

Regression with a Binary Dependent Variable (SW Chapter 11)

Outline

1. The Linear Probability Model
2. Probit and Logit Regression
3. Estimation and Inference in Probit and Logit
4. Application to Racial Discrimination in Mortgage Lending

Binary Dependent Variables: What's Different?

So far the dependent variable (Y) has been continuous:

- height
- BMI, weight
- wage

What if Y is binary?

- Y = get into college, or not; X = high school grades, ATAR scores, demographic variables
- Y = person smokes, or not; X = cigarette tax rate, income, demographic variables
- Y = mortgage application is accepted, or not; X = race, income, house characteristics, marital status

Example: Mortgage Denial and Race

The Boston Fed HMDA Dataset

- Individual applications for single-family mortgages made in 1990 in the greater Boston area
- 2380 observations, collected under Home Mortgage Disclosure Act (HMDA)

Variables

- Dependent variable:
 - Is the mortgage denied or accepted?
- Independent variables:
 - income, wealth, employment status
 - other loan, property characteristics
 - race of applicant

Binary Dependent Variables and the Linear Probability Model (SW Section 11.1)

A natural starting point is the linear regression model with a single regressor:

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

But:

- What does β_1 mean when Y is binary? Is $\beta_1 = \frac{\Delta Y}{\Delta X}$?
- What does the line $\beta_0 + \beta_1 X$ mean when Y is binary?
- What does the predicted value \hat{Y} mean when Y is binary? For example, what does $\hat{Y} = 0.26$ mean?

The linear probability model, ctd.

In the linear probability model, the predicted value of Y is interpreted as the predicted probability that $Y=1$, and β_1 is the change in that predicted probability for a unit change in X .

Here's the math:

Linear probability model: $Y_i = \beta_0 + \beta_1 X_i + u_i$

When Y is binary,

$$E(Y|X) = 1 \times \Pr(Y=1|X) + 0 \times \Pr(Y=0|X) = \Pr(Y=1|X)$$

Under LS assumption #1, $E(u_i|X_i) = 0$, so

$$E(Y_i|X_i) = E(\beta_0 + \beta_1 X_i + u_i|X_i) = \beta_0 + \beta_1 X_i,$$

so

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

The linear probability model, ctd.

When Y is binary, the linear regression model

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

is called the ***linear probability model*** because

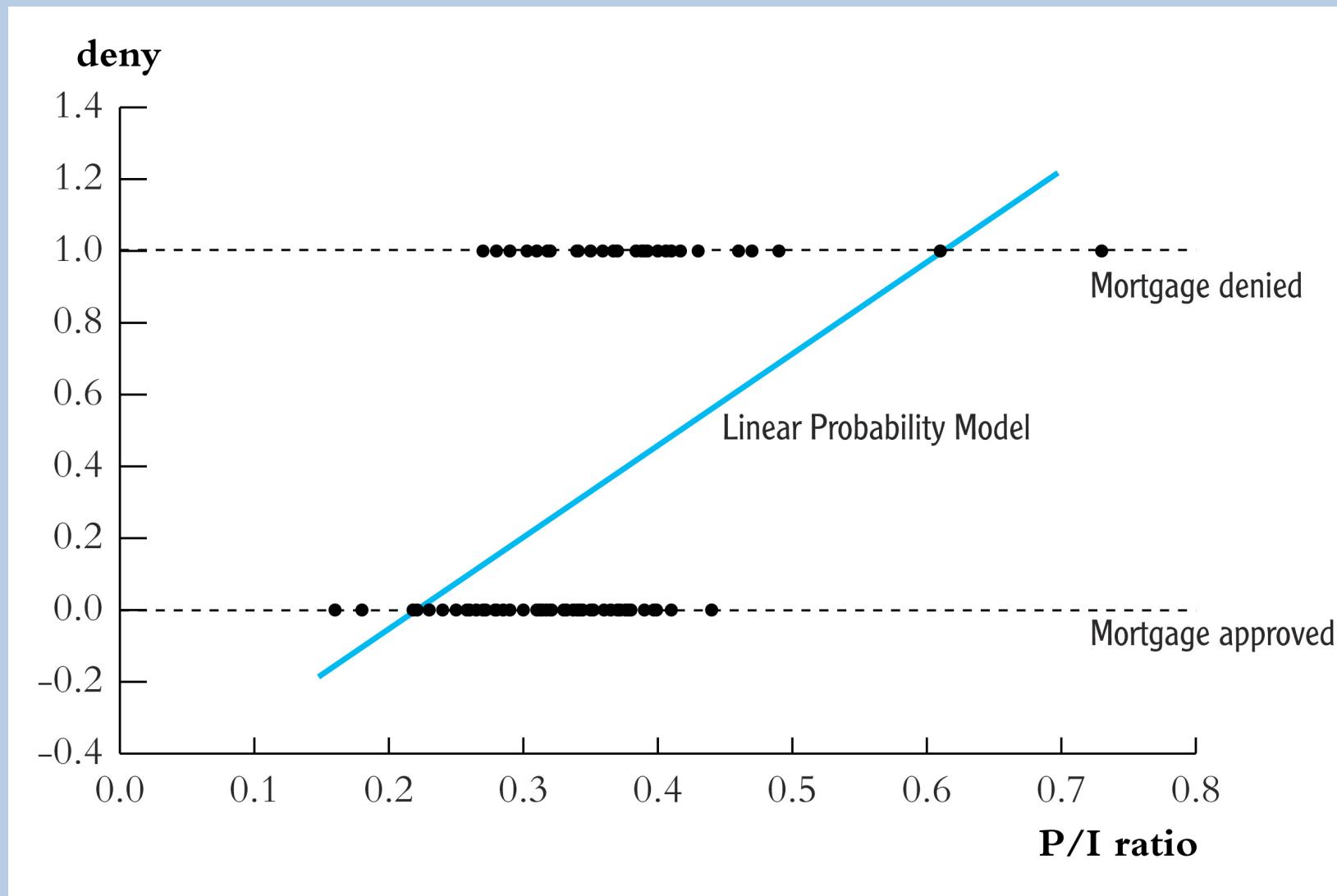
$$\Pr(Y=1|X) = \beta_0 + \beta_1 X_i$$

- The predicted value is a ***probability***:
 - $E(Y|X=x) = \Pr(Y=1|X=x)$ = prob. that $Y=1$ given x
 - \hat{Y} = the ***predicted probability*** that $Y_i=1$, given X
- β_1 = change in probability that $Y=1$ for a unit change in x :

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

Example: linear probability model, HMDA data

Mortgage denial v. ratio of debt payments to income (P/I ratio) in a subset of the HMDA data set ($n = 127$)



Linear probability model: full HMDA data set

$$\widehat{deny} = -.080 + .604P/I\ ratio \quad (n = 2380)$$
$$(.032) (.098)$$

- What is the predicted value for $P/I\ ratio = .3$?

$$\Pr(deny = 1 | P / Iratio = .3) = -.080 + .604 \times .3 = .151$$

- Calculating “effects:” increase $P/I\ ratio$ from .3 to .4:

$$\Pr(deny = 1 | P / Iratio = .4) = -.080 + .604 \times .4 = .212$$

The effect on the probability of denial of an increase in $P/I\ ratio$ from .3 to .4 is to increase the probability by .061, that is, by 6.1 *percentage points* (what?).

Linear probability model: HMDA data, ctd

Next include *black* as a regressor:

$$\widehat{deny} = -.091 + .559P/I\ ratio + .177black$$
$$(.032) \quad (.098) \quad (.025)$$

Predicted probability of denial:

- for black applicant with *P/I ratio* = .3:

$$\widehat{\Pr(deny = 1)} = -.091 + .559 \times .3 + .177 \times 1 = .254$$

- for white applicant, *P/I ratio* = .3:

$$\widehat{\Pr(deny = 1)} = -.091 + .559 \times .3 + .177 \times 0 = .077$$

- difference = .177 = 17.7 percentage points
- Coefficient on *black* is significant at the 5% level
- *Still plenty of room for omitted variable bias...*

The linear probability model: Summary

- The linear probability model models $\Pr(Y=1|X)$ as a linear function of X
- Advantages:
 - simple to estimate and to interpret
 - inference is the same as for multiple regression (need heteroskedasticity-robust standard errors)
- Disadvantages:
 - A LPM says that the change in the predicted probability for a given change in X is the same for all values of X , but that doesn't make sense. Think about the HMDA example...
 - Also, LPM predicted probabilities can be <0 or >1 !
- These disadvantages can be solved by using a *nonlinear* probability model: probit and logit regression

Probit and Logit Regression (SW Section 11.2)

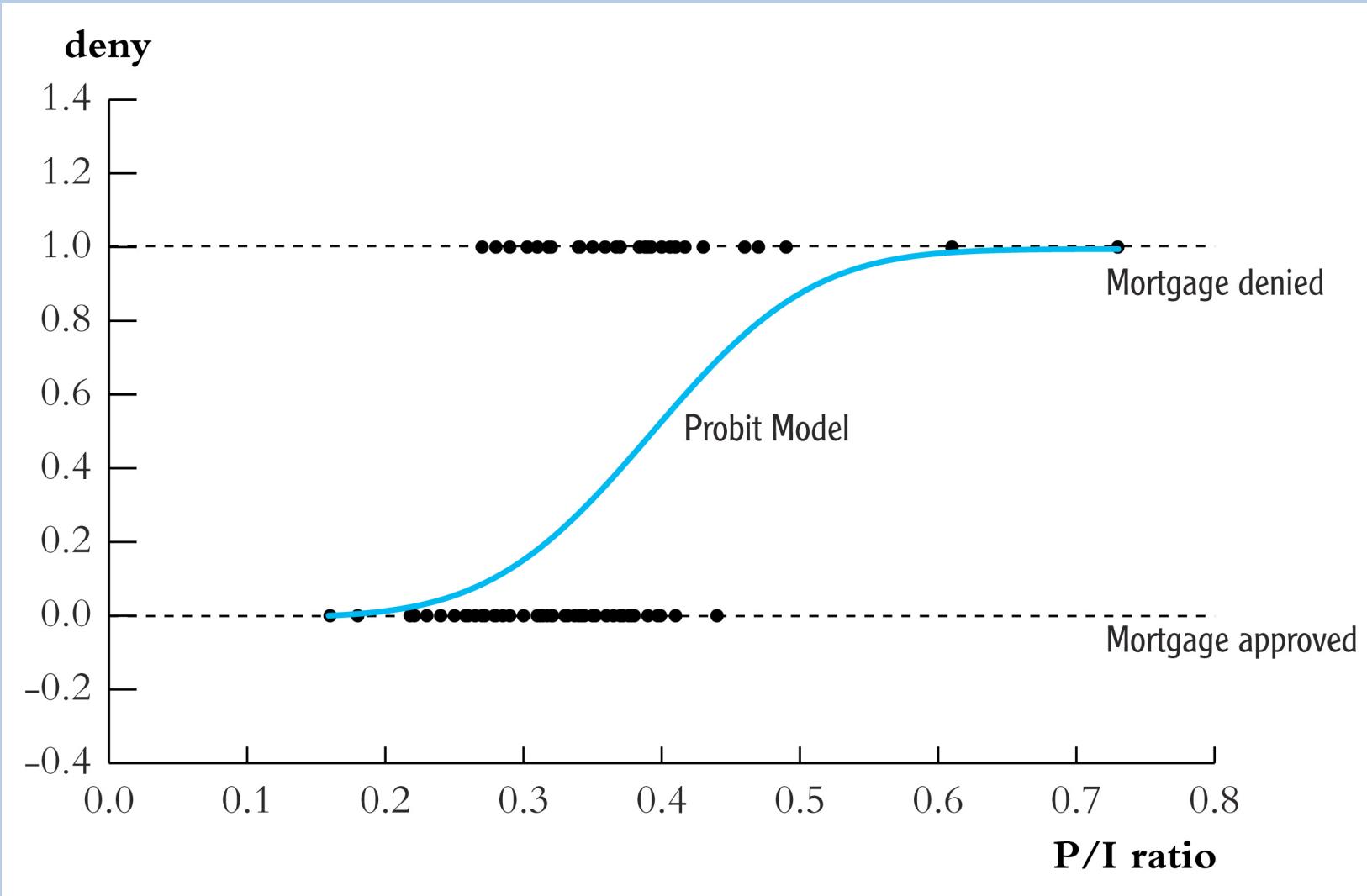
The problem with the linear probability model is that it models the probability of $Y=1$ as being linear:

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

Instead, we want:

- i. $\Pr(Y = 1|X)$ to be increasing in X for $\beta_1 > 0$, and
- ii. $0 \leq \Pr(Y = 1|X) \leq 1$ for all X

This requires using a *nonlinear* functional form for the probability. How about an “S-curve”...



The probit model satisfies these conditions:

- i. $\Pr(Y = 1|X)$ to be increasing in X for $\beta_1 > 0$, and
- ii. $0 \leq \Pr(Y = 1|X) \leq 1$ for all X

Probit regression models the probability that $Y=1$ using the cumulative standard normal distribution function, $\Phi(z)$, evaluated at $z = \beta_0 + \beta_1 X$. The probit regression model is,

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

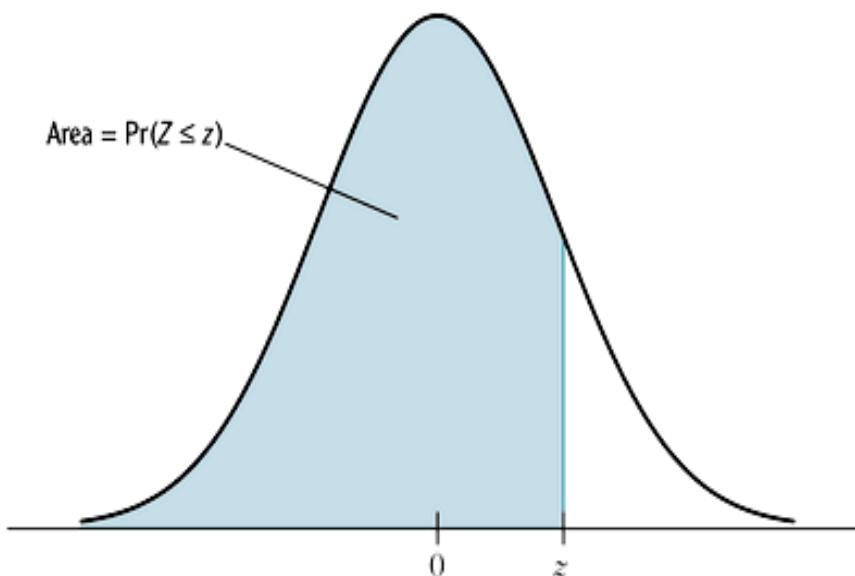
where Φ is the cumulative normal distribution function and $z = \beta_0 + \beta_1 X$ is the “z-value” or “z-index” of the probit model.

Example: Suppose $\beta_0 = -2$, $\beta_1 = 3$, $X = .4$, so

$$\Pr(Y = 1|X=.4) = \Phi(-2 + 3 \times .4) = \Phi(-0.8)$$

$\Pr(Y = 1|X=.4) =$ area under the standard normal density to left of $z = -.8$, which is...

TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z) = \Pr(Z \leq z)$



z	Second Decimal Value of z									
	0	1	2	3	4	5	6	7	8	9
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121

$$\Pr(z \leq -0.8) = .2119$$

Probit regression, ctd.

Why use the cumulative normal probability distribution?

- The “S-shape” gives us what we want:
 - i. $\Pr(Y = 1|X)$ is increasing in X for $\beta_1 > 0$
 - ii. $0 \leq \Pr(Y = 1|X) \leq 1$ for all X
- Easy to use – the probabilities are tabulated in the cumulative normal tables (and also are easily computed using regression software)
- Relatively straightforward interpretation:
 - $\beta_0 + \beta_1 X$ = z -value
 - $\hat{\beta}_0 + \hat{\beta}_1 X$ is the predicted z -value, given X
 - β_1 is the change in the z -value for a unit change in X

STATA Example: HMDA data

```
. probit deny p_irat, r;

Iteration 0:  log likelihood = -872.0853          We'll discuss this later
Iteration 1:  log likelihood = -835.6633
Iteration 2:  log likelihood = -831.80534
Iteration 3:  log likelihood = -831.79234

Probit estimates                                         Number of obs      =     2380
                                                               Wald chi2(1)      =      40.68
                                                               Prob > chi2       =     0.0000
Log likelihood = -831.79234                           Pseudo R2        =     0.0462

-----
```

	Robust				
deny	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
p_irat	2.967908	.4653114	6.38	0.000	2.055914 3.879901
_cons	-2.194159	.1649721	-13.30	0.000	-2.517499 -1.87082

$$\Pr(deny = 1 | P / I ratio) = \Phi(-2.19 + 2.97 \times P/I ratio)$$
$$(.16) \quad (.47)$$

STATA Example: HMDA data, ctd.

$$\Pr(deny = 1 | P / I ratio) = \Phi(-2.19 + 2.97 \times P/I ratio)$$

(.16) (.47)

- Positive coefficient: *Does this make sense?*
- Standard errors have the usual interpretation
- Predicted probabilities:

$$\begin{aligned}\Pr(deny = 1 | P / I ratio = .3) &= \Phi(-2.19 + 2.97 \times .3) \\ &= \Phi(-1.30) = .097\end{aligned}$$

- Effect of change in *P/I ratio* from .3 to .4:

$$\begin{aligned}\Pr(deny = 1 | P / I ratio = .4) &= \Phi(-2.19 + 2.97 \times .4) \\ &= \Phi(-1.00) = .159\end{aligned}$$

Predicted probability of denial rises from .097 to .159

Probit regression with multiple regressors

$$\Pr(Y = 1 | X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

- Φ is the cumulative normal distribution function.
- $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ is the “z-value” or “z-index” of the probit model.
- β_1 is the effect on the z-score of a unit change in X_1 , holding constant X_2

STATA Example: HMDA data

```
. probit deny p_irat black, r;
```

```
Iteration 0:  log likelihood = -872.0853  
Iteration 1:  log likelihood = -800.88504  
Iteration 2:  log likelihood = -797.1478  
Iteration 3:  log likelihood = -797.13604
```

```
Probit estimates                                         Number of obs      =      2380  
                                                               Wald chi2(2)      =     118.18  
                                                               Prob > chi2       =     0.0000  
Log likelihood = -797.13604                           Pseudo R2        =     0.0859
```

deny		Robust				
		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
p_irat	2.741637	.4441633	6.17	0.000	1.871092	3.612181
black	.7081579	.0831877	8.51	0.000	.545113	.8712028
_cons	-2.258738	.1588168	-14.22	0.000	-2.570013	-1.947463

We'll go through the estimation details later...

STATA Example, ctd.: Predicted probit probabilities

```
. probit deny p_irat black, r;
```

Probit estimates

Log likelihood = -797.13604

Number of obs	=	2380
Wald chi2(2)	=	118.18
Prob > chi2	=	0.0000
Pseudo R2	=	0.0859

deny	Robust					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
p_irat	2.741637	.4441633	6.17	0.000	1.871092	3.612181
black	.7081579	.0831877	8.51	0.000	.545113	.8712028
_cons	-2.258738	.1588168	-14.22	0.000	-2.570013	-1.947463

```
. scalar z1 = _b[_cons]+_b[p_irat]*.3+_b[black]*0;
```

```
. display "Pred prob, p_irat=.3, white: " normprob(z1);
```

Pred prob, p_irat=.3, white: .07546603

NOTE

`_b[_cons]` is the estimated intercept (-2.258738)

`_b[p_irat]` is the coefficient on `p_irat` (2.741637)

`scalar` creates a new scalar which is the result of a calculation

`display` prints the indicated information to the screen

STATA Example, ctd.

$$\begin{aligned}\Pr(deny = 1 | P/I, black) \\ &= \Phi(-2.26 + 2.74 \times P/I ratio + .71 \times black) \\ &\quad (.16) \quad (.44) \quad (.08)\end{aligned}$$

- Is the coefficient on *black* statistically significant?
- Estimated effect of race for *P/I ratio* = .3:

$$\Pr(deny = 1 | .3, 1) = \Phi(-2.26 + 2.74 \times .3 + .71 \times 1) = .233$$

$$\Pr(deny = 1 | .3, 0) = \Phi(-2.26 + 2.74 \times .3 + .71 \times 0) = .075$$

- Difference in rejection probabilities = .158 (15.8 percentage points)
- *Still plenty of room for omitted variable bias!*

Logit Regression

Logit regression models the probability of $Y=1$, given X , as the cumulative standard *logistic* distribution function, evaluated at $z = \beta_0 + \beta_1 X$:

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

where F is the cumulative logistic distribution function:

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Because logit and probit use different probability functions, the coefficients (β 's) are different in logit and probit.

Logit regression, ctd.

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

where

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}.$$

Example: $\beta_0 = -3, \beta_1 = 2, X = .4,$

$$\text{so } \beta_0 + \beta_1 X = -3 + 2 \times .4 = -2.2 \text{ so}$$

$$\Pr(Y = 1|X=.4) = 1/(1+e^{-(-2.2)}) = .0998$$

Why bother with logit if we have probit?

- The main reason is historical: logit is computationally faster & easier, but that doesn't matter nowadays
- In practice, logit and probit are very similar – since empirical results typically don't hinge on the logit/probit choice, both tend to be used in practice

STATA Example: HMDA data

```
. logit deny p_irat black, r;
```

```
Iteration 0:  log likelihood = -872.0853      Later...
Iteration 1:  log likelihood = -806.3571
Iteration 2:  log likelihood = -795.74477
Iteration 3:  log likelihood = -795.69521
Iteration 4:  log likelihood = -795.69521
```

```
Logit estimates                                         Number of obs     =      2380
                                                               Wald chi2(2)    =     117.75
                                                               Prob > chi2   =     0.0000
Log likelihood = -795.69521                           Pseudo R2       =     0.0876
```

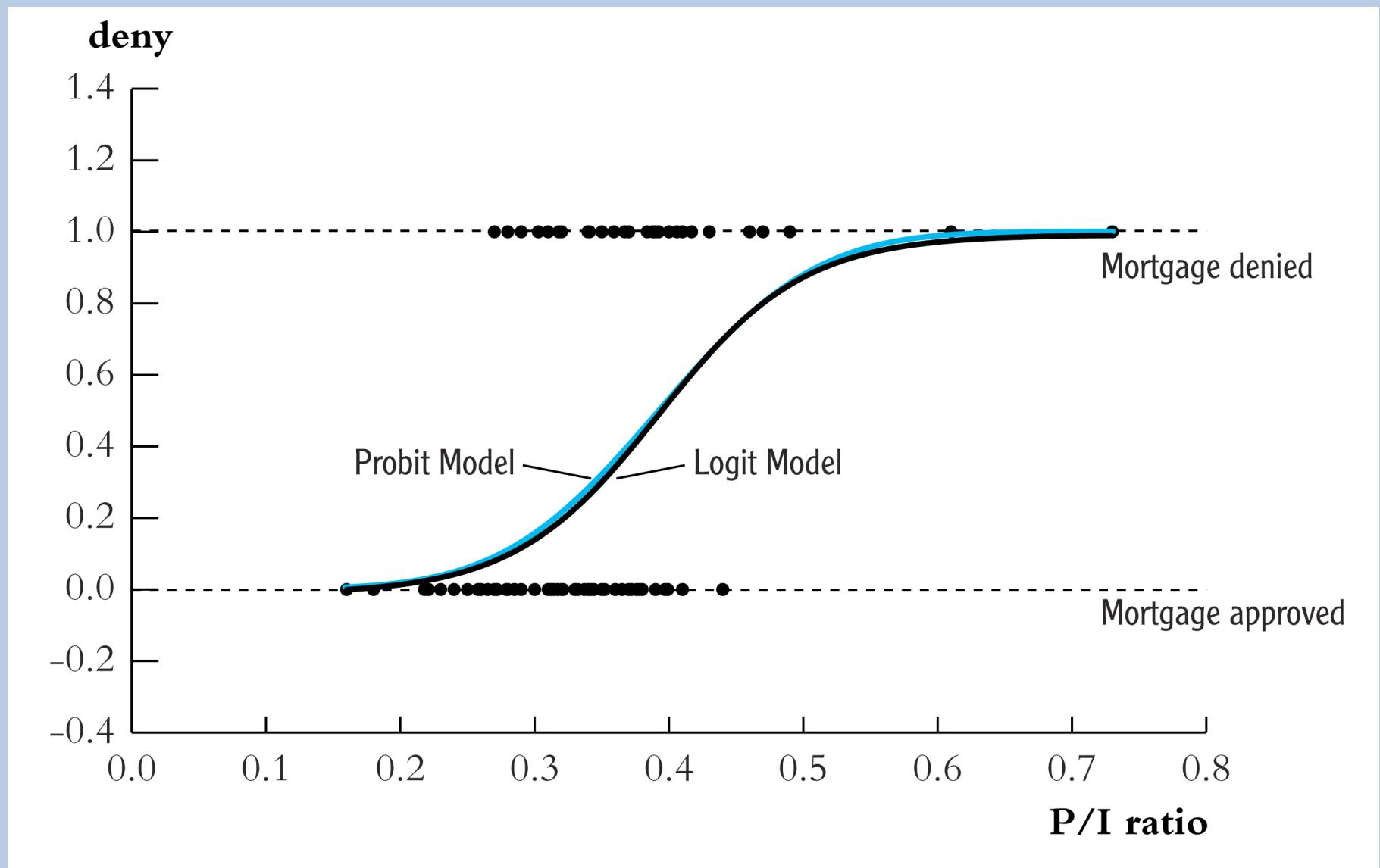
deny	Robust					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
p_irat	5.370362	.9633435	5.57	0.000	3.482244	7.258481
black	1.272782	.1460986	8.71	0.000	.9864339	1.55913
_cons	-4.125558	.345825	-11.93	0.000	-4.803362	-3.447753

```
. dis "Pred prob, p_irat=.3, white: "
>    1/(1+exp(-(_b[_cons]+_b[p_irat]*.3+_b[black]*0)));
```

```
Pred prob, p_irat=.3, white: .07485143
```

NOTE: the probit predicted probability is .07546603

The predicted probabilities from the probit and logit models are very close in these HMDA regressions:



Example for class discussion:

Who are Hezbollah Militants?

Source: Alan Krueger and Jitka Maleckova, “Education, Poverty and Terrorism: Is There a Causal Connection?” *Journal of Economic Perspectives*, Fall 2003, 119-144.

Research questions:

- Does poverty cause terrorism?
- Does low education cause terrorism?

Data set: See the article for details

Binary outcome variable: equal 1 if person died in Hezbollah military event

Table 4
Characteristics of Hezbollah Militants and Lebanese Population of Similar Age

<i>Characteristic</i>	<i>Deceased Hezbollah Militants</i>	<i>Lebanese Population Age 15–38</i>
<i>< Poverty</i>	28%	33%
<i>Education</i>		
Illiterate	0%	6%
Read and write	22%	7%
Primary	17%	23%
Preparatory	14%	26%
Secondary	33%	23%
University	13%	14%
High Studies	1%	1%
<i>Age</i>		
Mean	22.17	25.57
[std.dev.]	(3.99)	(6.78)
15–17	2%	15%
18–20	41%	14%
21–25	42%	23%
26–30	10%	20%
31–38	5%	28%
Hezbollah Education System	21%	NA
<i>Region of Residence</i>		
Beirut	42%	13%
Mount Lebanon	0%	36%
Bekaa	26%	13%
Nabatieh	2%	6%
South	30%	10%
North	0%	22%
<i>Marital Status</i>		
Divorced	1%	NA
Engaged	5%	NA
Married	39%	NA
Single	55%	NA

Notes: Sample size for Lebanese population sample is 120,796. Sample size for Hezbollah is 50 for poverty status, 78 for education, 81 for age (measured at death), 129 for education in Hezbollah system, 116 for region of residence and 75 for marital status.

Table 5

Logistic Estimates of Participation in Hezbollah

(dependent variable is 1 if individual is a deceased Hezbollah militant, and 0 otherwise;
standard errors shown in parentheses)

	All of Lebanon:			Heavily Shiite Regions:	
	Unweighted Estimates		Weighted Estimates	Weighted Estimates	
	(1)	(2)	(3)	(4)	(5)
Intercept	-4.886	-5.910	-5.965	-6.991	-4.658
	(0.365)	(0.391)	(0.230)	(0.255)	(0.232)
Attended Secondary School or Higher (1 = yes)	0.281	0.171	0.281	0.170	0.220
	(0.191)	(0.193)	(0.159)	(0.164)	(0.159)
Poverty (1 = yes)	-0.335	-0.167	-0.335	-0.167	-0.467
	(0.221)	(0.223)	(0.158)	(0.162)	(0.159)
Age	-0.083	-0.083	-0.083	-0.083	-0.083
	(0.015)	(0.015)	(0.008)	(0.008)	(0.008)
Beirut (1 = yes)	—	2.199	—	2.200	—
		(0.219)		(0.209)	
South Lebanon (1 = yes)	—	2.187	—	2.187	—
		(0.232)		(0.221)	
Pseudo R-Square	0.020	0.091	0.018	0.080	0.021
Sample Size	120,925	120,925	120,925	120,925	34,826
					34,826

Notes: Sample pools together observations on 129 deceased Hezbollah fighters and the general Lebanese population from 1996 PHS. Weights used in columns 3 and 4 are the relative share of Hezbollah militants in the population to their share in the sample and relative share of PHS respondents in the sample to their share in the population. Weight is 0.273 for Hezbollah sample and .093 for PHS sample.

Hezbollah militants example, ctd.

Compute the effect of **schooling** by comparing predicted probabilities using the logit regression in column (3):

$$\Pr(Y=1|\text{secondary} = 1, \text{poverty} = 0, \text{age} = 20)$$

$$- \Pr(Y=0|\text{secondary} = 0, \text{poverty} = 0, \text{age} = 20):$$

$$\Pr(Y=1|\text{secondary} = 1, \text{poverty} = 0, \text{age} = 20)$$

$$= 1/[1+e^{(-5.965+.281\times 1 - .335\times 0 - .083\times 20)}]$$

$$= 1/[1 + e^{7.344}] = .000646 \text{ } Does this make sense? \\$$

$$\Pr(Y=1|\text{secondary} = 0, \text{poverty} = 0, \text{age} = 20)$$

$$= 1/[1+e^{(-5.965+.281\times 0 - .335\times 0 - .083\times 20)}]$$

$$= 1/[1 + e^{7.625}] = .000488 \text{ } Does this make sense? \\$$

Predicted change in probabilities:

$$\Pr(Y=1 \mid \text{secondary} = 1, \text{poverty} = 0, \text{age} = 20)$$

$$\begin{aligned} & - \Pr(Y=1 \mid \text{secondary} = 1, \text{poverty} = 0, \text{age} = 20) \\ & = .000646 - .000488 = .000158 \end{aligned}$$

Both these statements are true:

- The probability of being a Hezbollah militant increases by 0.0158 percentage points, if secondary school is attended.
- The probability of being a Hezbollah militant increases by 32%, if secondary school is attended ($.000158/.000488 = .32$).

These sound so different! What is going on?

Hezbollah militants example, ctd.

Compute the effect of **poverty** by comparing predicted probabilities using the logit regression in column (3):

$$\Pr(Y=1|\text{secondary} = 1, \text{poverty} = 1, \text{age} = 20)$$

$$- \Pr(Y=0|\text{secondary} = 1, \text{poverty} = 0, \text{age} = 20):$$

$$\Pr(Y=1|\text{secondary} = 1, \text{poverty} = 1, \text{age} = 20)$$

$$= 1/[1+e^{(-5.965+.281\times 1 - .335\times 1 - .083\times 20)}]$$

$$= 1/[1 + e^{7.344}] = .000462 \text{ } Does this make sense? \\$$

$$\Pr(Y=1|\text{secondary} = 1, \text{poverty} = 0, \text{age} = 20)$$

$$= 1/[1+e^{(-5.965+.281\times 0 - .335\times 0 - .083\times 20)}]$$

$$= 1/[1 + e^{7.625}] = .000646 \text{ } Does this make sense? \\$$

Predicted change in probabilities:

$$\Pr(Y=1 \mid \text{secondary} = 1, \text{poverty} = 0, \text{age} = 20)$$

$$\begin{aligned} & - \Pr(Y=1 \mid \text{secondary} = 1, \text{poverty} = 0, \text{age} = 20) \\ & = .000462 - .000646 = -.000184 \end{aligned}$$

Both these statements are true:

- The probability of being a Hezbollah militant increases by 0.0158 percentage points, if secondary school is attended.
- The probability of being a Hezbollah militant decreases by 28%, if secondary school is attended ($-.000184/.000646=.28$).

Measures of Fit for Logit and Probit

The R^2 and \bar{R}^2 don't make sense here (*why?*). So, two other specialized measures are used:

1. The *fraction correctly predicted* = fraction of Y 's for which the predicted probability is $>50\%$ when $Y_i=1$, or is $<50\%$ when $Y_i=0$.
2. The *pseudo- R^2* measures the improvement in the value of the log likelihood, relative to having no X 's (see SW App. 11.2). The pseudo- R^2 simplifies to the R^2 in the linear model with normally distributed errors.

Application to the Boston HMDA Data (SW Section 11.4)

- 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 (if the incentives inside the bank or loan origination office are right – a big if during the mid-2000s housing bubble!) 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

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

- linear probability model
- probit

Main problem with the regressions so far: potential omitted variable bias. The following variables (i) enter the loan officer decision *and* (ii) are or could be correlated with race:

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

Fortunately, the HMDA data set is very rich...

TABLE 11.1 Variables Included in Regression Models of Mortgage Decisions

Variable	Definition	Sample Average
Financial Variables		
<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

Additional Applicant Characteristics

<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 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)
<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 ≤ <i>loan-value ratio</i> ≤ 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> (<i>loan-value ratio</i> ≥ 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 11.2, ctd.

<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 11.2 continued)

Table 11.2, ctd.

(Table 11.2 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 $n = 2380$ 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.

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?
2. Wrong functional form?
3. Errors-in-variables bias?
4. Sample selection bias?
5. Simultaneous causality bias?

What do you think?

External validity

These data are from Boston in 1990-91. Do you think the results also apply today, where *you* live?

Conclusion (SW Section 11.5)

- 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 ΔX 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