# Lecture 2: Maximum Likelihood and Friends

Chris Conlon

February 7, 2023

NYU Stern

# 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 $Y_1, \dots, Y_N$ which are (i.i.d.)

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

# 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.

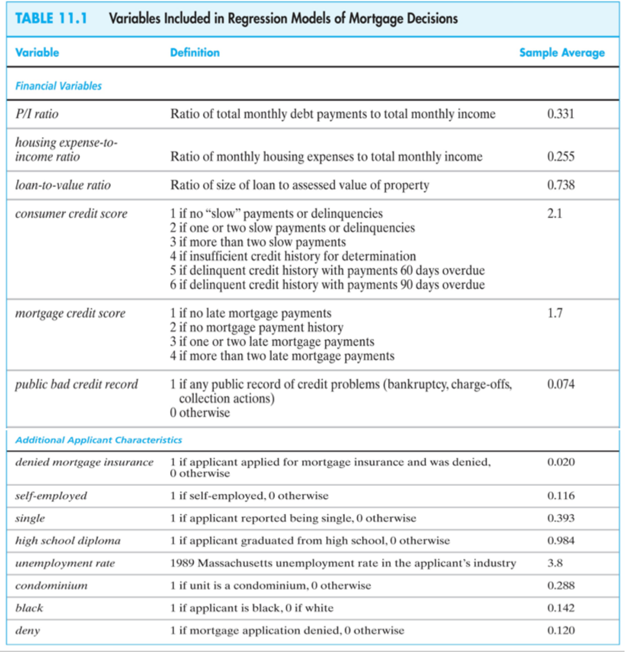


**TABLE 11.1** Variables Included in Regression Models of Mortgage Decisions

| 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<br>2 if one or two slow payments or delinquencies<br>3 if more than two slow payments<br>4 if insufficient credit history for determination<br>5 if delinquent credit history with payments 60 days overdue<br>6 if delinquent credit history with payments 90 days overdue | 2.1 |
| *mortgage credit score* | 1 if no late mortgage payments<br>2 if no mortgage payment history<br>3 if one or two late mortgage payments<br>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)<br>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 $\beta_1$ 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

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

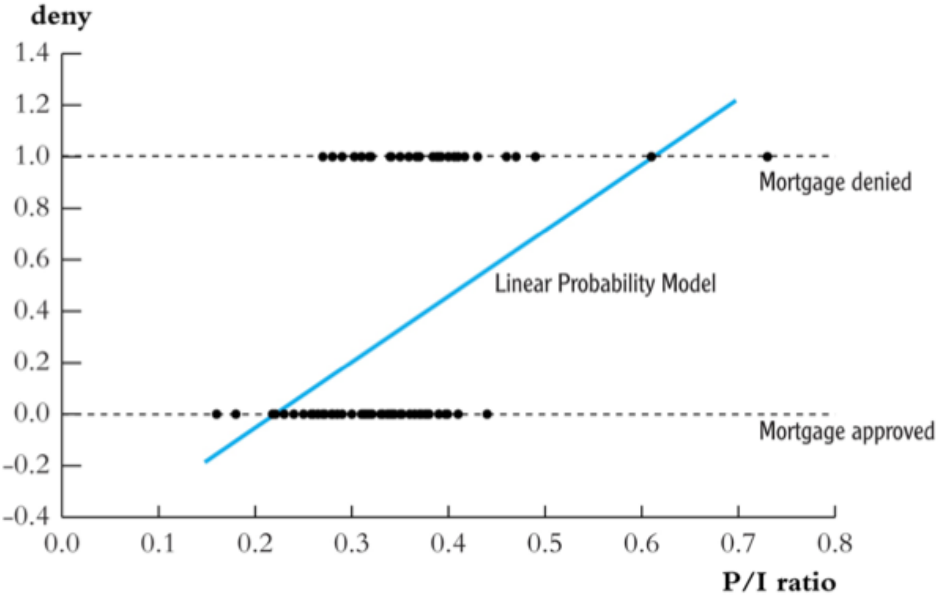
because:

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

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
Linear Probability Model
Mortgage approved

# 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?
- 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 → 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?

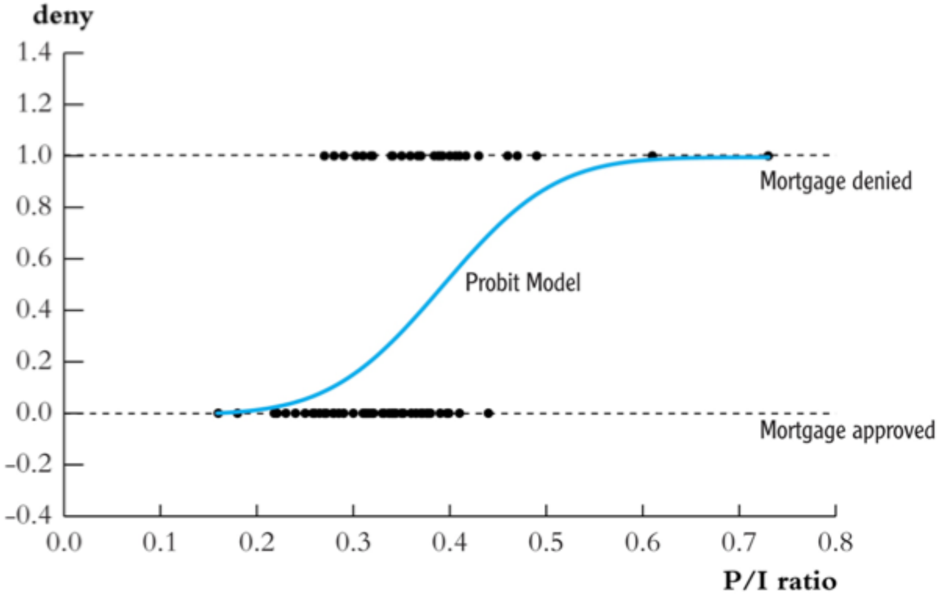
# Moving Away from LPM

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.)
- Want a function $F(z) : (-\infty, \infty) \longrightarrow [0, 1]$.
- 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
- $\mathbb{P}(Y = 1|X)$ is increasing in $X$ if $\beta_1 > 0$.
- $\mathbb{P}(Y = 1|X) \in [0, 1]$ 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.
  - $\beta_1$ is the change in the $z$-value for a change in $X_1$.

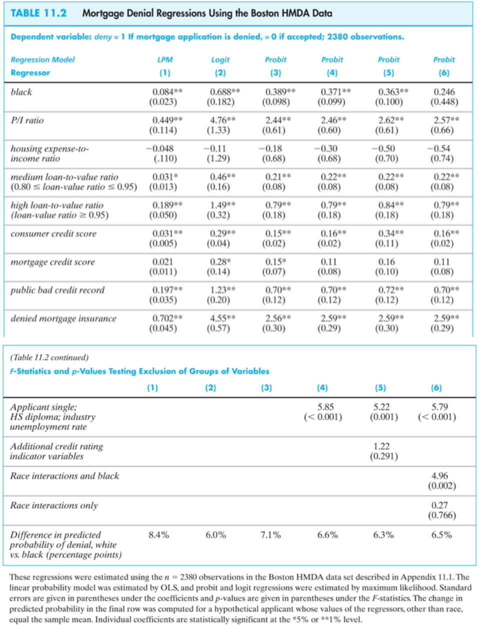


**TABLE 11.2** Mortgage Denial Regressions Using the Boston HMDA Data

**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**<br>(0.023) | 0.688**<br>(0.182) | 0.389**<br>(0.098) | 0.371**<br>(0.099) | 0.363**<br>(0.100) | 0.246<br>(0.448) |
| *P/I ratio* | 0.449**<br>(0.114) | 4.76**<br>(1.33) | 2.44**<br>(0.61) | 2.46**<br>(0.60) | 2.62**<br>(0.61) | 2.57**<br>(0.66) |
| *housing expense-to-income ratio* | −0.048<br>(.110) | −0.11<br>(1.29) | −0.18<br>(0.68) | −0.30<br>(0.68) | −0.50<br>(0.70) | −0.54<br>(0.74) |
| *medium loan-to-value ratio*<br>(0.80 ≤ *loan-value ratio* ≤ 0.95) | 0.031*<br>(0.013) | 0.46**<br>(0.16) | 0.21**<br>(0.08) | 0.22**<br>(0.08) | 0.22**<br>(0.08) | 0.22**<br>(0.08) |
| *high loan-to-value ratio*<br>*(loan-value ratio* ≥ 0.95) | 0.189**<br>(0.050) | 1.49**<br>(0.32) | 0.79**<br>(0.18) | 0.79**<br>(0.18) | 0.84**<br>(0.18) | 0.79**<br>(0.18) |
| *consumer credit score* | 0.031**<br>(0.005) | 0.29**<br>(0.04) | 0.15**<br>(0.02) | 0.16**<br>(0.02) | 0.34**<br>(0.11) | 0.16**<br>(0.02) |
| *mortgage credit score* | 0.021<br>(0.011) | 0.28*<br>(0.14) | 0.15*<br>(0.07) | 0.11<br>(0.08) | 0.16<br>(0.10) | 0.11<br>(0.08) |
| *public bad credit record* | 0.197**<br>(0.035) | 1.23**<br>(0.20) | 0.70**<br>(0.12) | 0.70**<br>(0.12) | 0.72**<br>(0.12) | 0.70**<br>(0.12) |
| *denied mortgage insurance* | 0.702**<br>(0.045) | 4.55**<br>(0.57) | 2.56**<br>(0.30) | 2.59**<br>(0.29) | 2.59**<br>(0.30) | 2.59**<br>(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<br>(< 0.001) | 5.22<br>(0.001) | 5.79<br>(< 0.001) |
| *Additional credit rating indicator variables* | | | | | 1.22<br>(0.291) | |
| *Race interactions and black* | | | | | | 4.96<br>(0.002) |
| *Race interactions only* | | | | | | 0.27<br>(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:

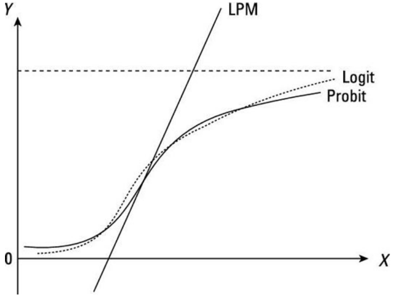
- Gives us more of the ‘S’ shape
- $\mathbb{P}(Y = 1|X)$ is increasing in $X$ if $\beta_1 > 0$.
- $\mathbb{P}(Y = 1|X) \in [0, 1]$ for all $X$
- Easy to compute: $\frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$ has analytic derivatives too.
- Log odds interpretation
  - $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X$
  - $\beta_1$ 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
- 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



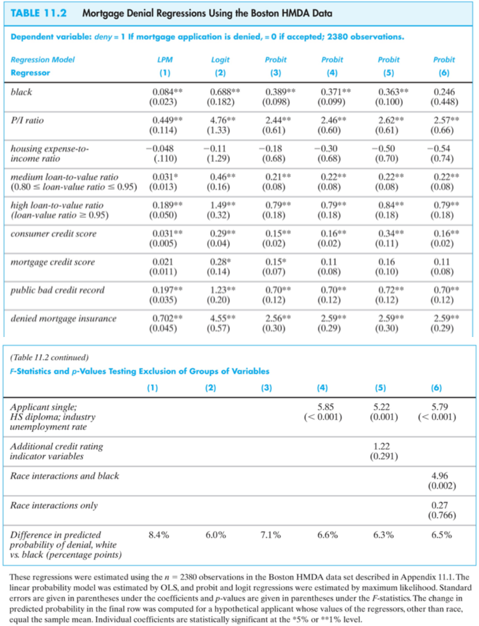


**TABLE 11.2** Mortgage Denial Regressions Using the Boston HMDA Data

**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**<br>(0.023) | 0.688**<br>(0.182) | 0.389**<br>(0.098) | 0.371**<br>(0.099) | 0.363**<br>(0.100) | 0.246<br>(0.448) |
| *P/I ratio* | 0.449**<br>(0.114) | 4.76**<br>(1.33) | 2.44**<br>(0.61) | 2.46**<br>(0.60) | 2.62**<br>(0.61) | 2.57**<br>(0.66) |
| *housing expense-to-income ratio* | −0.048<br>(.110) | −0.11<br>(1.29) | −0.18<br>(0.68) | −0.30<br>(0.68) | −0.50<br>(0.70) | −0.54<br>(0.74) |
| *medium loan-to-value ratio*<br>(0.80 ≤ *loan-value ratio* ≤ 0.95) | 0.031*<br>(0.013) | 0.46**<br>(0.16) | 0.21**<br>(0.08) | 0.22**<br>(0.08) | 0.22**<br>(0.08) | 0.22**<br>(0.08) |
| *high loan-to-value ratio*<br>(*loan-value ratio* ≥ 0.95) | 0.189**<br>(0.050) | 1.49**<br>(0.32) | 0.79**<br>(0.18) | 0.79**<br>(0.18) | 0.84**<br>(0.18) | 0.79**<br>(0.18) |
| *consumer credit score* | 0.031**<br>(0.005) | 0.29**<br>(0.04) | 0.15**<br>(0.02) | 0.16**<br>(0.02) | 0.34**<br>(0.11) | 0.16**<br>(0.02) |
| *mortgage credit score* | 0.021<br>(0.011) | 0.28*<br>(0.14) | 0.15*<br>(0.07) | 0.11<br>(0.08) | 0.16<br>(0.10) | 0.11<br>(0.08) |
| *public bad credit record* | 0.197**<br>(0.035) | 1.23**<br>(0.20) | 0.70**<br>(0.12) | 0.70**<br>(0.12) | 0.72**<br>(0.12) | 0.70**<br>(0.12) |
| *denied mortgage insurance* | 0.702**<br>(0.045) | 4.55**<br>(0.57) | 2.56**<br>(0.30) | 2.59**<br>(0.29) | 2.59**<br>(0.30) | 2.59**<br>(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<br>(< 0.001) | 5.22<br>(0.001) | 5.79<br>(< 0.001) |
| *Additional credit rating indicator variables* | | | | | 1.22<br>(0.291) | |
| *Race interactions and black* | | | | | | 4.96<br>(0.002) |
| *Race interactions only* | | | | | | 0.27<br>(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

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

# FOC of Log-Likelihood

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

# FOC of Log-Likelihood (Logit)

This is the score of the log-likelihood:

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

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

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

This comes from the hazard rate.

# FOC of Log-Likelihood (Probit)

This is the score of the log-likelihood:

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

Using the $q_i = 2y_i - 1$ trick

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

# 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 $\beta$.

# The Hessian Matrix (Logit)

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

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

This is negative semi definite

# The Hessian Matrix (Probit)

Recall

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

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

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

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:

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

- $\mathbf{H}_f(x_n)$ 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 $\left[\mathbf{H}_f(x_n)\right]^{-1}$ 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 $R^2$

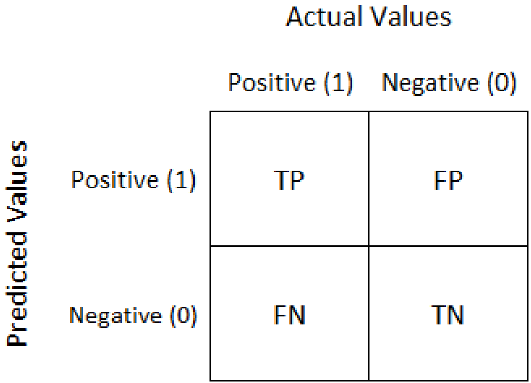
How well does the model fit the data?

- 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

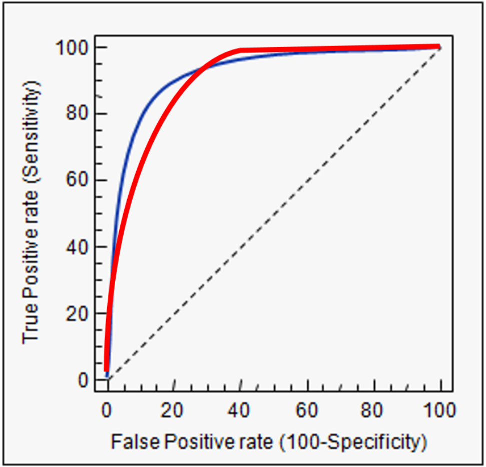
- Machine learning likes to think about this problem more like classification then regression.
- A caution: these are regression models not classification models.
- Predict either $\hat{y}_i = 1$ or $\hat{y}_i = 0$ for each observation.
- Predict $\hat{y}_i = 1$ if $Pr(y_i = 1|X_i = x) \geq 0.5$ or $F(X_i\hat{\beta})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

# ROC Curve/ AOC



| | | 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.
- AOC is area under the curve

# 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.

# Thanks!