Lecture 3: Extremum Estimators

Chris Conlon

February 10, 2020

NYU Stern

Introduction

Consider a linear regression with $\varepsilon_i | X_i \sim N(0, \sigma^2)$

$$Y_{it} = X_i' \beta_i + \varepsilon_i$$

We've discussed the least squares estimator:

$$\widehat{\beta}_{ols} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i' \beta)^2$$

$$\widehat{\beta}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Review: What is a Likelihood?

Suppose we write down the joint distribution of our data (y_i, x_i) for $i = 1, \ldots, n$.

$$Pr(y_1, \ldots, y_n, x_1, \ldots, x_n | \theta)$$

If (y_i, x_i) are I.I.D then we can write this as:

$$Pr(y_1, \ldots, y_n, x_1, \ldots, x_n | \theta) = \prod_{i=1}^{N} Pr(y_i, x_i | \theta) \propto \prod_{i=1}^{N} Pr(y_i | x_i, \theta) = L(\mathbf{y} | \mathbf{x}, \theta)$$

We call this $L(\mathbf{y}|\mathbf{x}, \theta)$ the likelihood of the observed data.

MLE: Example

If we know the distribution of ε_i we can construct a maximum likelihood estimator

$$(\widehat{\beta}_{MLE}, \widehat{\sigma}^2_{MLE}) = \arg\min_{\beta,\sigma^2} L(\beta, \sigma^2)$$

Where

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{i=1}^{N} p(y_i | x_i, \beta, \sigma^2) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (Y_i - X_i'\beta)^2 \right] \\ l(\beta, \sigma^2) &= \sum_{i=1}^{N} -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - X_i'\beta)^2 \end{aligned}$$

MLE: FOC's

Take the FOC's

$$l(\beta, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2$$

Where

$$\frac{\partial l(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(Y_i - X_i'\beta) = 0 \rightarrow \widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$$

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2} = -N\frac{1}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2 = 0$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2$$

Note: the unbiased estimator uses $\frac{1}{N-K-1}$.

MLE: General Case

1. Start with the joint density of the data $Z_1, \ldots, Z_N$ with density $f_Z(z, \theta)$
2. Construct the likelihood function of the sample $z_1, \ldots, z_n$

$$L(\mathbf{z}|\theta) = \prod_{i=1}^{N} f_Z(z_i, \theta)$$

3. Construct the log likelihood (this has the same arg max)

$$l(\mathbf{z}|\theta) = \sum_{i=1}^{N} \ln f_Z(z_i, \theta)$$

4. Take the FOC's to find $\widehat{\theta}_{MLE}$

$$\theta : \frac{\partial l(\theta)}{\partial \theta} = 0$$

MLE in Detail

Basic Setup: we know $F(z|\theta_0)$ but not θ_0. We know $\theta_0 \in \Theta \subset \mathbb{R}^K$.

- Begin with a sample of z_i from $i = 1, \ldots, N$ which are I.I.D. with CDF $F(z|\theta_0)$.
- The MLE chooses

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} l(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \ln f_Z(z_i, \theta)$$

MLE: Technical Details

1. Consistency. When is it true that for $\epsilon > 0$?

$$\lim_{N \to \infty} \Pr\left(\left\|\hat{\theta}_{mle} - \theta_0\right\| > \varepsilon\right) = 0$$

2. Asymptotic Normality. What else do we need to show?

$$\sqrt{N}\left(\hat{\theta}_{mle} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, -\left[E\frac{\partial^2}{\partial\theta\partial\theta'}\left(Z_i, \theta_0\right)\right]^{-1}\right)$$

3. Optimization. How to we obtain $\widehat{\theta}_{MLE}$ anyway?

MLE: Example # 1

- $Z_i \sim N(\theta_0, 1)$ and $\Theta = (-\infty, \infty)$. In this case:

$$l(\theta) = -N \cdot \ln(2\pi) - \sum_{i=1}^{N} (z_i - \theta)^2 / 2$$

- MLE is $\widehat{\theta}_{MLE} = \overline{z}$ which is consistent for $\theta_0 = E[Z_i]$
- Asymptotic distribution is $\sqrt{N}(\overline{z} - \theta_0) \sim N(0, 1)$.
- Calculating mean is easy!

MLE: Example # 2

- $Z_i = (Y_i, X_i)$— X_i has finite mean and variance (but arbitrary distribution)
- $(Y_i | X_i = x) \sim N(x'\beta_0, \sigma_0^2)$

$$\widehat{\beta}_{MLE} = (X'X)^{-1}X'Y$$
$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum(y_i - x_i\widehat{\beta}_{MLE})^2$$

- We already have shown consistency and AN for linear regression with normally distributed errors...

MLE: Example # 3

- $Z_i = (Y_i, X_i)$— X_i has finite mean and variance (but arbitrary distribution)
- $Pr(Y_i = 1|X_i x) = \frac{e^{x'\theta_0}}{1+e^{x'\theta_0}}$
- Solution is the logit model.
- No simple MLE solution, establishing properties is not obvious...

Jensen's Inequality

Let $g(z)$ be a convex function. Then $\mathbb{E}[g(Z)] \geq g(\mathbb{E}[Z])$, with equality only in the case of a linear function.

More Technical Details

Define Y as the ratio of the density at θ to the density at the true value θ_0 both evaluated at Z

$$Y = \frac{f_Z(Z;\theta)}{f_Z(Z;\theta_0)}$$

- Let $g(a) = -\ln(a)$ so that $g'(a) = \frac{-1}{a}$ and $g''(a) = \frac{1}{a^2}$.
- Then by Jensen's Inequality $\mathbb{E}[-\ln Y] \geq -\ln \mathbb{E}[Y]$.
- This gives us

$$\mathbb{E}_z\left[-\ln\left(\frac{f_Z(Z;\theta)}{f_Z(Z;\theta_0)}\right)\right] \geq -\ln\left(\mathbb{E}_z\left[\frac{f_Z(Z;\theta)}{f_Z(Z;\theta_0)}\right]\right)$$

- The RHS is

$$\mathbb{E}_z\left[\frac{f_Z(Z;\theta)}{f_Z(Z;\theta_0)}\right] = \int \frac{f_Z(z;\theta)}{f_Z(z;\theta_0)} \cdot f_Z(z;\theta_0)\, dz = \int f_Z(z;\theta) dz = 1$$

More Technical Details

Because $\log(1) = 0$ this implies:

$$\mathbb{E}_z\left[-\ln\left(\frac{f_Z(Z;\theta)}{f_Z(Z;\theta_0)}\right)\right] \geq 0$$

Therefore

$$-\mathbb{E}\left[\ln f_Z(Z;\theta)\right] + \mathbb{E}\left[\ln f_Z(Z;\theta_0)\right] \geq 0$$
$$\mathbb{E}\left[\ln f_Z(Z;\theta_0)\right] \geq \mathbb{E}\left[\ln f_Z(Z;\theta)\right]$$

- We maximize the expected value of the log likelihood at the true value of θ!
- Helpful to work with $E[\log f(z;\theta)]$ sometimes.

Information Matrix Equality

We can relate the Fisher Information to the Hessian of the log-likelihood

$$\mathcal{I}(\theta_0) = -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial\theta\partial\theta}(z;\theta_0)\right] = \mathbb{E}\left[\frac{\partial \ln f}{\partial\theta}(z;\theta_0) \times \frac{\partial \ln f}{\partial\theta}(z;\theta_0)'\right]$$

- This is sometimes known as the outer product of scores.
- This matrix is negative definite
- Recall that $\mathbb{E}\left[\frac{\partial \ln f}{\partial\theta}(z;\theta_0)\right] \approx 0$ at the maximum

Proof

$$1 = \int_z f_Z(z;\theta)dz \Rightarrow 0 = \frac{\partial}{\partial\theta}\int_z f_Z(z;\theta)dz$$

With some regularity conditions

$$0 = \int_z \frac{\partial f_Z}{\partial\theta}(z;\theta)dz = \underbrace{\int_z \frac{\partial \ln f_Z}{\partial\theta}(z;\theta)\cdot f_Z(z;\theta)dz}_{\mathbb{E}\left[\frac{\partial \ln f_Z}{\partial\theta}(z;\theta_0)\right]}$$

- This gives us the FOC we needed.
- Can get information identity with another set of derivatives.

The Cramer-Rao Bound

We can relate the Fisher Information to the Hessian of the log-likelihood

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial\theta\partial\theta'}(Z|\theta)\right]$$

It turns out this provides a bound on the variance

$$\text{Var}(\hat{\theta}(Z)) \geq \mathcal{I}(\theta_0)^{-1}$$

Because we can't do better than Fisher Information we know that MLE is most efficient estimator!

MLE: Discussion

Tradeoffs

- How does this compare to GM Theorem?
- If MLE is most efficient estimate, why ever use something else?

Exponential Example

$$f_{Y|X}(y|x,\beta_0) = e^{x'\beta_0}\exp\left(-ye^{x'\beta_0}\right)$$

With log likelihood

$$l(\beta) = \sum_{i=1}^{N} \ln f_{Y|X}\left(y_i|x_i,\beta\right) = \sum_{i=1}^{N} X_i'\beta - y_i \cdot \exp\left(x_i'\beta\right)$$

And Score, Hessian, and Information Matrix:

$$\begin{aligned}
\mathcal{S}_i(y_i,x_i,\beta) &= x_i'\left(1 - y_i\exp\left(x_i'\beta\right)\right) \\
\mathcal{H}_i(y_i,x_i,\beta) &= -y_i x_i x_i'\exp\left(x_i'\beta\right) \\
\mathcal{I}\left(\beta_0\right) &= \mathbb{E}\left[YXX'\exp\left(X'\beta_0\right)\right] = \mathbb{E}\left[XX'\right]
\end{aligned}$$

Computing Maximum Likelihood Estimators

Newton's Method for Root Finding

Consider the Taylor series for $f(x)$ approximated around $f(x_0)$:

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0) + f''(x_0) \cdot (x - x_0)^2 + o_p(3)$$

Suppose we wanted to find a root of the equation where $f(x^*) = 0$ and solve for x:

$$0 = f(x_0) + f'(x_0) \cdot (x - x_0)$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

This gives us an iterative scheme to find x^*:

1. Start with some x_k. Calculate $f(x_k), f'(x_k)$
2. Update using $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$
3. Stop when $|x_{k+1} - x_k| < \epsilon_{tol}$.

Newton-Raphson for Minimization

We can re-write optimization as root finding;

- We want to know $\hat{\theta} = \arg\max_\theta \ell(\theta)$.
- Construct the FOCs $\frac{\partial \ell}{\partial \theta} = 0 \rightarrow$ and find the zeros.
- How? using Newton's method! Set $f(\theta) = \frac{\partial \ell}{\partial \theta}$

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 \ell}{\partial \theta^2}(\theta_k)\right]^{-1} \cdot \frac{\partial \ell}{\partial \theta}(\theta_k)$$

The SOC is that $\frac{\partial^2 \ell}{\partial \theta^2} > 0$. Ideally at all θ_k.
This is all for a single variable but the multivariate version is basically the same.

Newton's Method: Multivariate

Start with the objective $Q(\theta) = -l(\theta)$:

- Approximate $Q(\theta)$ around some initial guess θ_0 with a quadratic function
- Minimize the quadratic function (because that is easy) call that θ_1
- Update the approximation and repeat.

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial Q}{\partial\theta}(\theta_k)$$

- The equivalent SOC is that the Hessian Matrix is positive semi-definite (ideally at all θ).
- In that case the problem is globally convex and has a unique maximum that is easy to find.

Newton's Method

We can generalize to Quasi-Newton methods:

$$\theta_{k+1} = \theta_k - \lambda_k \underbrace{\left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right]^{-1}}_{A_k} \frac{\partial Q}{\partial\theta}(\theta_k)$$

Two Choices:

- Step length λ_k
- Step direction $d_k = A_k \frac{\partial Q}{\partial\theta}(\theta_k)$
- Often rescale the direction to be unit length $\frac{d_k}{\|d_k\|}$.
- If we use A_k as the true Hessian and $\lambda_k = 1$ this is a full Newton step.

Newton's Method: Alternatives

Choices for A_k

- $A_k = I_k$ (Identity) is known as gradient descent or steepest descent
- BHHH. Specific to MLE. Exploits the Fisher Information.

$$A_k = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ln f}{\partial \theta} (\theta_k) \frac{\partial \ln f}{\partial \theta'} (\theta_k) \right]^{-1}$$
$$= -\mathbb{E} \left[\frac{\partial^2 \ln f}{\partial \theta \partial \theta'} (Z, \theta^*) \right] = \mathbb{E} \left[\frac{\partial \ln f}{\partial \theta} (Z, \theta^*) \frac{\partial \ln f}{\partial \theta'} (Z, \theta^*) \right]$$

- Alternatives SR1 and DFP rely on an initial estimate of the Hessian matrix and then approximate an update to A_k.
- Usually updating the Hessian is the costly step.
- Non invertible Hessians are bad news.

Extended Example: Binary Choice

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

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

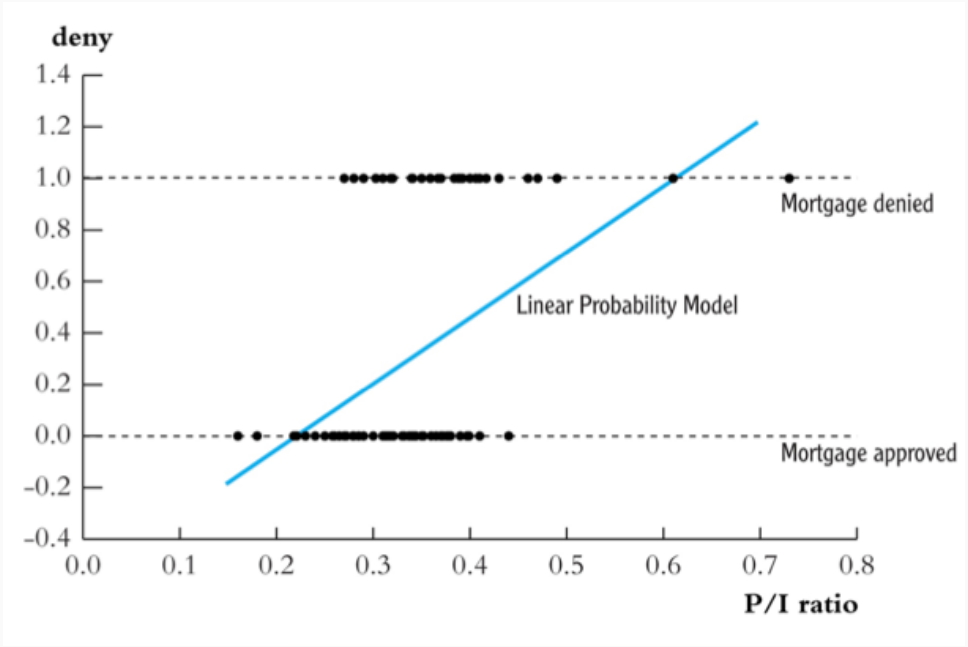
because:

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

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

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

Marginal Effects:

- Increasing P/I from $0.3 \to 0.4$ 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.)

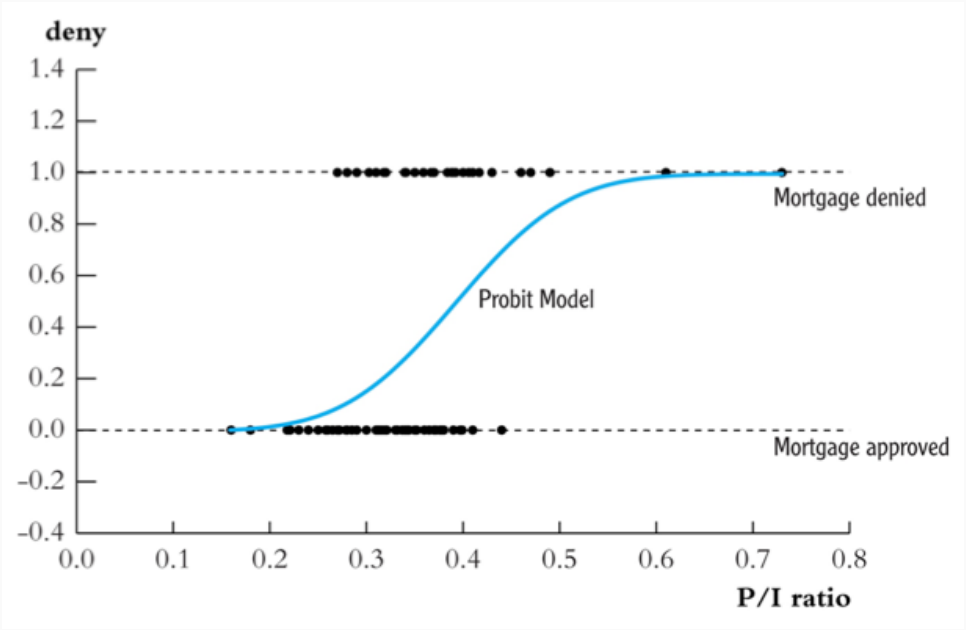
Moving Away from LPM

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	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
```

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

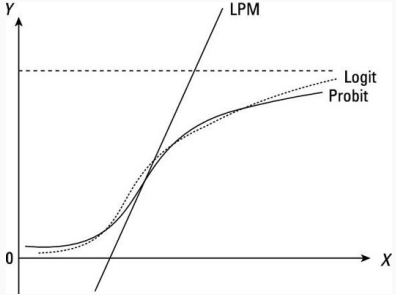


TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

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

Regression Model / 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 ≤ *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

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

This comes from the hazard rate.

FOC of Log-Likelihood (Probit)

This is the score of the log-likelihood:

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

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):

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

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 → not very precise estimates of parameters
  - Function is steep → 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

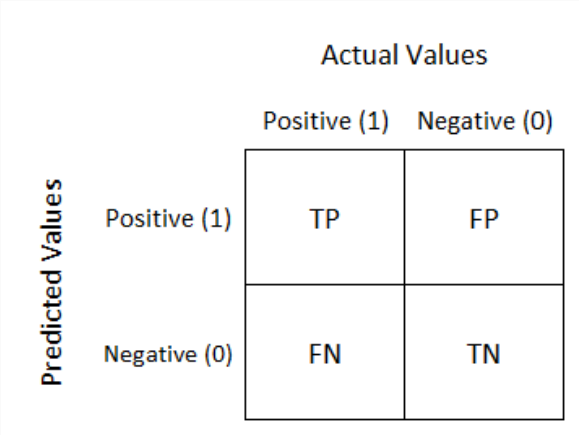
How well does the model fit the data?

- 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 $\rho = 1$.

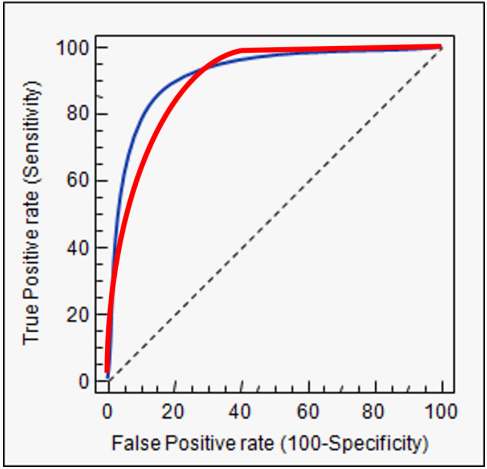
Goodness of Fit #2: Confusion Matrix

- 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



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:

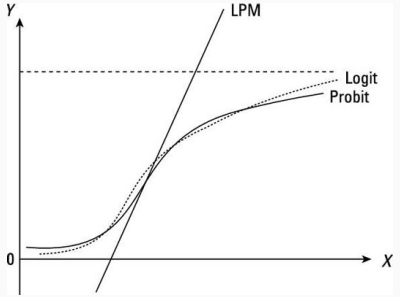
$$D = \mathbf{1}(X\beta + \epsilon \geq 0)$$

- 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. $ATE = \frac{1}{6}$.

Estimating Finite Mixtures

- In practice estimating finite mixture models can be tricky.
- A simple example is the mixture of normals (incomplete data likelihood)

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f(x_i|\mu_k, \sigma_k)$$

- We need to find both mixture weights $\pi_k = Pr(z_k)$ and the components (μ_k, σ_k) the weights define a valid probabiltiy measure $\sum_k \pi_k = 1$.
- Easy problem is label switching. Usually it helps to order the components by say decreasing $\pi_1 > \pi_2 > \ldots$ or $\mu_1 > \mu_2 > \ldots$
- The real problem is that which component you belong to is unobserved. We can add an extra indicator variable $z_{ik} \in \{0, 1\}$.
- We don't care about z_{ik} per-se so they are nuisance parameters.

Estimating Finite Mixtures

- We can write the complete data log-likelihood (as if we observed z_{ik}):

$$l(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} I[z_i = k] \pi_k f(x_i \mu_k, \sigma_k) \right)$$

- We can instead maximized the expected log-likelihood where we take the expectation $E_{z|\theta}$

$$\alpha_{ik}(\theta) = Pr(z_{ik} = 1|x_i, \theta) = \frac{f_k(x_i, z_k, \mu_k, \sigma_k)\pi_k}{\sum_{m=1}^{K} f_m(x_i, z_m, \mu_m, \sigma_m)\pi_m}$$

- Now we have a probability $\hat{\alpha}_{ik}$ that gives us the probability that i came from component k. We also compute $\hat{\pi}_k = \frac{1}{N} \sum_{i=1}^{N} \alpha_{ik}$

EM Algorithm

- Treat the $\hat{\alpha}_k(\theta^{(q)})$ as data and maximize to find μ_k, σ_k for each k

$$\hat{\theta}^{(q+1)} = \arg\max_{\theta} \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \hat{\alpha}_k(\theta^{(q)}) f(x_i|z_{ik}, \theta)\right)$$

- We iterate between updating $\hat{\alpha}_k(\theta^{(q)})$ (E-step) and $\hat{\theta}^{(q+1)}$ (M-step)
- For the mixture of normals we can compute the M-step very easily:

$$\begin{aligned}
\mu_k^{(q+1)} &= \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_k(\theta^{(q)}) x_i \\
\sigma_k^{(q+1)} &= \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_k(\theta^{(q)})(x_i - \overline{x})^2
\end{aligned}$$

EM Algorithm

- EM algorithm has the advantage that it avoids complicated integrals in computing the expected log-likelihood over the missing data.
- For a large set of families it is proven to converge to the MLE
- That convergence is monotonic and linear. (Newton's method is quadratic)
- This means it can be slow, but sometimes $\nabla_\theta f(\cdot)$ is really complicated.

Thanks!