Economics

Lecture #7

Regression with a Limited Dependent Variable

Outline (today and Wednesday)

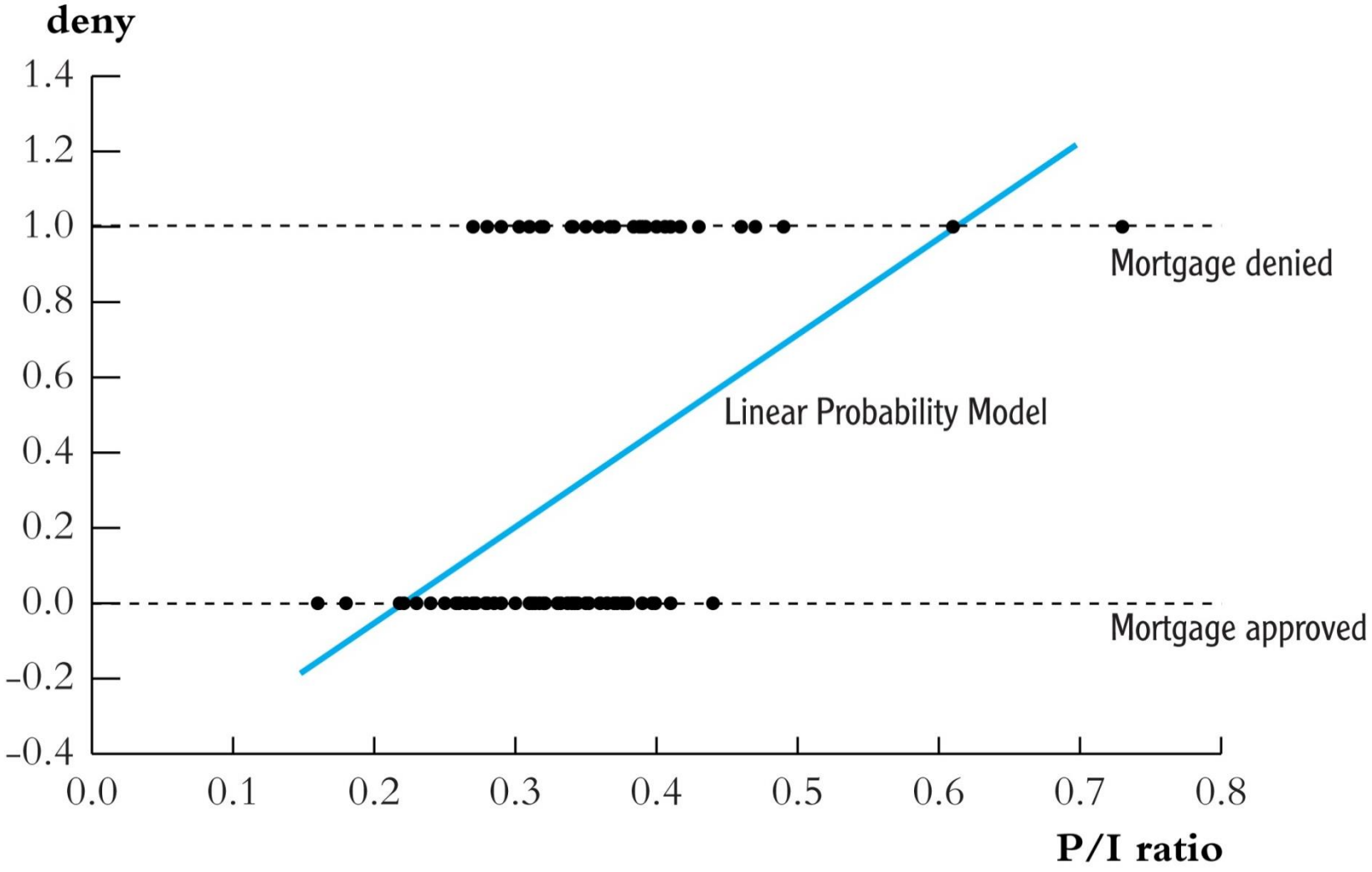
1. Causal inference and regression (wrap up)
2. Regression with a limited dependent variable
    a) Binary and other limited dependent variables
    b) The linear probability model
    c) Probit and logit regression
    d) Estimation and inference: maximum likelihood
    e) Ordered categorical data (ordered probit)

Binary Dependent Variable Extended Example: the HMDA Data Set

- Data on individual characteristics, property characteristics, and loan denial/acceptance; $N = 2380$
- 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.

Binary dependent variables

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



Linear probability model: HMDA data, ctd.

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

- Predicted value for *P/I ratio* = .3?

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

- Calculating "effects:" increase *P/I ratio* from .3 to .4:

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

Linear probability model: HMDA data, ctd

Next include *black* as a regressor:

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

Predicted probability of denial:

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

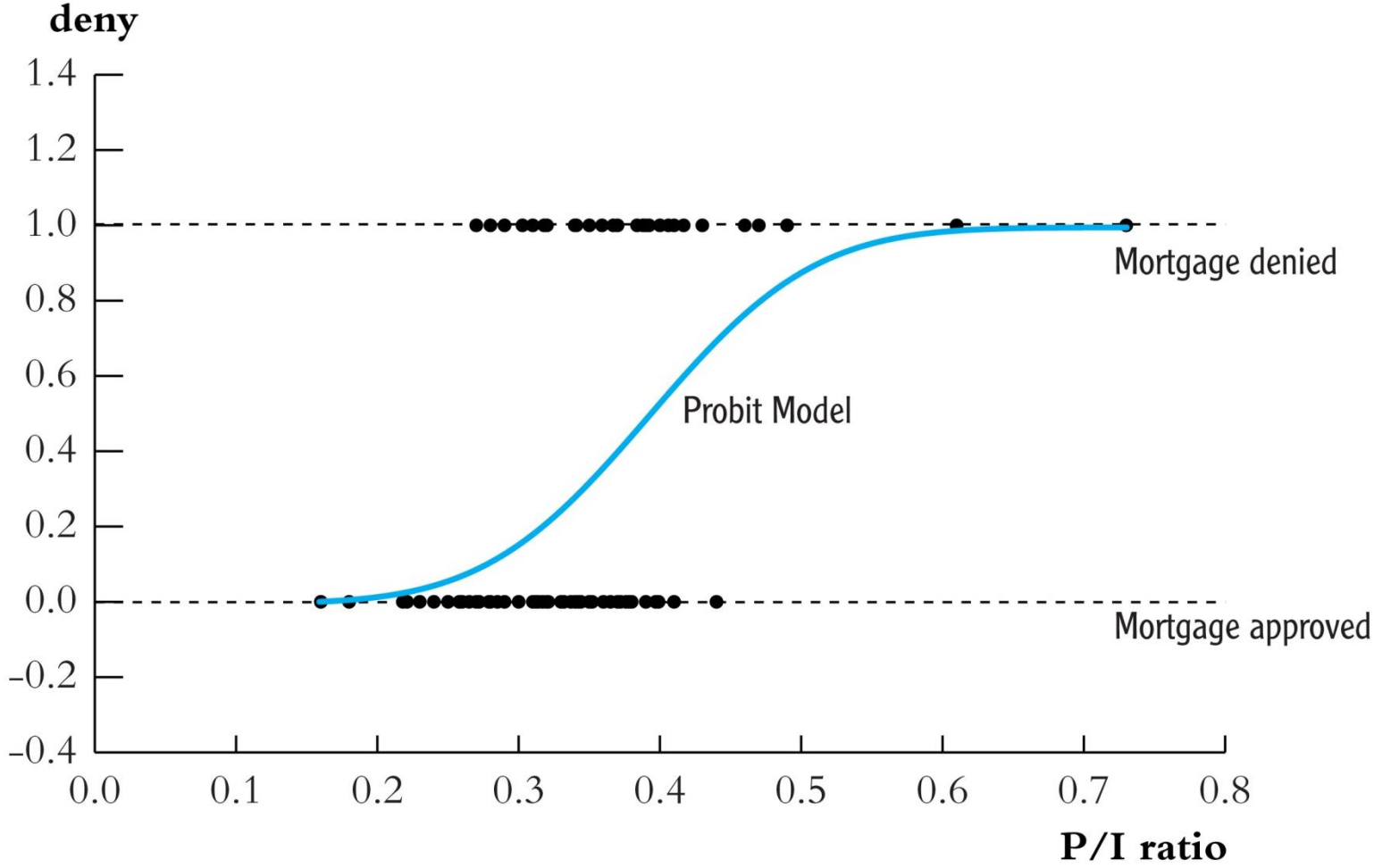
$$\Pr(deny = 1) = -.091 + .559 \times .3 + .177 \times 1 = .254$$

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

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

- difference = .177 = 17.7 **percentage points**
- Coefficient on *black* is significant at the 5% level

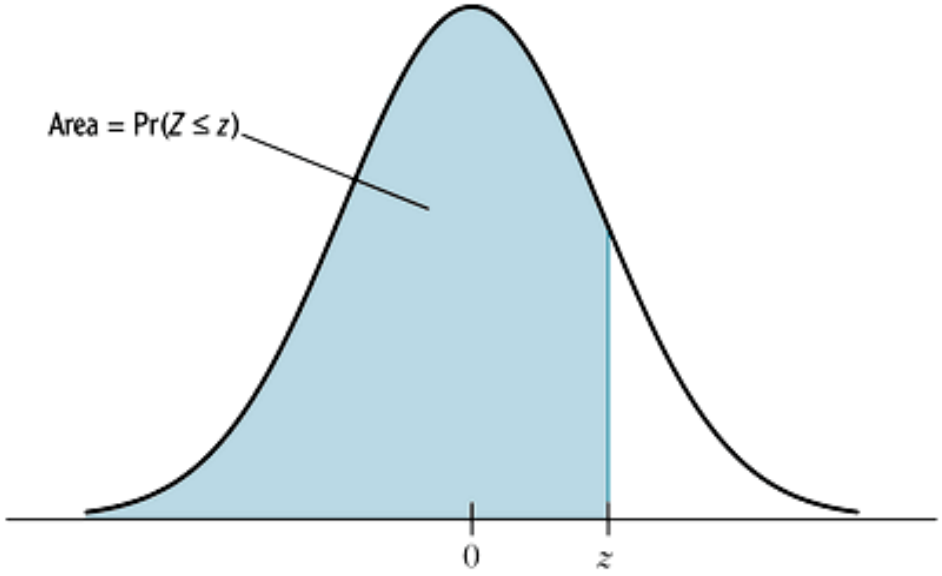
Probit Regression Model – HMDA data



The probit model satisfies these conditions:

- $0 \leq \Pr(Y = 1|X) \leq 1$ for all X
- $\Pr(Y = 1|X)$ is increasing in X (for $\beta_1 > 0$)

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



	Second Decimal Value of z									
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

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

STATA Example: HMDA data - Probit

```
. 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
------------------------------------------------------------------------------
```

$$\Pr(deny = 1 \mid P/Iratio) = \Phi(\underset{(.16)}{-2.19} + \underset{(.47)}{2.97} \times P/I\ ratio)$$

STATA Example: HMDA data, ctd.

$$\Pr(deny = 1 \mid P/Iratio) = \Phi(\underset{(.16)}{-2.19} + \underset{(.47)}{2.97} \times P/I\ ratio)$$

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

$$\Pr(deny = 1 \mid P/Iratio = .3) = \Phi(-2.19 + 2.97 \times .3)$$

$$= \Phi(-1.30) = .097$$

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

$$\Pr(deny = 1 \mid P/Iratio = .4) = \Phi(-2.19 + 2.97 \times .4)$$

$$= .159$$

Predicted probability of denial rises from .097 to .159

HMDA data – Probit with *P-I ratio* and *black*

Probit with multiple regressors

```
. 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

------------------------------------------------------------------------------
             |               Robust
        deny |      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                                  Number of obs   =       2380
                                                  Wald chi2(2)    =     118.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -797.13604                       Pseudo R2       =     0.0859

------------------------------------------------------------------------------
             |               Robust
        deny |      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
normprob(z1) computes the cumulative normal probability ≤ z1

STATA Example, ctd.

$$\Pr(deny = 1 \mid P/I, black)$$

$$= \Phi(\underset{(.16)}{-2.26} + \underset{(.44)}{2.74} \times P/I\ ratio + \underset{(.08)}{.71} \times black)$$

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

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

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

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

STATA Example: HMDA data – Logit regression

```
. logit deny p_irat black, r;

Iteration 0:   log likelihood =  -872.0853      Later…
Iteration 1:   log likelihood =  -806.3571…

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

------------------------------------------------------------------------------
             |               Robust
        deny |      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
```

Predicted probabilities from estimated probit and logit models usually are very close.

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%
  - credit score
- Illegal to use "protected class" information (gender, race…)

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

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

Table 11.2, ctd.

self-employed	0.060** (0.021)	0.67** (0.21)	0.36** (0.11)	0.35** (0.11)	0.34** (0.11)	0.35** (0.11)
single				0.23** (0.08)	0.23** (0.08)	0.23** (0.08)
high school diploma				−0.61** (0.23)	−0.60* (0.24)	−0.62** (0.23)
unemployment rate				0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
condominium					−0.05 (0.09)	
black × P/I ratio						−0.58 (1.47)
black × housing expense-to-income ratio						1.23 (1.69)
Additional credit rating indicator variables	no	no	no	no	yes	no
constant	−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)
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.