# Econometrics I

## Lecture 10: Binary Choice

Paul T. Scott
NYU Stern

Fall 2021

# Binary Choice: Overview

Many problems involve 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.

# Linear Probability Model

First thing we might try is OLS

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

- What does $\beta_1$ 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

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

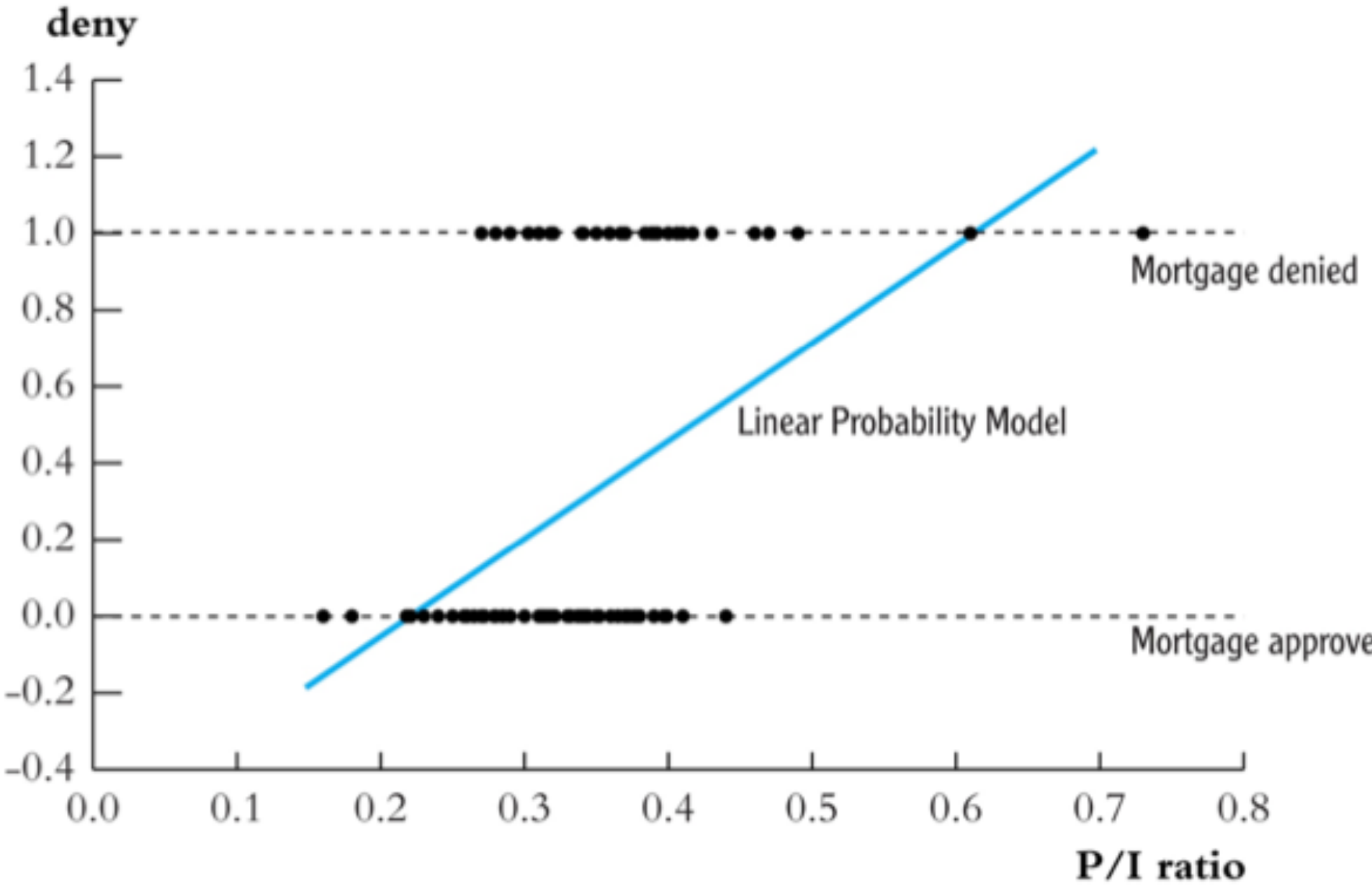
because:

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

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 $\beta_1$ 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
Mortgage approve
Linear Probability Model

# That didn't look great

- Is the marginal effect $\beta_1$ 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?

# 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:

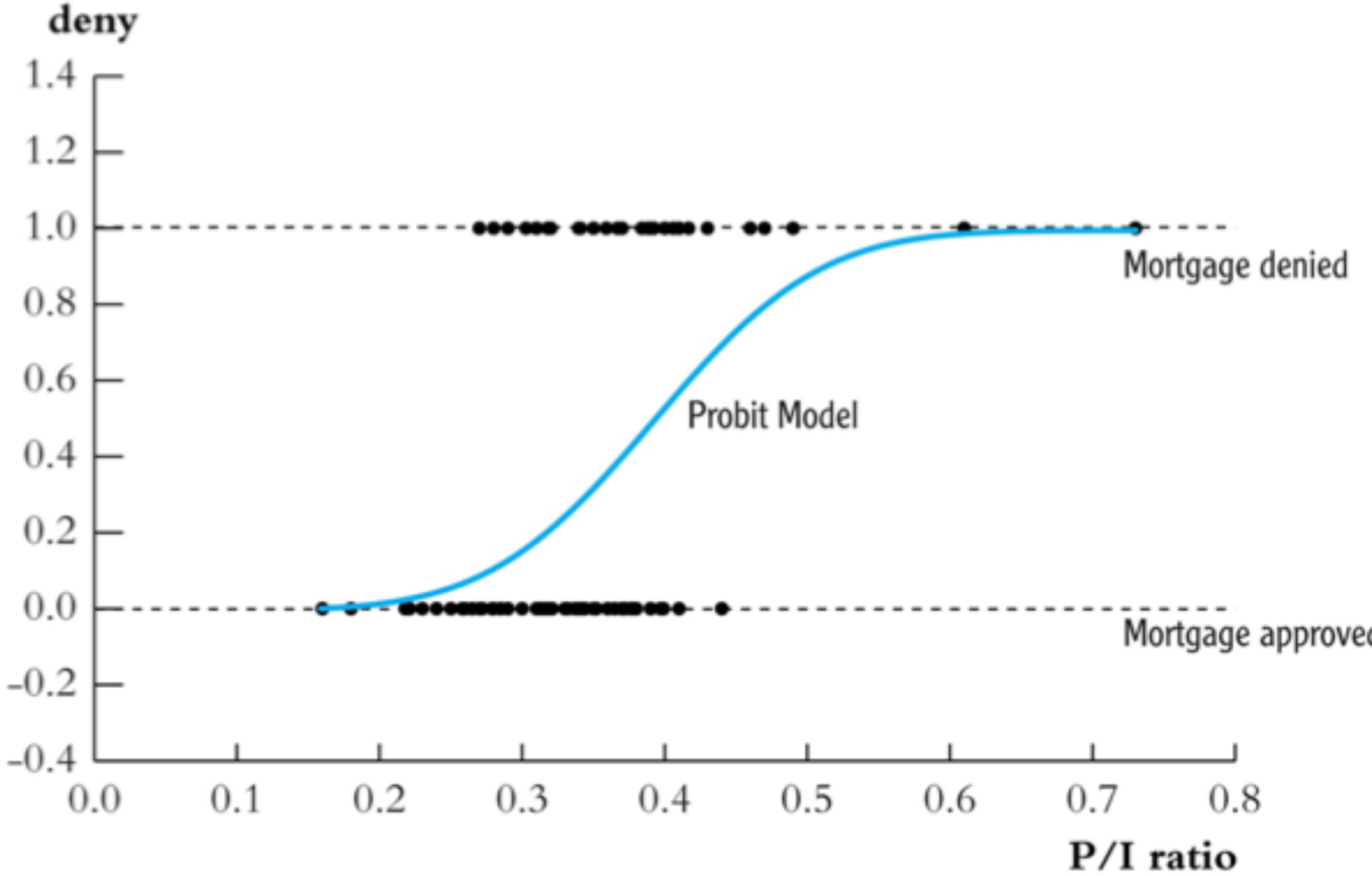
- Increasing $P/I$ from $0.3 \rightarrow 0.4$ increases probabilty of denial by 5.59 percentage points. (True at all level of $P/I$).
- At all $P/I$ levels blacks are 17.7 percentage points more likely to be denied.

# 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.

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.)


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 the $[0, 1]$ property).
- This $F(\cdot)$ is often called a link function.

# 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
  - $z_i = \beta \boldsymbol{x}_i$ (latent variable)
  - $\beta$ coefficients give derivative $dz/dx$
  - To map to changes in probabilities, use chain rule with $F$

**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 ≤ *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
```

- Probit predicts a 7.5% chance of mortgage denial for non-black applicants, and 23.3% chance for black ones.

# 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
  - $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X$
  - $\beta_1$ 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
- *Can* include fixed effects without incidental parameters problem (see **conditional logit**)
- Logit/Logistic are sometimes used interchangeably but sometimes mean different things depending on the literature.

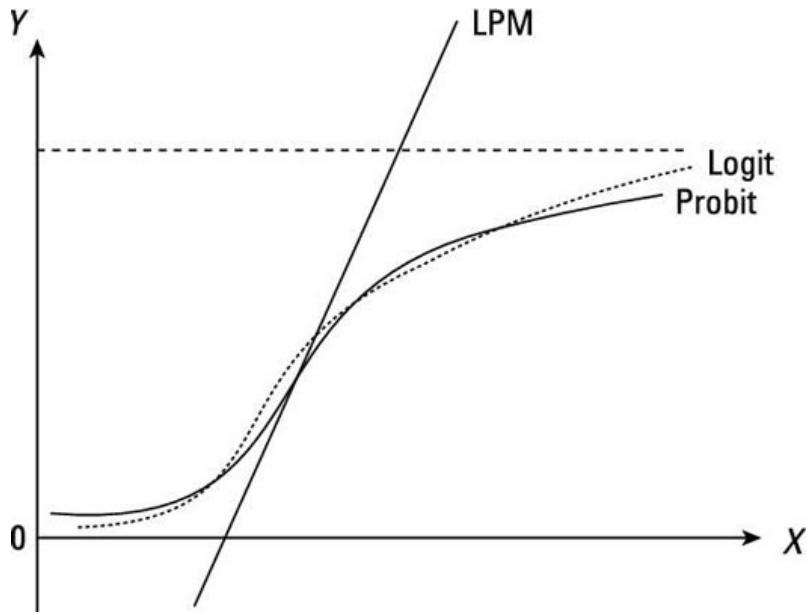
# Logit in R

```
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
```

- Logit predicts a 7.5% chance of mortgage denial for non-black applicants, and 22.4% chance for black ones. (Very similar to probit).

# 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



# 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?

- 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. Also symmetric.

# 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:

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

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

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

This comes from the hazard rate.

# FOC of Log-Likelihood (Probit)

This is the score of the log-likelihood:

$$\begin{aligned}\frac{\partial l}{\partial \beta} = \nabla_\beta \cdot l(\mathrm{y};\beta) &= \sum_{i=1}^{N}\left[y_i\frac{f(Z_i)}{F(Z_i)} - (1-y_i)\frac{f(Z_i)}{1-F(Z_i)}\right]\cdot X_i \\ &= \sum_{y_i=1}\frac{\phi(Z_i)}{\Phi(Z_i)}X_i + \sum_{y_i=0}\frac{-\phi(Z_i)}{1-\Phi(Z_i)}X_i\end{aligned}$$

Using the $q_i = 2y_i - 1$ trick

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

# 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 $\beta$.

# 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

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

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

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

Hard to show but this is negative definite too.

# Estimation

- We can try to find the values of $\beta$ 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 $\Delta x$:

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

Solve for $\Delta x$

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

# Estimation

- In multiple dimensions this becomes:

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

- $\mathrm{H}_f(x_n)$ 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 $[\mathrm{H}_f(x_n)]^{-1}$ 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.
- $H_f(\hat{\beta}^{MLE})$ 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.
- 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).
  - Note log-likelihood is negative; log-likelihood for model with only a constant is *more* negative.
  - 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`.

# Confusion Matrix for Binary Choice

| | | **Prediction** | | |
|---|---|---|---|---|
| | | $\hat{y}_i = 1$ | $\hat{y}_i = 1$ | **total** |
| **True outcome** | $y_i = 1$ | # true positives (TP) | # false negatives (FN) | PP |
| | $y_i = 0$ | # false positives (FP) | # true negatives (TN) | PN |
| | | P | N | |

- Accuracy: $(TP + TN)/(P + N)$, share of observations predicted correctly
- Recall: $TP/P$, rate at which positives are predicted correctly
- Precision: $TP/PP$, rate at which positive predictions are correct

# 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.

# Conditional logit with fixed effects

- Consider logit model with

$$Z_{it} = \boldsymbol{x}'_{it}\beta + \gamma_i.$$

$$Pr\left(Y_{it} = 1\right) = \frac{1}{1 + \exp\left(-\boldsymbol{x}'_{it}\beta - \gamma_i\right)}$$

- Consider $Pr\left(Y_{it} = 1 | Y_{it} + Y_{it+1} = 1\right)$. By Bayes' Rule,

$$Pr\left(Y_{it} = 1, Y_{it+1} = 0 | Y_{it} + Y_{it+1} = 1\right) = \frac{Pr\left(Y_{it} = 1\right) Pr\left(Y_{it+1} = 0\right)}{Pr\left(Y_{it} + Y_{it+1} = 1\right)}.$$

# Conditional MLE for logit with FE

- We have

$$\begin{aligned} Pr\left(Y_{it}=1\right) &= \frac{1}{1+\exp\left(-\boldsymbol{x}_{it}'\beta-\gamma_i\right)} \\ Pr\left(Y_{it+1}=0\right) &= \frac{\exp\left(-\boldsymbol{x}_{it+1}'\beta-\gamma_i\right)}{1+\exp\left(-\boldsymbol{x}_{it+1}'\beta-\gamma_i\right)} \end{aligned}$$

- Also,

$$\begin{aligned} Pr\left(Y_{it}+Y_{it+1}=1\right) &= Pr\left(Y_{it}=1,Y_{it+1}=0\right) \\ &\quad +Pr\left(Y_{it}=0,Y_{it+1}=1\right) \\ &= \frac{1}{1+\exp\left(-\boldsymbol{x}_{it}'\beta-\gamma_i\right)}\frac{\exp\left(-\boldsymbol{x}_{it+1}'\beta-\gamma_i\right)}{1+\exp\left(-\boldsymbol{x}_{it+1}'\beta-\gamma_i\right)} \\ &\quad +\frac{\exp\left(-\boldsymbol{x}_{it}'\beta-\gamma_i\right)}{1+\exp\left(-\boldsymbol{x}_{it}'\beta-\gamma_i\right)}\frac{1}{1+\exp\left(-\boldsymbol{x}_{it+t}'\beta-\gamma_i\right)}. \\ &= \frac{\exp\left(-\boldsymbol{x}_{it}'\beta-\gamma_i\right)+\exp\left(-\boldsymbol{x}_{it+1}'\beta-\gamma_i\right)}{\left(1+\exp\left(-\boldsymbol{x}_{it}'\beta-\gamma_i\right)\right)\left(1+\exp\left(-\boldsymbol{x}_{it+t}'\beta-\gamma_i\right)\right)} \end{aligned}$$

# Conditional MLE for logit with FE

- Plugging everything in,

$$
\begin{aligned}
Pr\left(Y_{it}=1, Y_{it+1}=0 \middle| Y_{it}+Y_{it+1}=1\right) &= \frac{\exp\left(-\boldsymbol{x}_{it+1}'\beta-\gamma_i\right)}{\exp\left(-\boldsymbol{x}_{it}'\beta-\gamma_i\right)+\exp\left(-\boldsymbol{x}_{it+1}'\beta-\gamma_i\right)} \\
&= \frac{\exp\left(-\boldsymbol{x}_{it+1}'\beta\right)}{\exp\left(-\boldsymbol{x}_{it}'\beta\right)+\exp\left(-\boldsymbol{x}_{it+1}'\beta\right)} \\
&= \frac{\exp\left(\left(\boldsymbol{x}_{it}-\boldsymbol{x}_{it+1}'\right)\beta\right)}{1+\exp\left(\left(\boldsymbol{x}_{it}-\boldsymbol{x}_{it+1}'\right)\beta\right)}
\end{aligned}
$$

- On the last line, we have a (conditional) likelihood that does not depend on the value of $\gamma_i$.

# Conditional MLE for logit with FE

- Like in the linear model, we're able to construct an estimator for $\beta$ that doesn't involve the fixed effect.
- This avoids the incidental parameter problem, meaning we can consistently estimate $\beta$ with fixed $T$ and $N \to \infty$.
- In linear models with fixed effects, the fixed effects estimator (with de-meaning) and the least squares with dummy variables (LSDV) estimators are equivalent. Neither will suffer from the incidental parameters problem.
- That's not the case for logit models. The conditional likelihood estimator is different from an unconditional likelihood estimator with fixed effects as parameters to be estimated. Only the conditional likelihood estimator escapes the incidental parameters problem.

# Problem #1: Endogeneity

Five 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. Probability Inversion
5. 'Special Regressor' Methods

# Problem #1: Endogeneity

Setup:

- Binary variable $Y$: the outcome of interest
- $X$ is a vector of observed regressors with coefficient $\beta$
  - (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 $X$
- $\varepsilon$ is unobserved error. Specifying $f(\varepsilon)$ can give logit/probit.
- Threshold Crossing / Latent Variable Model:

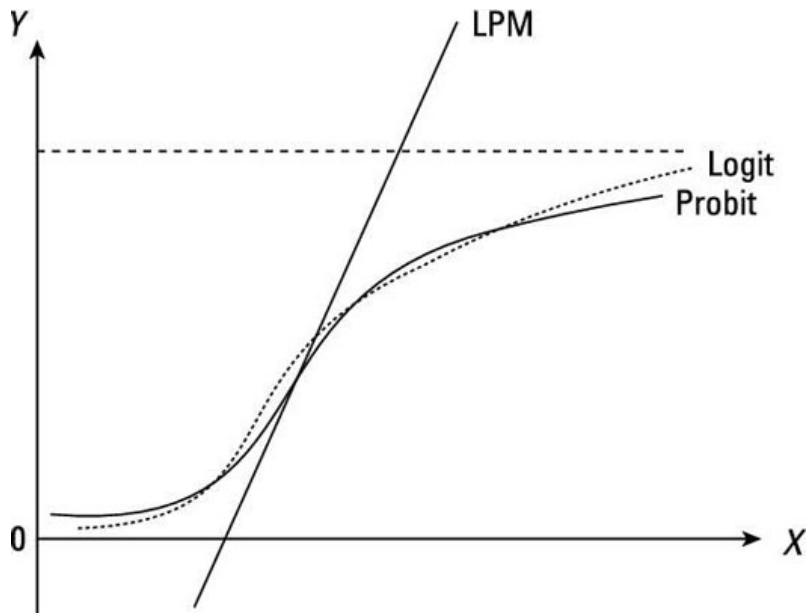
$$Y = 1(X\beta + \epsilon \geq 0)$$

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

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.)*

# 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)*

# Linear Probability Model

How does the LPM work?

$$Y = X\beta + \varepsilon$$

- Estimated $\hat{\beta}$ are the MFX.
- With exogenous $X$ we have $E[Y|X] = Pr[Y = 1|X] = X\beta$.
- If some elements of $X$ (including treatment indicators) are endogenous or mismeasured, they will be correlated with $\varepsilon$.
- 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?

$$Y = X\beta + \varepsilon$$

- Estimated $\hat{\beta}$ are the MFX.
- With exogenous $X$ we have $E[Y|X] = Pr[Y = 1|X] = X\beta$.
- If some elements of $X$ (including treatment indicators) are endogenous or mismeasured, they will be correlated with $\varepsilon$.
- 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: $\varepsilon|X$ must equal either $1 - X\beta$ or $-X\beta$ which are functions of $X$
- Difficult to satisfy $E(X\varepsilon) = 0$ or $E(Z\varepsilon) = 0$ unless we have a single binary regressor.

# Alarming Example: Lewbel Dong and Yang (2012)

- Three treated observations, three untreated
- Assume that $f(\varepsilon) \sim N(0, \sigma^2)$ with $\sigma^2$ **very small**

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

- 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 $(R, \varepsilon)$.
- Construct a sample where true effect $\approx 1$ for 5th individual, $\approx 0$ otherwise. $ATE \approx \frac{1}{6}$.

# 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]
------------------------------------------
           D |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |        -3.2   10.23042    -0.31   0.754    -23.25125    16.85125
------------------------------------------
```

# Alarming Example: Lewbel Dong and Yang (2012)

- That went well, except that:
  - we got the wrong sign of $\beta_T$
  - $\beta_1/\beta_2$ 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]
-------------+----------------------------------------------------------------
     Treated |  -.1550841   .0760907    -2.04   0.043    -.3052458   -.0049224
           R |   .0484638   .0054572     8.88   0.000     .0376941    .0592334
       _cons |   .7251463    .047868    15.15   0.000     .6306808    .8196117
------------------------------------------------------------------------------
```

# Solution #0 : LPM

## Advantages

- Just 2SLS
- Computationally easy (no numerical searches)
- Only need moment and rank conditions; don't need to specify distribution of endogenous regressors and instruments.

## Disadvantages

- linear approximation typically only valid for small range of $X$. Lewbel, Dong, and Yang show that the "local" approximation idea does not mean that we're always getting some sort of average effect.
- for asymptotics, no element of $X$ can have $\infty$ support (e.g. no normally distributed regressors).
- $\varepsilon$ not independent of any regressors (even the exogenous ones). How do we also get $E[X^0\varepsilon] = 0$ ?

# Solution #1 : MLE

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

- 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.

# Solution #1 : MLE

## Advantages

- Nests logit, probit, etc. as special cases.
- Can have any kind of $X^e$
- 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 $(\varepsilon, e)$.

# Solution #2 : Control Functions

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

Simple Case:

- Estimate a vector of functions $G$ in the $X^e$ models, get estimated errors $\hat{e}$.
- Estimate the $Y$ 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 $(e, \varepsilon)$ jointly normal, independent of $Z$.
- It is actually Control Function not IV

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

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

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

and do a conventional probit estimator.

# Solution #2 : Control Functions

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

Much stronger requirements that 2SLS

- Must be able to solve part of $X^e$ that causes endogeneity problem (not just orthogonality)
- Endogeneity must be caused only by $\varepsilon$ 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 from model of $X^e$

Not Quite MLE

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

# Solution #2 : Control Functions

## 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 $(\varepsilon, e)$.
- 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 : Probability Inversion

- Consider logit model with latent variable

$$Y_i^* = \beta X_i + \varepsilon_i,$$

- We can show that

$$\log\left(\frac{Pr\left(Y_i = 1|X_i\right)}{1 - Pr\left(Y_i = 1|X_i\right)}\right) = \beta X_i$$

- We can recover the value of the latent variable from the probabilities. Can we think about above equation as regression equation? What would be error term?

# Solution #3 : Probability Inversion

- Let's assume we have data that can be grouped by $t$, and

$$Y_{i,t}^* = \beta X_t + \xi_t + \varepsilon_{i,t}.$$

  We will maintain the assumption that $\varepsilon$ is i.i.d. and uncorrelated with $X_t$.

- Now, we have an equation with an error term:

$$\log\left(\frac{Pr\left(Y_{i,t}=1|X_t,\xi_t\right)}{1-Pr\left(Y_{i,t}=1|X_t,\xi_t\right)}\right) = \beta X_t + \xi_t$$

- If $E\left(\xi_t X_t\right) = 0$, we can estimate using OLS. If $E\left(\xi_t X_t\right) \neq 0$ but we have a valid instrument, we can use 2SLS.

# Solution #3 : Probability Inversion Interpretation

$$Y_{i,t}^* = \beta X_t + \xi_t + \varepsilon_{i,t}.$$

$$\log\left(\frac{Pr\left(Y_{i,t}=1|X_t,\xi_t\right)}{1-Pr\left(Y_{i,t}=1|X_t,\xi_t\right)}\right) = \beta X_t + \xi_t$$

- NB: we're in a setting where the regressors don't vary across individuals within $t$, and the endogeneity problem only comes through an unobservable variable that doesn't vary within $t$.
- What could $t$ represent?
- Potential applications?

## Advantages

- Linear regression: easy to compute, standard asymptotics
- When the framework is correct, naive use of MLE would give invalid standard errors; we need to cluster by $t$.
- Doesn't always require individual-level data. We can often construct probabilities from aggregated data.

## Disadvantages

- Requires large number of individuals that can be aggregated based on common values of $X$ and $\xi$. NB: not just aggregating conditional on values of $X$.
- Need to be able to estimate first-stage probabilities precisely. Probabilities estimates should be $p \in (0, 1)$.

# Solution #4: Special Regressor

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-Chavez (2009), Khan and Tamer (2010), and Khan and Nekipelov (2010a, 2010b).

# Ordered Probit/Logit

- Consider situation where dependent variable has discrete distribution but more than two outcomes:

$$Y_i \in \mathcal{J} = \{1, 2, 3, \dots\},$$

  Furthermore, suppose that the outcomes have a clear ranking (e.g., product reviews).

- Ordered probit/logit are extensions of the basic model that have several cutoffs for $Y_i^*$. These can still be estimated with MLE, and there are standard packages in most software systems.

# Multinomial Choice Models

- Consider situation where dependent variable has discrete distribution but more than two outcomes:

$$Y_i \in \mathcal{J} = \{1, 2, 3, \dots\},$$

but they aren't ranked in any particular way

- Things are no longer as simple as thinking of a link function (i.e., a CDF for $Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$)

- But we can use the latent variable formulation. Consider a value of the latent variable for each potential outcome $j \in \mathcal{J}$:

$$Y^*_{i,j} = \beta X_{i,j} + \varepsilon_{i,j},$$

and $Y_i = j$ if and only if $Y^*_{i,j} \geq Y^*_{i,j'}$ for all $j' \in \mathcal{J}$.

# Multinomial Choice Models

- NB: error term works differently. For logit model, error term now has Extreme Value Type I (Gumbel) distribution. Difference of two of those is logistic.
- Can still apply MLE, or use probability inversion to recover latent variable.