Review of Binary Choice

Chris Conlon

April 26, 2020

Applied Econometrics

Intro

Binary Choice: Overview

Many problems we are interested in look at discrete rather than continuous outcomes:

- Entering a Market/Opening a Store
- Working or a not
- Being married or not
- Exporting to another country or not
- Going to college or not
- Smoking or not
- etc.

Simplest Example: Flipping a Coin

Suppose we flip a coin which is yields heads ($Y = 1$) and tails ($Y = 0$). We want to estimate the probability p of heads:

$$Y_i = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p \end{cases}$$

We see some data $Y_1, \ldots, Y_N$ which are (i.i.d.)

We know that $Y_i \sim Bernoulli(p)$.

Simplest Example: Flipping a Coin

We can write the likelihood of N Bernoulli trials as

$$Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) = f(y_1, y_2, \ldots, y_N | p)$$

$$\begin{aligned} &= \prod_{i=1}^{N} p^{y_i} (1-p)^{1-y_i} \\ &= p^{\sum_{i=1}^{N} y_i} (1-p)^{N - \sum i=1^N y_i} \end{aligned}$$

And then take logs to get the log likelihood:

$$\ln f(y_1, y_2, \ldots, y_N | p) = \left(\sum_{i=1}^{N} y_i \right) \ln p + \left(N - \sum_{i=1}^{N} y_i \right) (1-p)$$

Simplest Example: Flipping a Coin

Differentiate the log-likelihood to find the maximum:

$$
\begin{aligned}
\ln f(y_1, y_2, \ldots, y_N | p) &= \left(\sum_{i=1}^{N} y_i\right) \ln p + \left(N - \sum_{i=1}^{N} y_i\right) \ln(1-p) \\
\rightarrow 0 &= \frac{1}{\hat{p}}\left(\sum_{i=1}^{N} y_i\right) + \frac{-1}{1-\hat{p}}\left(N - \sum_{i=1}^{N} y_i\right) \\
\frac{\hat{p}}{1-\hat{p}} &= \frac{\sum_{i=1}^{N} y_i}{N - \sum_{i=1}^{N} y_i} = \frac{\overline{Y}}{1-\overline{Y}} \\
\hat{p}^{MLE} &= \overline{Y}
\end{aligned}
$$

That was a lot of work to get the obvious answer: fraction of heads.

More Complicated Example: Adding Covariates

We probably are interested in more complicated cases where p is not the same for all observations but rather $p(X)$ depends on some covariates. Here is an example from the Boston HMDA Dataset:

- 2380 observations from 1990 in the greater Boston area.
- Data on: individual Characteristics, Property Characteristics, Loan Denial/Acceptance (1/0).
- Mortgage Application process circa 1990-1991:
  - Go to bank
  - Fill out an application (personal+financial info)
  - Meet with loan officer
  - Loan officer makes decision
    - Legally in race blind way (discrimination is illegal but rampant)
    - Wants to maximize profits (ie: loan to people who don't end up defeaulting!)

Loan Officer's Decision

Financial Variables:

- P/I ratio
- housing expense to income ratio
- loan-to-value ratio
- personal credit history (FICO score, etc.)
- Probably some nonlinearity:
  - Very high $LTV > 80\%$ or $> 95\%$ is a bad sign (strategic defaults?)
  - Credit Score Thresholds

Loan Officer's Decision

Goal $Pr(Deny = 1|black, X)$

- Lots of potential omitted variables which are correlated with race
  - Wealth, type of employment
  - family status
  - credit history
  - zip code of property
- Lots or redlining cases hinge on whether or not black applicants were treated in a discriminatory way.

TABLE 11.1 Variables Included in Regression Models of Mortgage Decisions

Variable	Definition	Sample Average
Financial Variables		
P/I ratio	Ratio of total monthly debt payments to total monthly income	0.331
housing expense-to-income ratio	Ratio of monthly housing expenses to total monthly income	0.255
loan-to-value ratio	Ratio of size of loan to assessed value of property	0.738
consumer credit score	1 if no "slow" payments or delinquencies 2 if one or two slow payments or delinquencies 3 if more than two slow payments 4 if insufficient credit history for determination 5 if delinquent credit history with payments 60 days overdue 6 if delinquent credit history with payments 90 days overdue	2.1
mortgage credit score	1 if no late mortgage payments 2 if no mortgage payment history 3 if one or two late mortgage payments 4 if more than two late mortgage payments	1.7
public bad credit record	1 if any public record of credit problems (bankruptcy, charge-offs, collection actions) 0 otherwise	0.074
Additional Applicant Characteristics		
denied mortgage insurance	1 if applicant applied for mortgage insurance and was denied, 0 otherwise	0.020
self-employed	1 if self-employed, 0 otherwise	0.116
single	1 if applicant reported being single, 0 otherwise	0.393
high school diploma	1 if applicant graduated from high school, 0 otherwise	0.984
unemployment rate	1989 Massachusetts unemployment rate in the applicant's industry	3.8
condominium	1 if unit is a condominium, 0 otherwise	0.288
black	1 if applicant is black, 0 if white	0.142
deny	1 if mortgage application denied, 0 otherwise	0.120

Linear Probability Model

First thing we might try is OLS

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- What does β_1 mean when Y is binary? Is $\beta_1 = \frac{\Delta Y}{\Delta X}$?
- What does the line $\beta_0 + \beta_1 X$ when Y is binary?
- What does the predicted value $\hat{Y}$ mean when Y is binary? Does $\hat{Y} = 0.26$ mean that someone gets approved or denied for a loan?

Linear Probability Model

OLS is called the linear probability model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

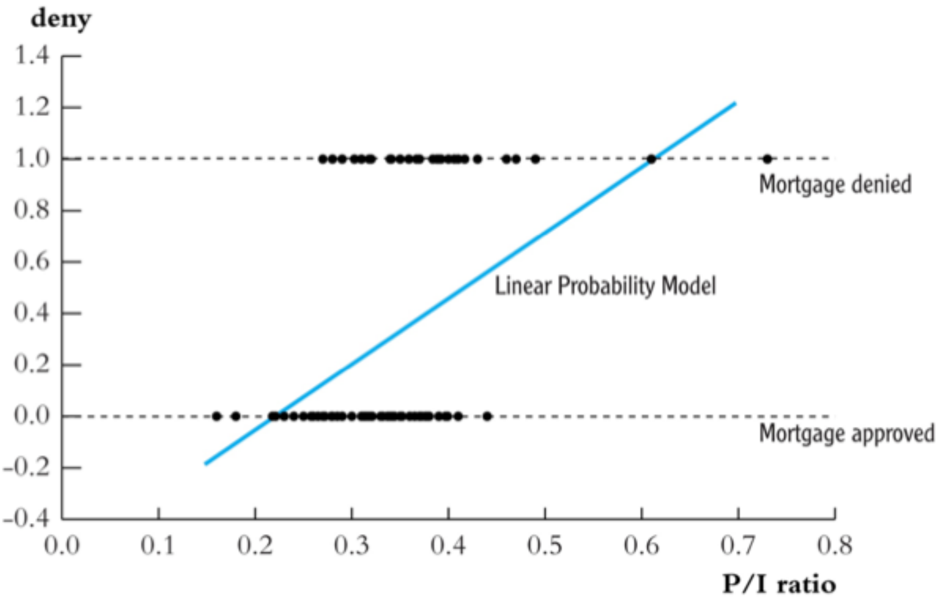
because:

$$\begin{aligned} E[Y|X] &= 1 \times Pr(Y = 1|X) + 0 \times Pr(Y = 0|X) \\ Pr(Y = 1|X) &= \beta_0 + \beta_1 X_i + \varepsilon_i \end{aligned}$$

The predicted value is a probability and

$$\beta_1 = \frac{Pr(Y = 1|X = x + \Delta x) - Pr(Y = 1|X = x)}{\Delta x}$$

So β_1 represents the average change in probability that $Y = 1$ for a unit change in X.


deny
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
P/I ratio
Mortgage denied
Linear Probability Model
Mortgage approved

That didn't look great

- Is the marginal effect β_1 actually constant or does it depend on X?
- Sometimes we predict $\hat{Y} > 1$ or $\hat{Y} < 0$. What does that even mean? Is it still a probability?
- Fit in the middle seems not so great – what does $\hat{Y} = 0.5$ mean?

Results

$$\widehat{deny_i} = \underset{(0.32)}{-.091} \quad \underset{(.098)}{+.559} \cdot \text{P/I ratio} + \quad \underset{(.025)}{.177} \cdot \text{black}$$

Marginal Effects:

- Increasing P/I from $0.3 \rightarrow 0.4$ increases probabilty of denial by 5.59 percentage points. (True at all level of P/I).
- At all P/I levels blacks are 17.7 percentage points more likely to be denied.
- But still some omitted factors.
- True effects are likely to be nonlinear can we add polynomials in P/I? Dummies for different levels?

Moving Away from LPM

Problem with the LPM/OLS is that it requires that marginal effects are constant or that probability can be written as linear function of parameters.

$$Pr(Y = 1|X) = \beta_0 + \beta_1 X + \epsilon$$

Some desirable properties:

- Can we restrict our predictions to $[0, 1]$?
- Can we preserve monotonicity so that $Pr(Y = 1|X)$ is increasing in X for $\beta_1 > 0$?
- Some other properties (continuity, etc.)

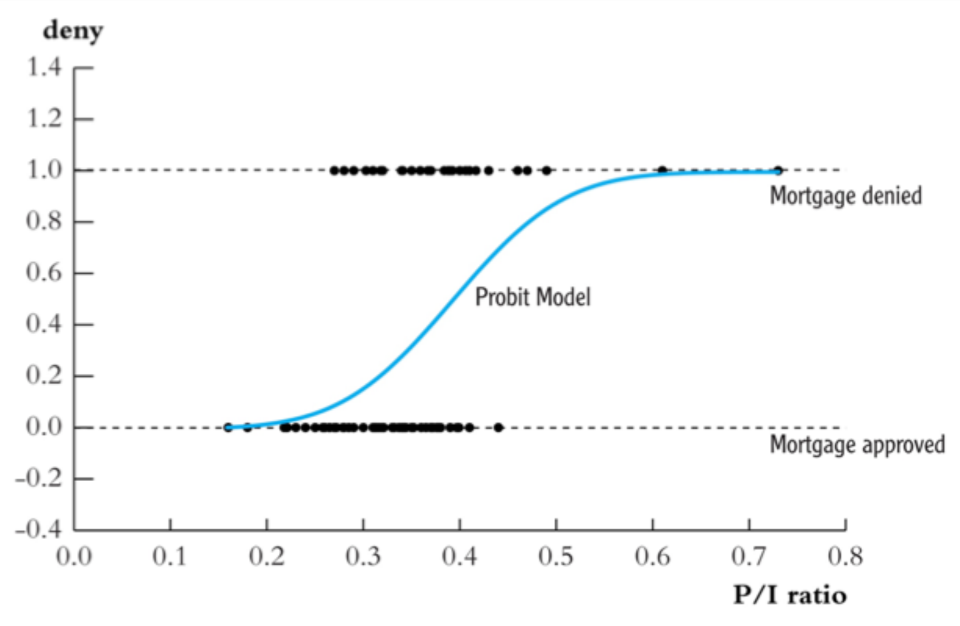
Moving Away from LPM

Problem with the LPM/OLS is that it requires that marginal effects are constant or that probability can be written as linear function of parameters.

$$Pr(Y = 1|X) = \beta_0 + \beta_1 X + \epsilon$$

Some desirable properties:

- Can we restrict our predictions to $[0, 1]$?
- Can we preserve monotonicity so that $Pr(Y = 1|X)$ is increasing in X for $\beta_1 > 0$?
- Some other properties (continuity, etc.)
- Want a function $F(z) : (-\infty, \infty) \rightarrow [0, 1]$.
- What function will work?


deny
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
P/I ratio
Mortgage denied
Probit Model
Mortgage approved

Choosing a transformation

$$Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

- One $F(\cdot)$ that works is $\Phi(z)$ the normal CDF. This is the probit model.
  - Actually any CDF would work but the normal is convenient.
- One $F(\cdot)$ that works is $\frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$ the logistic function . This is the logit model.
- Both of these give 'S'-shaped curves.
- The LPM is $F(\cdot)$ is the identity function (which doesn't satisfy my $[0, 1]$ property).
- This $F(\cdot)$ is often called a link function. Why?

Why use the normal CDF?

Has some nice properties:

- Gives us more of the 'S' shape
- $Pr(Y = 1|X)$ is increasing in X if $\beta_1 > 0$.
- $Pr(Y = 1|X) \in [0, 1]$ for all X
- Easy to use – you can look up or use computer for normal CDF.
- Relatively straightforward interpretation
  - $Z = \beta_0 + \beta_1 X$ is the z-value.
  - β_1 is the change in the z-value for a change in X_1.

TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: *deny* = 1 if mortgage application is denied, = 0 if accepted; 2380 observations.

Regression Model / **Regressor**	*LPM* (1)	*Logit* (2)	*Probit* (3)	*Probit* (4)	*Probit* (5)	*Probit* (6)
black	0.084** (0.023)	0.688** (0.182)	0.389** (0.098)	0.371** (0.099)	0.363** (0.100)	0.246 (0.448)
P/I ratio	0.449** (0.114)	4.76** (1.33)	2.44** (0.61)	2.46** (0.60)	2.62** (0.61)	2.57** (0.66)
housing expense-to-income ratio	−0.048 (.110)	−0.11 (1.29)	−0.18 (0.68)	−0.30 (0.68)	−0.50 (0.70)	−0.54 (0.74)
medium loan-to-value ratio ($0.80 \leq$ *loan-value ratio* ≤ 0.95)	0.031* (0.013)	0.46** (0.16)	0.21** (0.08)	0.22** (0.08)	0.22** (0.08)	0.22** (0.08)
high loan-to-value ratio (*loan-value ratio* ≥ 0.95)	0.189** (0.050)	1.49** (0.32)	0.79** (0.18)	0.79** (0.18)	0.84** (0.18)	0.79** (0.18)
consumer credit score	0.031** (0.005)	0.29** (0.04)	0.15** (0.02)	0.16** (0.02)	0.34** (0.11)	0.16** (0.02)
mortgage credit score	0.021 (0.011)	0.28* (0.14)	0.15* (0.07)	0.11 (0.08)	0.16 (0.10)	0.11 (0.08)
public bad credit record	0.197** (0.035)	1.23** (0.20)	0.70** (0.12)	0.70** (0.12)	0.72** (0.12)	0.70** (0.12)
denied mortgage insurance	0.702** (0.045)	4.55** (0.57)	2.56** (0.30)	2.59** (0.29)	2.59** (0.30)	2.59** (0.29)

(Table 11.2 continued)

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
Applicant single; HS diploma; industry unemployment rate				5.85 (< 0.001)	5.22 (0.001)	5.79 (< 0.001)
Additional credit rating indicator variables					1.22 (0.291)	
Race interactions and black						4.96 (0.002)
Race interactions only						0.27 (0.766)
Difference in predicted probability of denial, white vs. black (percentage points)	8.4%	6.0%	7.1%	6.6%	6.3%	6.5%

These regressions were estimated using the $n = 2380$ observations in the Boston HMDA data set described in Appendix 11.1. The linear probability model was estimated by OLS, and probit and logit regressions were estimated by maximum likelihood. Standard errors are given in parentheses under the coefficients and *p*-values are given in parentheses under the *F*-statistics. The change in predicted probability in the final row was computed for a hypothetical applicant whose values of the regressors, other than race, equal the sample mean. Individual coefficients are statistically significant at the *5% or **1% level.

Probit in R

```
bm1 <- glm(deny ~ pi_rat+black, data=hmda, family = binomial(link="probit"))
coeftest(bm1)

z test of coefficients:

             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -2.258787   0.136691 -16.5248 < 2.2e-16 ***
pi_rat       2.741779   0.380469   7.2063 5.749e-13 ***
blackTRUE    0.708155   0.083352   8.4959 < 2.2e-16 ***
--
    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

predict(bm1, data.frame(pi_rat=.3,black=FALSE),type = "response")
    0.07546516
predict(bm1, data.frame(pi_rat=.3,black=TRUE),type = "response")
    0.2332769
```

Why use the logistic CDF?

Has some nice properties:

- Gives us more of the 'S' shape
- $Pr(Y = 1|X)$ is increasing in X if $\beta_1 > 0$.
- $Pr(Y = 1|X) \in [0, 1]$ for all X
- Easy to compute: $\frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$ has analytic derivatives too.
- Log odds interpretation
    - $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X$
    - β_1 tells us how log odds ratio responds to X.
    - $\frac{p}{1-p} \in (-\infty, \infty)$ which fixes the $[0, 1]$ problem in the other direction.
    - more common in other fields (epidemiology, biostats, etc.).
- Also has the property that $F(z) = 1 - F(-z)$.
- Similar to probit but different scale of coefficients
- Logit/Logistic are sometimes used interchangeably but sometimes mean different things depending on the literature.

Logit in R

```
bm1 <-glm(deny~pi_rat+black,data=hmda, family=binomial(link="logit"))
coeftest(bm1)

z test of coefficients:

            Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -4.12556    0.26841 -15.3701 < 2.2e-16 ***
pi_rat       5.37036    0.72831   7.3737  1.66e-13 ***
blackTRUE    1.27278    0.14620   8.7059 < 2.2e-16 ***
--
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> predict(bm1, data.frame(pi_rat=.3,black=TRUE),type = "response")
0.2241459
> predict(bm1, data.frame(pi_rat=.3,black=FALSE),type = "response")
0.07485143
```

A quick comparison

- LPM prediction departs greatly from CDF long before $[0, 1]$ limits.
- We get probabilities that are too extreme even for $X\hat{\beta}$ “in bounds”.
- Some (MHE) argue that though $\hat{Y}$ is flawed, constant marginal effects are still OK.
- Logit and Probit are highly similar

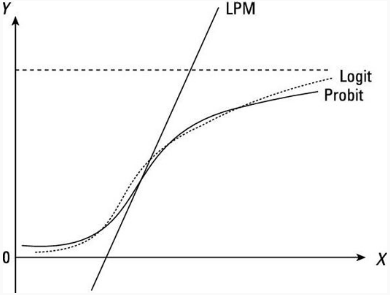


TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: *deny* = 1 if mortgage application is denied, = 0 if accepted; 2380 observations.

Regression Model	*LPM*	*Logit*	*Probit*	*Probit*	*Probit*	*Probit*
Regressor	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
black	0.084** (0.023)	0.688** (0.182)	0.389** (0.098)	0.371** (0.099)	0.363** (0.100)	0.246 (0.448)
P/I ratio	0.449** (0.114)	4.76** (1.33)	2.44** (0.61)	2.46** (0.60)	2.62** (0.61)	2.57** (0.66)
housing expense-to-income ratio	−0.048 (.110)	−0.11 (1.29)	−0.18 (0.68)	−0.30 (0.68)	−0.50 (0.70)	−0.54 (0.74)
medium loan-to-value ratio ($0.80 \le$ *loan-value ratio* ≤ 0.95)	0.031* (0.013)	0.46** (0.16)	0.21** (0.08)	0.22** (0.08)	0.22** (0.08)	0.22** (0.08)
high loan-to-value ratio (*loan-value ratio* ≥ 0.95)	0.189** (0.050)	1.49** (0.32)	0.79** (0.18)	0.79** (0.18)	0.84** (0.18)	0.79** (0.18)
consumer credit score	0.031** (0.005)	0.29** (0.04)	0.15** (0.02)	0.16** (0.02)	0.34** (0.11)	0.16** (0.02)
mortgage credit score	0.021 (0.011)	0.28* (0.14)	0.15* (0.07)	0.11 (0.08)	0.16 (0.10)	0.11 (0.08)
public bad credit record	0.197** (0.035)	1.23** (0.20)	0.70** (0.12)	0.70** (0.12)	0.72** (0.12)	0.70** (0.12)
denied mortgage insurance	0.702** (0.045)	4.55** (0.57)	2.56** (0.30)	2.59** (0.29)	2.59** (0.30)	2.59** (0.29)

(Table 11.2 continued)

***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**

	(1)	(2)	(3)	(4)	(5)	(6)
Applicant single; HS diploma; industry unemployment rate				5.85 (< 0.001)	5.22 (0.001)	5.79 (< 0.001)
Additional credit rating indicator variables					1.22 (0.291)	
Race interactions and black						4.96 (0.002)
Race interactions only						0.27 (0.766)
Difference in predicted probability of denial, white vs. black (percentage points)	8.4%	6.0%	7.1%	6.6%	6.3%	6.5%

These regressions were estimated using the $n = 2380$ observations in the Boston HMDA data set described in Appendix 11.1. The linear probability model was estimated by OLS, and probit and logit regressions were estimated by maximum likelihood. Standard errors are given in parentheses under the coefficients and *p*-values are given in parentheses under the *F*-statistics. The change in predicted probability in the final row was computed for a hypothetical applicant whose values of the regressors, other than race, equal the sample mean. Individual coefficients are statistically significant at the *5% or **1% level.

Latent Variables/ Limited Dependent Variables

An alternative way to think about this problem is that there is a continuously distributed Y^* that we as the econometrician don't observe.

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

- Instead we only see whether Y^* exceeds some threshold (in this case 0).
- We can think about Y^* as a latent variable.
- Sometimes you will see this description in the literature, everything else is the same!

Index Models

We sometimes call these single index models or threshold crossing models

$$Z_i = X_i \beta$$

- We start with a potentially large number of regressors in X_i but $X_i \beta = Z_i$ is a scalar
- We can just calculate $F(Z_i)$ for Logit or Probit (or some other CDF).
- Z_i is the index. if $Z_i = X_i \beta$ we say it is a linear index model.

What does software do?

- One temptation might be nonlinear least squares:

$$\hat{\beta}^{NLLS} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - \Phi(X_i\beta))^2$$

- Turns out this isn't what people do.
- We can't always directly estimate using the log-odds

$$\log\left(\frac{p}{1-p}\right) = \beta X_i + \varepsilon_i$$

- The problem is that p or $p(X_i)$ isn't really observed.

What does software do?

- Can construct an MLE:

$$\begin{aligned}\hat{\beta}^{MLE} &= \arg\max_{\beta} \prod_{i=1}^{N} F(Z_i)^{y_i}(1-F(Z_i))^{1-y_i} \\ Z_i &= \beta_0 + \beta_1 X_i\end{aligned}$$

- Probit: $F(Z_i) = \Phi(Z_i)$ and its derivative (density) $f(Z_i) = \phi(Z_i)$. Also is symmetric so that $1 - F(Z_i) = F(-Z_i)$.
- Logit: $F(Z_i) = \frac{1}{1+e^{-z}}$ and its derivative (density) $f(Z_i) = \frac{e^{-z}}{(1+e^{-z})^2}$ a more convenient property is that $\frac{f(z)}{F(z)} = 1 - F(z)$ this is called the hazard rate.

A probit trick

Let $q_i = 2y_i - 1$

$$F(q_i \cdot Z_i) = \begin{cases} F(Z_i) & \text{when } y_i = 1 \\ F(-Z_i) = 1 - F(Z_i) & \text{when } y_i = 0 \end{cases}$$

So that

$$l(y_1, \ldots, y_n | \beta) = \sum_{i=1}^{N} \ln F(q_i \cdot Z_i)$$

FOC of Log-Likelihood

$$
\begin{aligned}
l(y_1, \ldots, y_n|\beta) &= \sum_{i=1}^{N} y_i \ln F(Z_i) + (1-y_i)\ln(1-F(Z_i)) \\
\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{N} \frac{y_i}{F(Z_i)}\frac{dF}{d\beta}(Z_i) - \frac{1-y_i}{1-F(Z_i)}\frac{dF}{d\beta}(Z_i) \\
&= \sum_{i=1}^{N} \frac{y_i \cdot f(Z_i)}{F(Z_i)}\frac{dZ_i}{d\beta} - \sum_{i=1}^{N} \frac{(1-y_i)\cdot f(Z_i)}{1-F(Z_i)}\frac{dZ_i}{d\beta} \\
&= \sum_{i=1}^{N} \left[\frac{y_i \cdot f(Z_i)}{F(Z_i)}X_i - \frac{(1-y_i)\cdot f(Z_i)}{1-F(Z_i)}X_i\right]
\end{aligned}
$$

FOC of Log-Likelihood (Logit)

This is the score of the log-likelihood:

$$\frac{\partial l}{\partial \beta} = \nabla_\beta \cdot l(\mathbf{y};\beta) \quad = \quad \sum_{i=1}^{N} \left[y_i \frac{f(Z_i)}{F(Z_i)} - (1-y_i) \frac{f(Z_i)}{1-F(Z_i)} \right] \cdot X_i$$

It is technically also a moment condition. It is easy for the logit

$$\begin{aligned}
\nabla_\beta \cdot l(\mathbf{y};\beta) \quad &= \quad \sum_{i=1}^{N} [y_i(1-F(Z_i)) - (1-y_i)F(Z_i)] \cdot X_i \\
&= \quad \sum_{i=1}^{N} \underbrace{[y_i - F(Z_i)]}_{\varepsilon_i} \cdot X_i
\end{aligned}$$

This comes from the hazard rate.

FOC of Log-Likelihood (Probit)

This is the score of the log-likelihood:

$$\frac{\partial l}{\partial \beta} = \nabla_\beta \cdot l(\mathbf{y};\beta) = \sum_{i=1}^{N} \left[y_i \frac{f(Z_i)}{F(Z_i)} - (1-y_i)\frac{f(Z_i)}{1-F(Z_i)} \right] \cdot X_i$$

$$= \sum_{y_i=1} \frac{\phi(Z_i)}{\Phi(Z_i)} X_i + \sum_{y_i=0} \frac{-\phi(Z_i)}{1-\Phi(Z_i)} X_i$$

Using the $q_i = 2y_i - 1$ trick

$$\nabla_\beta \cdot l(\mathbf{y};\beta) = \sum_{i=1}^{N} \underbrace{\frac{q_i \phi(q_i Z_i)}{\Phi(Z_i)}}_{\lambda_i} X_i$$

The Hessian Matrix

We could also take second derivatives to get the Hessian matrix:

$$\frac{\partial l^2}{\partial\beta\partial\beta'} = -\sum_{i=1}^{N} y_i \frac{f(Z_i)f(Z_i) - f'(Z_i)F(Z_i)}{F(Z_i)^2} X_i X_i'$$
$$+\sum_{i=1}^{N}(1-y_i)\frac{f(Z_i)f(Z_i) - f'(Z_i)(1-F(Z_i))}{(1-F(Z_i))^2} X_i X_i'$$

This is a $K \times K$ matrix where K is the dimension of X or β.

The Hessian Matrix (Logit)

For the logit this is even easier (use the simplified logit score):

$$\begin{aligned}\frac{\partial l^2}{\partial\beta\partial\beta'} &= -\sum_{i=1}^{N} f(Z_i)X_iX_i' \\ &= -\sum_{i=1}^{N} F(Z_i)(1-F(Z_i))X_iX_i'\end{aligned}$$

This is negative semi definite

The Hessian Matrix (Probit)

Recall

$$\nabla_\beta \cdot l(\mathbf{y};\beta) = \sum_{i=1}^{N} \underbrace{\frac{q_i\phi(q_i Z_i)}{\Phi(Z_i)}}_{\lambda_i} X_i$$

Take another derivative and recall $\phi'(z_i) = -z_i\phi(z_i)$

$$\begin{aligned} \nabla^2_\beta \cdot l(\mathbf{y};\beta) &= \sum_{i=1}^{N} \frac{q_i\phi'(q_i Z_i)\Phi(z_i) - q_i\phi(z_i)^2}{\Phi(z_i)^2} X_i X_i' \\ &= -\lambda_i(z_i+\lambda_i)\cdot X_i X_i' \end{aligned}$$

Hard to show but this is negative definite too.

Estimation

- We can try to find the values of β which make the average score $= 0$ (the FOC).
- But no closed form solution!
- Recall Taylor's Rule:

$$f(x + \Delta x) = f(x_0) + f'(x_0)\Delta x + \frac{1}{2} f''(x_0)(\Delta x)^2$$

Goal is to find the case where $f'(x) \approx 0$ so take derivative w.r.t Δx:

$$\frac{d}{d\Delta x}\left[f(x_0) + f'(x_0)\Delta x + \frac{1}{2} f''(x_0)(\Delta x)^2\right] = f'(x_0) + f''(x_0)(\Delta x) = 0$$

Solve for Δx

$$\Delta x = -f'(x_0)/f''(x_0)$$

Estimation

- In multiple dimensions this becomes:

$$x_{n+1} = x_n - \alpha \cdot [\mathbf{H}_f(x_n)]^{-1} \nabla f(x_n)$$

- $\mathbf{H}_f(x_n)$ is the Hessian Matrix. $\nabla f(x_n)$ is the gradient.
- $\alpha \in [0, 1]$ is a parameter that determines step size
- Idea is that we approximate the likelihood with a quadratic function and minimize that (because we know how to solve those).
- Each step we update our quadratic approximation.
- If problem is convex this will always converge (and quickly)
- Most software "cheats" and doesn't compute $[\mathbf{H}_f(x_n)]^{-1}$ but uses tricks to update on the fly (BFGS, Broyden, DFP, SR1). Mostly you see these options in your software.

Marginal effects

$$\frac{\partial E[Y_i|X_i]}{\partial X_{ik}} = f(Z_i)\beta_k$$

- The whole point was that we wanted marginal effects not to be constant
- So where do we evaluate?
  - Software often plugs in mean or median values for each component
  - Alternatively we can integrate over X and compute:

$$E_{X_i}[f(Z_i)\beta_k]$$

  - The right thing to do is probably to plot the response surface (either probability) or change in probability over all X.

Inference

- If we have the Hessian Matrix, inference is straightforward.
- $\mathbf{H}_f(\hat{\beta}^{MLE})$ tells us about the curvature of the log-likelihood around the maximum.
  - Function is flat $\rightarrow$ not very precise estimates of parameters
  - Function is steep $\rightarrow$ precise estimates of parameters
- Construct Fisher Information $I(\hat{\beta}^{MLE}) = E[H_f(\hat{\beta}^{MLE})]$ where expectation is over the data.
  - Logit does not depend on y_i so $E[H_f(\hat{\beta}^{MLE})] = H_f(\hat{\beta}^{MLE})$.
  - Probit does depend on y_i so $E[H_f(\hat{\beta}^{MLE})] \neq H_f(\hat{\beta}^{MLE})$.
- Inverse Fisher information $E[H_f(\hat{\beta}^{MLE})]^{-1}$ is an estimate of the variance covariance matrix for $\hat{\beta}$.
- $\sqrt{diag[E[H_f(\hat{\beta}^{MLE})]^{-1}]}$ is an estimate for $SE(\hat{\beta})$.

Goodness of Fit #1: Pseudo R^2

How well does the model fit the data?

- No R^2 measure (why not?).
- Well we have likelihood units so average likelihood tells us something but is hard to interpret.
- $\rho = 1 - \frac{LL(\hat{\beta}^{MLE})}{LL(\beta_0)}$ where $LL(\beta_0)$ is the likelihood of a model with just a constant (unconditional probability of success).
  - If we don't do any better than unconditional mean then $\rho = 0$.
  - Won't ever get all of the way to $\rho = 1$.

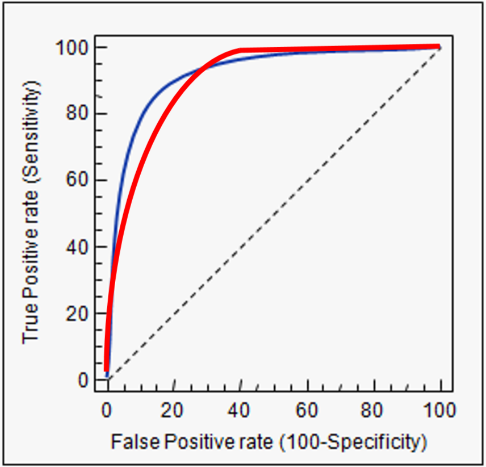
Goodness of Fit #2: Confusion Matrix

- Machine learning likes to think about this problem more like classification then regression.
- A caution: these are regression models not classification models.
- Predict either $\hat{y}_i = 1$ or $\hat{y}_i = 0$ for each observation.
- Predict $\hat{y}_i = 1$ if $Pr(y_i = 1|X_i = x) \geq 0.5$ or $F(X_i\hat{\beta}) > 0.5$.
- Imagine for cells Prediction: $\{Success, Failure\}$, Outcome $\{Success, Failure\}$
- Can construct this using the R package `caret` and command `caret`.

ROC Curve/ AOC

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

ROC Curve/ AOC



- At each predicted probability calculate both True Positive Rate and False Positive Rate.

Binary Choice: Overview

Many problems we are interested in look at discrete rather than continuous outcomes.

- We are familiar with limitations of the linear probability model (LPM)
  - Predictions outside of $[0, 1]$
  - Estimates of marginal effects need not be consistent.
- What about the case where Y is binary and a regressor X is endogenous?
  - The usual 2SLS estimator is NOT consistent.
  - Or we can ignore the fact that Y is binary...
  - Neither seems like a good option
- Suppose we have panel data on repeated binary choices
  - Adding FE to the probit model produces biased estimates.

Problem #1: Endogeneity

Four possible solutions (maybe there are more?)

1. Close eyes, run the LPM with instruments (Suggested by MHE).
2. Specify the distribution of errors in first and second stage and do MLE (`biprobit` in STATA).
3. Control Function Estimation
4. 'Special Regressor' Methods

Problem #1: Endogeneity

Setup:

- Binary variable D: the outcome of interest
- X is a vector of observed regressors with coefficient β
  - (Can think about X^e: endogenous and X^0: exogenous).
  - In an treatment model we might have that T is a binary treatment indicator within X
- ϵ is unobserved error. Specifying $f(e)$ can give logit/probit.
- Threshold Crossing / Latent Variable Model:

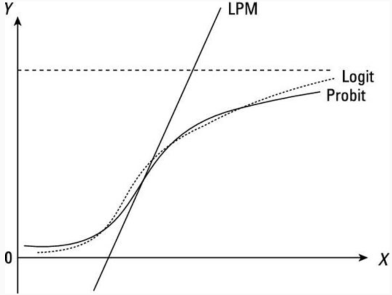
$$D = \mathbf{1}(X\beta + \epsilon \geq 0)$$

- Goal is not usually $\hat{\beta}$ or it's CI, but rather $P(D = 1|X)$ or $\frac{\partial P[D=1|x]}{\partial X}$ (marginal effects).

Linear Probability Model

Consider the LPM with a single continuous regressor

- LPM prediction departs greatly from CDF long before $[0, 1]$ limits.
- We get probabilities that are too extreme even for $X\hat{\beta}$ "in bounds".
- Some (MHE) argue that though $\hat{Y}$ is flawed, constant marginal effects are still OK.



Some well known textbooks

(Baby) Wooldrige:

> *"Even with these problems, the linear probability model is useful and often applied in economics. It usually works well for values of the independent variables that are near the averages in the sample." (2009, p. 249)*

- Mentions heteroskedasticity of error (which is binomial given X) but does not address the violation of the first LSA.

Some well known textbooks

Angrist and Pischke (MHE)

- several examples where marginal effects of probit and LPM are "indistinguishable".

  ...while a nonlinear model may fit the CEF (conditional expectation function) for LDVs (limited dependent variable models) more closely than a linear model, when it comes to marginal effects, this probably matters little. This optimistic conclusion is not a theorem, but as in the empirical example here, it seems to be fairly robustly true.(2009, p. 107)

and continue...

...extra complexity comes into the inference step as well, since we need standard errors for marginal effects. (ibid.)

Linear Probability Model

How does the LPM work?

$$D = X\beta + \varepsilon$$

- Estimated $\hat{\beta}$ are the MFX.
- With exogenous X we have $E[D|X] = Pr[D = 1|X] = X\beta$.
- If some elements of X (including treatment indicators) are endogenous or mismeasured they will be correlated with E.
- In that case we can do IV via 2SLS or IV-GMM given some instruments Z.
- We need the usual $E[\varepsilon|X] = 0$ or $E[\varepsilon|Z] = 0$.

Linear Probability Model

How does the LPM work?

$$D = X\beta + \varepsilon$$

- Estimated $\hat{\beta}$ are the MFX.
- With exogenous X we have $E[D|X] = Pr[D = 1|X] = X\beta$.
- If some elements of X (including treatment indicators) are endogenous or mismeasured they will be correlated with E.
- In that case we can do IV via 2SLS or IV-GMM given some instruments Z.
- We need the usual $E[\varepsilon|X] = 0$ or $E[\varepsilon|Z] = 0$.
- An obvious flaw: Given any $\varepsilon|X$ must equal either $1 - X\beta$ or $-X\beta$ which are functions of X
- Only the trivial binary X with no other regressors satisfies this!

Alarming Example: Lewbel Dong and Yang (2012)

- LPM is not just about taste and convenience.
- Three treated observations, three untreated
- Assume that $f(\varepsilon) \sim N(0, \sigma^2)$

$$D = I(1 + Treated + R + \varepsilon \geq 0)$$

- Each individual treatment effect given by:

$$I(2 + R + \varepsilon \geq 0) - I(1 + R + \varepsilon \geq 0) = I(0 \leq 1 + R + \varepsilon \leq 1)$$

- All treatment effects are positive for all (R, ε).
- Construct a sample where true effect $= 1$ for 5th individual, 0 otherwise. $ATE = \frac{1}{6}$.

Alarming Example: Lewbel Dong and Yang (2012)

```
. list
     |     R   Treated   D |
  1. |  -1.8         0   0 |
  2. |   -.9         0   1 |
  3. |  -.92         0   1 |
  4. |  -2.1         1   0 |
  5. | -1.92         1   1 |
  6. |    10         1   1 |

. reg D Treated R, robust

Linear regression                               Number of obs     =          6
                                                F(2, 3)           =       1.02
                                                Prob > F          =     0.4604
                                                R-squared         =     0.1704
                                                Root MSE          =     .60723

---------------------------------------
             |               Robust
           D |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------+--------------------------------
     Treated |  -.1550841   .5844637    -0.27   0.808    -2.015108     1.70494
           R |   .0484638   .0419179     1.16   0.331    -.0849376    .1818651
       _cons |   .7251463   .3676811     1.97   0.143    -.4449791    1.895272
---------------------------------------
. nlcom _b[Treated]/_b[R]
       _nl_1:  _b[Treated]/_b[R]
```

Alarming Example: Lewbel Dong and Yang (2012)

- That went well, except that:
  - we got the wrong sign of β_T
  - β_1/β_2 was the wrong sign and three times too big.
- this is not because of small sample size or $\beta_1 \approx 0$.
- As $n \to \infty$ we can get an arbitrarily precise wrong answer.
- We don't even get the sign right!
- This is still in OLS (not much hope for 2SLS).

```
. expand 30
(...)
. reg D Treated R, robust

Linear regression                               Number of obs     =        180
                                                F(2, 177)         =      59.93
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1704
                                                Root MSE          =       .433

---------------------------------------
             |               Robust
           D |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

Solution #0 : LPM

Advantages

- Just like 2SLS.
- Computationally easy (no numerical searches)
- Missing Z is about efficiency not consistency.
- X^e can be discrete or continuous (same estimator)
- allows general heteroskedasticity (random coefficients)

Disadvantages

- $\hat{F}_D$ is linear not S-shaped, approximation valid for small range of X.
- $\hat{F}_D$ can be outside $[0, 1]$.
- no element of X can have ∞ support (e.g. no normally distributed regressors).
- ε not independent of any regressors (even the exogenous ones). How do we also get $E[X^0\varepsilon] = 0$?

Solution #1 : MLE

$$D = I(X'\beta + \varepsilon \geq 0) \quad \text{and} \quad X^e = G(Z, \theta, e)$$

- Fully specified G (could be vector). Could be linear if X^e continuous or probit if X^e binary.
- Need to fully specify distribution of $(\varepsilon, e, |Z)$, parametrized.
- Implementation (see book), `biprobit` for joint normal in Stata.

Solution #1 : MLE

Advantages

- Nests logit, probit, etc. as special cases.
- Can have any kind of X^e
- Allows heteroskedasticity, random coefficients
- Asymptotically efficient (if correctly specified)

Disadvantages

- Need to parametrize everything G, $F_{\varepsilon,e|Z}$.
- Numerical optimization issues
- Many nusiance parameters, sometimes poorly identified, especially with discrete X^e, correlation between latent (ε, e).
- Need to know all required instruments Z. Omitting just one Z causes inconsistency in G. (Not sure if something is exogenous, too bad!).

Solution #2 : Control Functions

$$
\begin{aligned}
D &= I(X'\beta + \varepsilon \geq 0) \quad \text{and} \\
X^e &= G(Z) + e \quad \text{or} \quad X^e = G(Z, e) \quad \text{identified and invertible in } e \\
\varepsilon &= \lambda' e + U \quad \text{or} \quad \varepsilon = H(U, e) \quad \text{with conditions and } U \perp Z, e.
\end{aligned}
$$

Simple Case:

- Estimate a vector of functions G in the X^e models, get estimated errors $\hat{e}$.
- Estimate the D model including $\hat{e}$ as additional regressors in addition to X.
- This "cleans" the errors in U.

Solution #2 : Control Functions (Stata Version)

- `ivprobit` assumes that $G(Z, e)$ is linear, and (e, ε) jointly normal, independent of Z.
- not exactly the semi-parametric flexibility we were looking for...
- It is actually Control Function not IV

$$
\begin{aligned}
D &= I(X^e \beta_e + X^0 \beta_0 + \varepsilon \geq 0) \\
X^e &= \gamma Z + e
\end{aligned}
$$

Run first-stage OLS and get residuals $\hat{e}$. Then plug into

$$
D = I(X^e \beta_e + X^0 \beta_0 + + \lambda \hat{e} + U \geq 0)
$$

and do a conventional probit estimator.

- If you forget a Z_2 the resulting model isn't even probit!
- God help you if X^e isnt continuous.

Solution #2 : Control Functions

$$D = I(X'\beta + \varepsilon \geq 0), \quad X^e = G(Z, e), \varepsilon = H(U, e), \quad U \perp X, e.$$

Much stronger requirements that 2SLS

- Must be able to solve for errors e in X^e equations (not just orthogonality)
- Endogeneity must be caused only by ε relation to e so after conditioning on e must be that $f(\varepsilon|e, X^e) = f(\varepsilon|e)$.
- I need a consistent estimator for e which means nothing is omitted.

Not Quite MLE

- First stage can be semi/non-parametric .
- Don't need to fully specify joint distribution of (ε, e) (Stata does though!).

Advantages

- Nests logit, probit, etc. as special cases
- Requires less parametric information than MLE
- Some versions are computationally easy without numerical optimization (Bootstrap!)
- Less efficient than MLE due to less restrictions,but can be semiparametrically efficient given information.

Solution #2 : Control Functions (Disadvantages: Not well known)

- Only allows limited heteroskedasticity
- Need to correctly specify vector $G(Z, e)$ including all Z. Omitting a Z or misspecified G causes inconsistency because we need to have joint conditions on (ε, e).
- Generally inconsistent for X^e that is discrete, censored, limited, or not continuous.
- If you cannot solve for a latent e in $G(Z, e)$ then you can't get $\hat{e}$ for the censored observations (e.g.: $X^e = \max(0, Z'\gamma + e)$.
- An observable e is $e = X^e - E[X^e|Z]$ but for discontinuous X^e that e violates assumptions (except in very strange cases)
    - Ex: $\varepsilon = [X^e - E[X^e|Z]]\lambda + U$ satisfies CF, but if X^e is discrete then e has some strange distribution that depends on regressors.
    - Hard to generate a model of behavior that justifies this!

Solution #2 : Control Functions Generalized Residuals

What if X^e isn't continuous? Technically possible...

- Given the probit estimate in first stage we could construct a generalized residual (see Imbens and Wooldridge notes)
- $e^g \propto E[\varepsilon|Z, e]$. An estimate $\hat{e}^g$ of e^g can be included as a regressor in the model to fix the endogeneity problem, just as $\hat{e}$ would have been used if the endogenous regressor were continuous.

Why would you ever want to do this..

- In the linear model we should just do IV with far fewer restrictions
- In the nonlinear model, $\hat{e}^g$ requires almost as many assumptions as MLE which is efficient!

Solution #3 : Special Regressor

This approach draws on these two papers:

- Lewbel (JoE 2002)
- Dong and Lewbel (Econometric Reviews 2015)

$$D = I(X'\beta + V + \varepsilon \geq 0)$$

Special Regressor V has three properties

- Exogenous $E[\varepsilon|V] = 0$
- Additively separable in the model
- Continuous distribution and large support (such as Normal)
- Helpful to have thick-tails (kurtosis). Why? We want to trace $Pr(D|V)$ from $[0, 1]$.

Learn about SR

Binary, ordered, and multinomial choice, censored regression, selection, and treatment models (Lewbel 1998, 2000, 2007a), truncated regression models (Khan and Lewbel 2007), binary panel models with FE (Honore and Lewbel 2002), dynamic choice models (Heckman and Navarro 2007, Abbring and Heckman 2007), contingent valuation models (Lewbel, Linton, and McFadden 2008), market equilibrium models of multinomial choice (Berry and Haile 2009a, 2009b), models with (partly) nonseparable errors (Lewbel 2007b, Matzkin 2007, Briesch, Chintagunta, and Matzkin 2009).

Other empirical applications: Anton, Fernandez Sainz, and Rodriguez-Poo (2002), Cogneau and Maurin (2002), Goux and Maurin (2005), Stewart (2005), Lewbel and Schennach (2007), and Tiwari, Mohnen, Palm, and van der Loeff (2007).

Precursors: Matzkin (1992, 1994) and Lewbel (1997).

Recent theory: Magnac and Maurin (2007, 2008), Jacho-Ch·vez (2009), Khan and Tamer (2010), and Khan and Nekipelov (2010a, 2010b).

Solution #3 : Special Regressor

Requirements:

$$D = I(X'\beta + V + \varepsilon \geq 0)$$

- Exogenous $E[\varepsilon|V] = 0$ (Strict Exogeneity) This is the key!
- Additively separable in the model
- Continuous distribution and large support (such as Normal)
- NOT interacted with other regressors.
- enters LINEARLY, e.g. V must be continuously distributed after conditioning on other regressors
- Can normalize its coefficient to 1
- 2SLS Assumptions: $E[\varepsilon|Z] = 0$ and $E[Z'X]$ is full rank.

Solution #3 : Special Regressor: How it works

1. Demean or center V at zero.
2. Assume that $f_v(V|Z, X^e) = f_v(V|Z)$ and let $\hat{f}_v(V|Z)$ be a nonparametric kernel estimator of $f_v(V|Z)$. Or just use a kernel of the whole thing $\hat{f}_v(V|Z, X^e)$
3. For each observation i, Construct $\hat{T}_i = I[D_i - I(V_i \geq 0)]/\hat{f}_v(V_i|Z_i)$
4. Linear 2SLS regression of $\hat{T}$ on X using instruments Z to get the estimated coefficients $\hat{\beta}$.

Here $\hat{f}_v(V|Z)$ is high dimensional. So will consider some simpler parametric or semi-parametric version of f_v.
By properly adjusting T_i we guarantee to stay in $[0, 1]$.

Solution #3 : Special Regressor Advantages

- Unlike LPM it stays "in bounds" and is consistent with threshold crossing models.
- Unlike MLE and CF, does not require correctly specified first stage model: any valid set of instruments may be used, with only efficiency at stake.
- Unlike MLE, the SR method has a linear form, not requiring iterative search
- Unlike CF, the SR method can be used when endogenous regressors X^e are discrete or limited; unlike ML there is a single estimation method, regardless of the characteristics of X^e
- Unlike MLE, the SR method permits unknown heteroskedasticity in the model errors.

Solution #3 : Special Regressor Disadvantages

$$D = I(X'\beta + V + \varepsilon \geq 0)$$

- Because assumptions are weaker we give up a lot of potential efficiency (larger SEs).
- Of course this presumes the assumptions were valid and alternatives were consistent.
- SR Methods are generally valid under more general conditions.

The average index function (AIF)/ Propensity Score

In the original problem

$$D \quad = \quad I(X'\beta + \varepsilon \geq 0)$$

- V is part of X with coefficient $= 1$
- When $\varepsilon \perp x$ write the propensity score: $E[D|X] = E[D|X\beta] = F_{-\varepsilon}(X\beta) = Pr(-\varepsilon \leq X\beta)$.
- Under independence $X \perp \varepsilon$ these are the same, under endogeneity or even heteroskedasticity they are not.

The average index function (AIF)/ Propensity Score

- Blundell and Powell (ReStud 2004, this is actually the most important control function paper) use the average structural function (ASF) $= F_{-\varepsilon}(X\beta)$ to summarize choice probabilities. But when $\varepsilon \perp X$ is violated then they have to compute $F_{-\varepsilon|X}(X\beta)$ which is quite difficult (especially semiparametrically).
- Lewbel, Dong and Tang (CJE 2012) propose using the AIF estimator $E[D|X\beta]$ instead.
- Like ASF the AIF is based on the estimated index $X\beta$ and is equal to the propensity score if $\varepsilon \perp X$. However, when this is violated (endogeneity, heteroskedasticity) the AIF is easier to estimate, via unidimensional nonparametric regression of D on $X\beta$.

What is the difference

Propensity Score: Conditions on ALL covariates using $F_{-\varepsilon|X}$.

ASF: Conditions on no covariates using $F_{-\varepsilon}$.

AIF: Conditions on index only using $F_{-\varepsilon|X\beta}$.

- Unlike ASF, AIF is always identified and easy to estimate.
- Unlike Propensity score AIF uses β and isn't high dimensional
- ASF, AIF and propensity score all coincide under exogeneity.

Marginal Effects

- With exogenous X: MFX are $m(X) = p'(X) = \frac{\partial E[D|X\beta]}{\partial X}$.
- Let $f_{-\varepsilon}$ be marginal pdf of $-\varepsilon$. If $D = I(X'\beta + \varepsilon \geq 0)$ with $\varepsilon \perp X$ then:

$$m(X\beta)\beta = \frac{\partial E[D|X]}{\partial X} = \frac{\partial E[D|X\beta]}{\partial X'\beta}\beta = f_{-\varepsilon}(X'\beta)\beta$$

With endogenous X:

- Propensity Score marginal effects are $m(X) = p'(X) = \frac{\partial E[D|X]}{\partial X}$.
- ASF marginal effects are $m(X) = \frac{\partial ASF(X'\beta)}{\partial X'\beta}\beta = f_{-\varepsilon}(X'\beta)\beta$.
- AIF marginal effects are $m(X) = \frac{\partial ASF(X'\beta)}{\partial X'\beta}\beta = \frac{\partial E[D|X'\beta]}{\partial X'\beta}\beta$

Given $\hat{\beta}$ ASF and AIF mfx require just one dimensional index derivative.

Binary Choice with Endogenous Regressors

- Linear probability models, Maximum Likelihood, and Control functions (including `ivprobit` have more drawbacks and limitations than are usually recognized.
- Special Regressor estimators are a viable alternative (or at least they have completely different drawbacks and may be more generally applicable than has been recognized).
- In practice, best might be to try all estimators and check robustness of results. Can use marginal effects to normalize them the same when comparing.
- Average Index Functions can be used to construct estimated probabilities and comparable marginal effects across estimators, often simpler to calculate than Average Structural Functions.
- Implementation of special regressor in Stata is done in `sspecialreg`.

Empirical Example: Dong and Lewbel (2015)

- Binary dependent variable: does i migrate from one state to another.
- Special Regressor V_i: age. Human capital theory suggests it should appear linearly (or at least monotonic) in a threshold crossing model
- Migration is drive by maximizing expected lifetime income and potential gain from a permanent change in income declines linearly in age.
- V_i is defined as negative of age, demeaned so that coefficient is positive with mean zero.
- Other endogenous regressors: family income pre migration, home ownership.

Empirical Example: Dong and Lewbel (2015)

As a reminder, normally we would be in trouble here:

- MLE would be very complicated with multiple endogenous variables
- Control functions `ivprobit` won't work with 0/1 homeowner variable.

Empirical Example: Dong and Lewbel (2015)

1990 PSID

- male head of household (23-59 years), completed education and not-retired (key!)
- $D = 1$ indicates migration during 1991-1993.
- 4689 Individuals, 807 migrants.
- Exogenous regressors: years of education, # of children, white, disabled, married.
- Instruments: level of govt benefits in 1989-1990, state median residential tax rate.

Empirical Example: Dong and Lewbel (2015)

Specifications

- Special Regressors: kernel density vs. sorted data density.
- Special Regressor: homoskedastic vs. heteroskedastic errors.
- LPM vs 2SLS
- Probit (assuming exogeneity)
- Control Function (`ivprobit`) misspecified for `homeowner` endogenous binary variable.

Empirical Example

Table: Marginal effects: binary outcome, binary endogenous regressor

	kdens	sortdens	kdens_hetero	sortdens_hetero	IV-LPM	probit	ivprobit
age	0.0146	0.0112	0.0071	0.0104	-0.0010	0.0019	-0.0005
	(0.003)***	(0.003)***	(0.003)*	(0.003)***	(0.002)	(0.001)**	(0.007)
log income	-0.0079	0.0024	0.0382	0.0176	0.0550	-0.0089	0.1406
	(0.028)	(0.027)	(0.024)	(0.026)	(0.080)	(0.007)	(0.286)
homeowner	0.0485	-0.0104	-0.0627	-0.0111	-0.3506	-0.0855	-1.0647
	(0.072)	(0.065)	(0.059)	(0.061)	(0.204)	(0.013)***	(0.708)
white	0.0095	0.0021	0.0021	0.0011	0.0086	-0.0099	0.0134
	(0.008)	(0.010)	(0.007)	(0.008)	(0.018)	(0.012)	(0.065)
disabled	0.1106	0.0730	0.0908	0.0916	0.0114	-0.0122	0.0104
	(0.036)**	(0.042)	(0.026)***	(0.037)*	(0.055)	(0.033)	(0.203)
education	-0.0043	-0.0023	-0.0038	-0.0036	0.0015	0.0004	0.0047
	(0.002)*	(0.003)	(0.002)*	(0.002)	(0.004)	(0.002)	(0.015)
married	0.0628	0.0437	0.0258	0.0303	0.0322	-0.0064	0.0749
	(0.020)**	(0.028)	(0.013)	(0.020)	(0.031)	(0.017)	(0.114)
nr. children	-0.0169	-0.0117	0.0006	-0.0021	0.0137	0.0097	0.0502
	(0.005)***	(0.005)*	(0.002)	(0.003)	(0.006)*	(0.005)*	(0.023)*

Note: bootstrapped standard errors in parentheses (100 replications)

Empirical Example: Dong and Lewbel (2015)

- SEs of MFX are computed from 100 bootstrap replications
- MFX of special regressor (age) is estimated as positive and significant but LPM and ivprobit estimate negative effects!
- household income and home ownership status do not seem to play significant roles in migration decision.
- Kernel density estimator seems to give most significant results.