{\overline{\pi} - 0.5}{N\overline{\pi}(1-\overline{\pi})} \tag{85}$$

where $\overline{\pi}$ is the mean of π_i. When events are rare, so that $\overline{\pi}_i < 0.5$, this means that:

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

As we will see, King and Zeng point out that the standard way of computing predicted probabilities is suboptimal in finite samples and leads to us underestimate the probability of an event even more.

9.2 An Alternative Approach

Given these problems with rare events data, King and Zeng outline an alternative approach to using a standard logit. King and Zeng argue that we should collect data on all the possible 1s in the data and a random sample of 0s. This is called *choice-based sampling* in econometrics or *case control sampling* in statistics and biostatistics.[14] The basic idea is the following:

[14]Note that we are sampling on the dependent variable here, something that we are often told not to do. However, it turns out that selecting on the dependent variable via case control sampling can be a useful strategy so long as we make some corrections to our estimates.

1. Figure out the proportion τ of the population with 1s.
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 analysis on the data.[15]
5. Correct the coefficients and standard errors after the fact.

9.2.1 Sampling

The first thing to do is to get our sample. To keep things simple, assume that we know the proportion of 1s in the population; call this proportion τ.[16] 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 mean proportion $\overline{y}$. As an example, we might observe wars in only 0.1% of the cases ($\tau = 0.0001$) but actually select a sample with all, say, 1,000 instances of conflict, as well as 2,000 randomly selected non-wars. Thus, we would have $N = 3,000$ and $\overline{y} = 0.333$.

King and Zeng make suggestions for how many 0s to include relative to the number of 1s. As they note, the marginal contribution to the explanatory variables' information content for each additional 0 starts to drop as the number of 0s passes the number of 1s. Their general suggestion is that you will not want to collect more than 2-5 times as many 0s as 1s; a ratio of 1:1 ($\overline{y} = 0.5$) could even work. In fact, they state that a useful practice is to start with a ratio of 1:1. Then if the standard errors and confidence intervals are narrow enough, stop. Otherwise, continue to sample zeros randomly and stop when the confidence intervals get sufficiently small for the substantive purposes at hand.

9.2.2 Correcting Estimates for Selection on Y

Designs that select on the dependent variable can be consistent and efficient, but only with the appropriate statistical corrections. There are at least two approaches to correcting the estimates:(i) weighting and (ii) prior correction.

Weighting

One approach to correcting for case control sampling involves weighting the observations in a particular way (Greene 2003, 673). Intuitively, the goal is to increase the weight of 0s and reduce the weight of the 1s in proportion to their frequency in the population.

Formally, we can think of $w_1 = \frac{\tau}{\overline{y}}$ as the weights for the 1s and $w_0 = \frac{1-\tau}{1-\overline{y}}$ as the weight for the 0s. To the extent that $\tau < \overline{y}$, as it always will be for rare events data, this will have the desired effect.

[15] It has to be a logit as you will see.

[16] If you don't know τ, then you need to consider an alternative approach outlined in King and Zeng (2002).

The log-likelihood for the weighted logit is then:[17]

$$\ln\mathcal{L} = \sum_{i=1}^{N} w_1 y_i \ln\Lambda(x_i\beta) + w_0(1-y_i)\ln[1-\Lambda(x_i\beta)] \tag{86}$$

Weighting is a good approach if the sample is large and there is a chance that the model is misspecified (when is it not misspecified?). Despite this, King and Zeng argue that weighting, while intuitively reasonable, may not be optimal in some circumstances. They argue that:

- Weighting is less efficient that the prior correction strategy that we are about to get to, thereby making it less useful in small samples.
- Standard errors from the weighted regression are way off and, before King and Zeng, finite sample corrections were not an option.

Prior Correction

The second approach to correcting for case control sampling is 'prior correction'. With case control sampling, it can be shown that all the usual estimates for $\beta_1 \ldots \beta_k$ are consistent; the only problem comes with the intercept, which is biased. Thus, 'prior correction' involves correcting the standard MLE of the intercept $\hat{\beta}_0$. The following prior-corrected estimate is consistent for β_0:

$$\hat{\beta}_{0pc} = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\overline{y}}{1-\overline{y}}\right)\right] \tag{87}$$

9.2.3 Bias Correction

Even after correcting for case control sampling, King and Zeng note that we still have bias in our usual estimates of β in finite samples. This needs to be corrected. King and Zeng show that the bias in $\hat{\beta}$ can be estimated by the following weighted least-squares expression:

$$\text{bias}(\hat{\beta}) = (X'WX)^{-1}X'W\xi \tag{88}$$

where ξ is a complicated combination of the weights w_i, the predicted probabilities $\hat{\pi}_i$, and the X matrix. The bias-corrected estimate is then $\tilde{\beta} = \hat{\beta} - \text{bias}(\hat{\beta})$.

9.2.4 Probability Calculations

We know that biased-corrected estimate $\tilde{\beta}$ is less biased and has smaller variance, and hence a smaller mean square error, than the original estimate $\hat{\beta}$. This might lead you to think that we could simply plug in the biased-corrected estimate into our standard logit equation to get predicted

[17] This is called the weighted exogenous sampling maximum-likelihood estimator (Manski & Lerman 1977).

probabilities:

$$\Pr(\hat{y}_i = 1|\tilde{\beta}) = \frac{\exp(x_i\tilde{\beta})}{1 + \exp(x_i\tilde{\beta})} \tag{89}$$

However, this is not optimal because it ignores the uncertainty in $\tilde{\beta}$. This uncertainty can be thought of as sampling error or the fact that $\tilde{\beta}$ is estimated rather than known; it is reflected in the fact that the standard errors are greater than zero. In many cases, ignoring estimation uncertainty leaves the point estimate unaffected and changes only its standard error. However, because of the nature of $\Pr(\hat{y}_i = 1)$ as a quantity to be estimated, ignoring uncertainty affects the point estimate too.[18] Ignoring estimation uncertainty generates too small an estimated probability of a rare event.

So, how can we calculate the correct predicted probability? One way, which you have already seen, is to use simulation: take a random draw of β from $P(\beta)$ – the sampling distribution of β, insert it into $\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$, repeat, and average over the simulations. Increasing the number of simulations enables us to approximate $\Pr(y_i = 1)$ to any desired degree of accuracy. A second way is to use an analytical method for including the estimation uncertainty back into the predicted probabilities. The basic idea is to use the corrected estimates, $\tilde{\beta}$ to get predicted probabilities $\tilde{\pi}_i$, and then correct these predicted probabilities. As King and Zeng illustrate, we have:

$$\Pr(y_i = 1) \approx \tilde{\pi}_i + C_i \tag{90}$$

where the correction factor is:

$$C_i = (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)X_iV(\tilde{\beta})X_i' \tag{91}$$

There are a number of things worth noting about this correction factor. The first is that the uncertainty is captured in the $V(\tilde{\beta})$ term. This reflects the extent to which there is intrinsic uncertainty in $\tilde{\beta}$. Thus, the correction factor grows with uncertainty in $\tilde{\beta}$. In the presence of some uncertainty, the direction of the bias is determined by the first factor in C_i, $(0.5 - \tilde{\pi}_i)$. When $\tilde{\pi}_i < 0.5$, as is the case for rare events, the correction factor adds to the estimated probability of an event. Hence, using $\tilde{\pi}_i$ alone underestimates the predicted probability.

9.3 Estimation

As you can see, this is quite complicated. Fortunately, King and Zeng have made things easy by providing a STATA routine called RELOGIT to implement the corrections they note in their article. You will get to use the RELOGIT command in the homework.[19] The basic command is:

[18] Note that this point is not just about rare events data. Calculating predicted probabilities, predicted counts, predicted durations etc. all require us to plug in the estimated parameters. But simply plugging in the estimated parameters ignores that these parameters are estimated and not known. One solution is to use simulation to calculate the quantity of interest.

[19] You will need to get the RELOGIT.ADO file and auxiliary documents from Gary King's webpage. Once you have the files, copy the content. Then go to your STATA10 (STATA9) folder on the c: drive, go into the ADO folder, into the BASE folder, and into the R folder, and then paste the content. If you open up STATA and type HELP RELOGIT, you will find out about the ado file.

```
relogit Y X, wc()
```

or

```
relogit Y X, pc()
```

depending on whether you are going to use the weight correction (wc) or the prior correction (pc). The homework will walk you through one full example.

References

Alvarez, R. Michael & John Brehm. 1995. "American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values." *American Journal of Political Science* 39:1055–1082.

Beck, Nathaniel, Gary King & Langche Zeng. 2004. "Theory and Evidence in International Conflict: A Response to de Marchi, Gelpi and Grynaviski." *American Political Science Review* 98:379–389.

Berry, Frances Stokes & William Berry. 1991. "Specifying a Model of State Policy Innovation." *American Political Science Review* 85:573–579.

Cleves, Mario A. 2002. "From the Help Desk: Comparing Areas Under Receiver Operating Characteristics Curves from Two or More Probit or Logit Models." *The STATA Journal* 2:301–313.

Dubin, Jeffrey & Langche Zeng. 1991. "The Heterogenous Logit Model." Caltech Working Papers, #759.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80:27–38.

Gerber, Elizabeth & Arhtur Lupia. 1993. "When Do Campaigns Matter? Informed Votes, the Heteroskedastic Logit and the Responsiveness of Electoral Outcomes." Caltech Working Papers, #814.

Greene, William. 2003. *Econometric Analysis.* New Jersey: Prentice Hall.

Heinze, Georg & Michael Schemper. 2002. "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21:2409–2419.

Herron, Michael. 1999. "Postestimation Uncertainty in Limited Dependent Variable Models." *Political Analysis* 8:83–98.

Keele, Luke & David Park. 2006. "Ambivalent about Ambivalence: A Re-Examinatino of Heteroskedastic Probit Models." Ohio State University.

King, Gary & Langche Zeng. 2001*a*. "Explaining Rare Events in International Relations." *International Organization* 55:693–715.

King, Gary & Langche Zeng. 2001*b*. "Logistic Regression in Rare Events Data." *Political Analysis* 12:137–163.

King, Gary & Langche Zeng. 2002. "Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies." *Statistics in Medicine* 21:1409–1427.

Manski, Charles F. & Steven R. Lerman. 1977. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica* 45:1977–1988.

Nagler, Jonathan. 1991. "The Effect of Registration Laws and Education on U.S. Voter Turnout." *American Political Science Review* 85:1393–1405.

Nagler, Jonathan. 1994. "Scobit: An Alternative to Logit and Probit." *American Journal of Political Science* 38:230–255.

Train, Kenneth E. 2007. *Discrete Choice Models with Simulation.* New York: Cambridge University Press.

Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models." *Political Analysis* 13:157–170.