

# Estimation and Inference in Probit (and Logit) Models (SW Section 9.3)

Probit model:

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

- Estimation and inference
  - How to estimate  $\beta_0$  and  $\beta_1$ ?
  - What is the sampling distribution of the estimators?
  - Why can we use the usual methods of inference?
- First discuss *nonlinear least squares* (easier to explain)
- Then discuss *maximum likelihood* estimation (what is actually done in practice)

# Probit estimation by nonlinear least squares

Recall OLS:

$$\min_{b_0, b_1} \sum_{i=1}^n [Y_i - (b_0 + b_1 X_i)]^2$$

- The result is the OLS estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$

In probit, we have a different regression function – the nonlinear probit model. So, we could estimate  $\beta_0$  and  $\beta_1$  by *nonlinear least squares*:

$$\min_{b_0, b_1} \sum_{i=1}^n [Y_i - \Phi(b_0 + b_1 X_i)]^2$$

Solving this yields the *nonlinear least squares* estimator of the probit coefficients.

## Nonlinear least squares, ctd.

$$\min_{b_0, b_1} \sum_{i=1}^n [Y_i - \Phi(b_0 + b_1 X_i)]^2$$

How to solve this minimization problem?

- Calculus doesn't give an explicit solution since the first-order conditions will be nonlinear.
- Must be solved *numerically* using the computer, e.g. by “trial and error” method of trying one set of values for  $(b_0, b_1)$ , then trying another, and another,...
- Better idea: use specialized minimization algorithms that search more efficiently than “trial and error.”

In practice, nonlinear least squares isn't used because it isn't efficient – an estimator with a smaller variance is...

## Probit estimation by maximum likelihood

The *likelihood function* is the conditional density of  $Y_1, \dots, Y_n$  given  $X_1, \dots, X_n$ , treated as a function of the unknown parameters  $\beta_0$  and  $\beta_1$ .

- The maximum likelihood estimator (MLE) is the value of  $(\beta_0, \beta_1)$  that maximize the likelihood function.
- The MLE is the value of  $(\beta_0, \beta_1)$  that best describe the full distribution of the data.
- Like the nonlinear-least squares estimator, the MLE of  $\beta_0$  and  $\beta_1$  in the probit and logit models must be found numerically.

- In large samples, the MLE is:
  - Consistent,
  - normally distributed, and
  - efficient (has the smallest variance of all estimators)

# **Application to the Boston HMDA Data**

## **(SW Section 9.4)**

- Mortgages (home loans) are an essential part of buying a home.
- Is there differential access to home loans by race?
- If two otherwise identical individuals, one white and one black, applied for a home loan, is there a difference in the probability of denial?

## The HMDA Data Set

- Data on individual characteristics, property characteristics, and loan denial/acceptance
- The mortgage application process circa 1990-1991:
  - Go to a bank or mortgage company
  - Fill out an application (personal+financial info)
  - Meet with the loan officer
- Then the loan officer decides – by law, in a race-blind way. Presumably, the bank wants to make profitable loans, and the loan officer doesn't want to originate defaults.

# The loan officer's decision

- Loan officer uses key financial variables:
  - *P/I ratio*
  - housing expense-to-income ratio
  - loan-to-value ratio
  - personal credit history
- The decision rule is nonlinear:
  - loan-to-value ratio  $> 80\%$
  - loan-to-value ratio  $> 95\%$  (what happens in default?)
  - credit score

## Regression specifications

$\Pr(\text{deny}=1|\text{black, other } X\text{'s}) = \dots$

- linear probability model
- probit

Main problem with the regressions so far: potential omitted variable bias. All these (i) enter the loan officer decision function, all (ii) are or could be correlated with race:

- wealth, type of employment
- credit history
- family status

Variables in the HMDA data set...

**TABLE 9.1 Variables Included in Regression Models of Mortgage Decisions**

Variable	Definition	Sample Average
<b>Financial Variables</b>		
<i>P/I ratio</i>	Ratio of total monthly debt payments to total monthly income	0.331
<i>housing expense-to-income ratio</i>	Ratio of monthly housing expenses to total monthly income	0.255
<i>loan-to-value ratio</i>	Ratio of size of loan to assessed value of property	0.738
<i>consumer credit score</i>	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
<i>mortgage credit score</i>	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
<i>public bad credit record</i>	1 if any public record of credit problems (bankruptcy, charge-offs, collection actions) 0 otherwise	0.074

**TABLE 9.1 Variables Included in Regression Models of Mortgage Decisions**

Variable	Definition	Sample Average
<b>Additional Applicant Characteristics</b>		
<i>denied mortgage insurance</i>	1 if applicant applied for mortgage insurance and was denied, 0 otherwise	0.020
<i>self-employed</i>	1 if self-employed, 0 otherwise	0.116
<i>single</i>	1 if applicant reported being single, 0 otherwise	0.393
<i>high school diploma</i>	1 if applicant graduated from high school, 0 otherwise	0.984
<i>unemployment rate</i>	1989 Massachusetts unemployment rate in the applicant’s industry	3.8
<i>condominium</i>	1 if unit is a condominium, 0 otherwise	0.288
<i>black</i>	1 if applicant is black, 0 if white	0.142
<i>deny</i>	1 if mortgage application denied, 0 otherwise	0.120

**TABLE 9.2 Mortgage Denial Regressions Using the Boston HMDA Data****Dependent Variable: deny = 1 If Mortgage Application Is Denied, = 0 If Accepted; 2,380 observations.**

<b>Regression Model</b>	<b>LPM</b>	<b>Logit</b>	<b>Probit</b>	<b>Probit</b>	<b>Probit</b>	<b>Probit</b>
<b>Regressor</b>	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>	<b>(4)</b>	<b>(5)</b>	<b>(6)</b>
<i>black</i>	0.084** (0.023)	0.688** (0.182)	0.389** (0.098)	0.371** (0.099)	0.363** (0.100)	0.246 (0.448)
<i>P/I ratio</i>	0.449** (0.114)	4.76** (1.33)	2.44** (0.61)	2.46** (0.60)	2.62** (0.61)	2.57** (0.66)
<i>housing expense-to-income ratio</i>	-0.048 (.110)	-0.11 (1.29)	-0.18 (0.68)	-0.30 (0.68)	-0.50 (0.70)	-0.54 (0.74)
<i>medium loan-to-value ratio</i> (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)
<i>high loan-to-value ratio</i> (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)
<i>consumer credit score</i>	0.031** (0.005)	0.29** (0.04)	0.15** (0.02)	0.16** (0.02)	0.34** (0.11)	0.16** (0.02)
<i>mortgage credit score</i>	0.021 (0.011)	0.28* (0.14)	0.15* (0.07)	0.11 (0.08)	0.16 (0.10)	0.11 (0.08)
<i>public bad credit record</i>	0.197** (0.035)	1.23** (0.20)	0.70** (0.12)	0.70** (0.12)	0.72** (0.12)	0.70** (0.12)
<i>denied mortgage insurance</i>	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 9.2 Mortgage Denial Regressions Using the Boston HMDA Data****Dependent Variable: deny = 1 If Mortgage Application Is Denied, = 0 If Accepted; 2,380 observations.**

<b>Regression Model</b>	<b>LPM</b>	<b>Logit</b>	<b>Probit</b>	<b>Probit</b>	<b>Probit</b>	<b>Probit</b>
<b>Regressor</b>	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>	<b>(4)</b>	<b>(5)</b>	<b>(6)</b>
<i>self-employed</i>	0.060** (0.021)	0.67** (0.21)	0.36** (0.11)	0.35** (0.11)	0.34** (0.11)	0.35** (0.11)
<i>single</i>				0.23** (0.08)	0.23** (0.08)	0.23** (0.08)
<i>high school diploma</i>				-0.61** (0.23)	-0.60* (0.24)	-0.62** (0.23)
<i>unemployment rate</i>				0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
<i>condominium</i>					-0.05 (0.09)	
<i>black × P/I ratio</i>						-0.58 (1.47)
<i>black × housing expense-to-income ratio</i>						1.23 (1.69)
<i>Additional credit rating indicator variables</i>	no	no	no	no	yes	no
<i>constant</i>	-0.183** (0.028)	-5.71** (0.48)	-3.04** (0.23)	-2.57** (0.34)	-2.90** (0.39)	-2.54** (0.35)

(Table 9.2 continued)

(Table 9.2 continued)

**F-statistics and p-values Testing Exclusion of Groups of Variables**

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

These regressions were estimated using the  $n = 2,380$  observations in the Boston HMDA data set described in Appendix 9.1. 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.

## Summary of Empirical Results

- Coefficients on the financial variables make sense.
- *Black* is statistically significant in all specifications
- Race-financial variable interactions aren't significant.
- Including the covariates sharply reduces the effect of race on denial probability.
- LPM, probit, logit: similar estimates of effect of race on the probability of denial.
- Estimated effects are large in a “real world” sense.

# Remaining threats to internal, external validity

- Internal validity

1. omitted variable bias

- what *else* is learned in the in-person interviews?

2. functional form misspecification (no...)

3. measurement error (originally, yes; now, no...)

4. selection

- random sample of loan applications
- define population to be loan applicants

5. simultaneous causality (no)

- External validity

This is for Boston in 1990-91. What about today?

# Summary

## (SW Section 9.5)

- If  $Y_i$  is binary, then  $E(Y|X) = \Pr(Y=1|X)$
- Three models:
  - linear probability model (linear multiple regression)
  - probit (cumulative standard normal distribution)
  - logit (cumulative standard logistic distribution)
- LPM, probit, logit all produce predicted probabilities
- Effect of  $\Delta X$  is change in conditional probability that  $Y=1$ . For logit and probit, this depends on the initial  $X$

- Probit and logit are estimated via maximum likelihood
  - Coefficients are normally distributed for large  $n$
  - Large- $n$  hypothesis testing, conf. intervals is as usual