ECON 7310 Elements of Econometrics
Week 9: Regression with a Binary Dependent Variable

David Du[1]

[1] University of Queensland

Draft

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:
  - district-wide average test score
  - traffic fatality rate
- What if Y is binary?
  - $Y =$ get into college, or not;
    $X =$ high school grades, SAT 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 include:
  - Dependent variable: Is the mortgage denied or accepted?
  - Independent variables: income, wealth, employment status, other characteristics of applicant like race.

Binary Dependent Variables and the Linear Probability Model

SW Section 11.1

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

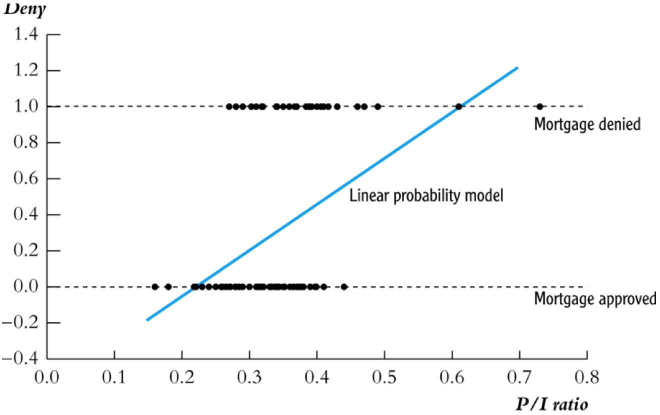
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- But:
  - What does β_1 mean when Y is binary? Is $\beta_1 = \frac{Y}{X}$?
  - What does the line $\beta_0 + \beta_1 X$ mean when Y is binary?
  - What does the predicted value $\widehat{Y}$ mean when Y is binary? Ex: $\widehat{Y} = 0.26$?
- When Y is binary, we have

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

- This is because $E(Y|X) = 1 \times Pr(Y = 1|X) + 0 \times Pr(Y = 0|X)$. And, LS assumption #1, $E(u|X) = 0 \Rightarrow E(Y|X) = E(\beta_0 + \beta_1 X + u|X) = \beta_0 + \beta_1 X$
- So, $\widehat{Y}$ is **the predicted probability** that $Y = 1$, given X
  β_1 is the change in probability that $Y = 1$ for a unit change in X:

Example: linear probability model, HMDA data
Mortgage denial vs. 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

$$\underset{(0.032)}{Deny = -0.080+} \quad \underset{(0.098)}{0.604} \times (PIratio), \qquad n = 2,380$$

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

$$Pr(deny = 1|PIratio = .3) = -.080 + .604 \times .3 = .101$$

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

$$Pr(deny = 1|PIratio = .4) = -.080 + .604 \times .4 = .161$$

- The effect on the probability of denial of an increase in P/I ratio from .3 to .4 is to increase the probability by .06, that is, by 6.0 percentage points.

Linear probability model: full HMDA data set

Next include black as a regressor:

$$\underset{}{Deny} = \underset{(0.032)}{-0.091+} \; \underset{(0.098)}{0.559 \times (PIratio)+} \; \underset{(0.025)}{0.177 \times Black,} \qquad n = 2,380$$

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
- 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 (heteroskedasticity robust SE!!)
- 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.
  - 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:
  1. $Pr(Y = 1|X)$ to be increasing in X for $\beta_1 > 0$, and
  2. $0 \leq Pr(Y = 1|X) \leq 1$ for all X
- This requires using a nonlinear functional form for the probability.
- The probit model and logit model always satisfy these conditions:

Probit Regression

- Probit model considers the structure

$$z_i = \beta_0 + \beta_1 X_i + u_i,$$
$$u_1, \ldots, u_n \overset{iid}{\sim} \mathcal{N}(0,1)$$

where we do not observe z_i but we observe

$$Y_i = \begin{cases} 1 \text{ if } z_i \geq 0 \\ 0 \text{ if } z_i < 0 \end{cases}$$

- Then, a simple algebra shows that

$$\Pr(Y_i = 1 | X_i) = \Phi(\beta_0 + \beta_1 X_i),$$

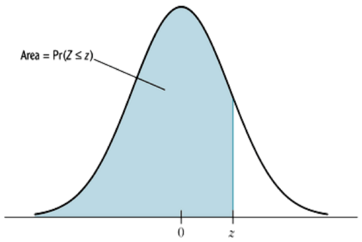
where Φ is the CDF of $\mathcal{N}(0,1)$.

- For example, if $\beta_0 = -2$, $\beta_1 = 3$, and $X_i = 0.4$,

$$\Pr(Y_i = 1 | X_i = 0.4) = \Phi(-2 + 3 \times 0.4) = \Phi(-0.8)$$

Probit Regression, continued

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

So, $\Pr(Y_i = 1 | X_i = 0.4) = \Phi(-0.8) = 0.2119$.

Probit Regression, continued

- The “S-shape” gives us what we want:
  1. $Pr(Y = 1|X)$ to be increasing in X for $\beta_1 > 0$, and
  2. $0 \leq Pr(Y = 1|X) \leq 1$ 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
  - $\widehat{\beta}_0 + \widehat{\beta}_1 X$ 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
------------------------------------------------------------------------------
```

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

Stata example: HMDA data

- Positive coefficient β_1: Does this make sense?
- Standard errors have the usual interpretation
- Predicted probabilities:

$$\Pr(\textit{deny} = 1 | \textit{PIratio} = 0.3) = \Phi(-2.19 + 2.97 \times 0.3) = \Phi(-1.30) = .097$$

- Effect of change in P/I ratio from 0.3 to 0.4:

$$\Pr(\textit{deny} = 1 | \textit{PIratio} = 0.4) = \Phi(-2.19 + 2.97 \times 0.4) = \Phi(-1.00) = .159$$

- Predicted probability of denial rises from .097 to .159

Probit regression with multiple regressors

- A slight extension gives

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

  where Φ the cumulative normal distribution function.
- $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ is the z-value (or z-index, z-score) 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: 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
------------------------------------------------------------------------------
.  sca 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)
sca creates a new scalar which is the result of a calculation
display prints the indicated information to the screen
```

STATA Example, continued

- Is the coefficient on black statistically significant?
- Estimated effect of race for P/I ratio = 0.3:

$$\Pr(deny = 1|0.3, 1) = \Phi(-2.26 + 2.74 \times 0.3 + 0.71 \times 1) = 0.233$$
$$\Pr(deny = 1|0.3, 0) = \Phi(-2.26 + 2.74 \times 0.3 + 0.71 \times 0) = 0.075$$

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

Logit regression

- Logit model is the same as Probit model except that it uses the CDF of logistic distribution, i.e.,

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X))}$$

- Because logit and probit use different probability functions, the coefficients (β's) are different in logit and probit, but predicted probabilities are often very similar.
- Why bother with logit if we have probit?
  - The main reason is historical: logit is computationally faster & easier
  - In practice, logit and probit are very similar – since empirical results typically do not hinge on the logit/probit choice, both tend to be used in practice
  - So, we use probit or logit depending on which method is easiest to use in the software package at hand (both are easy in Stata)

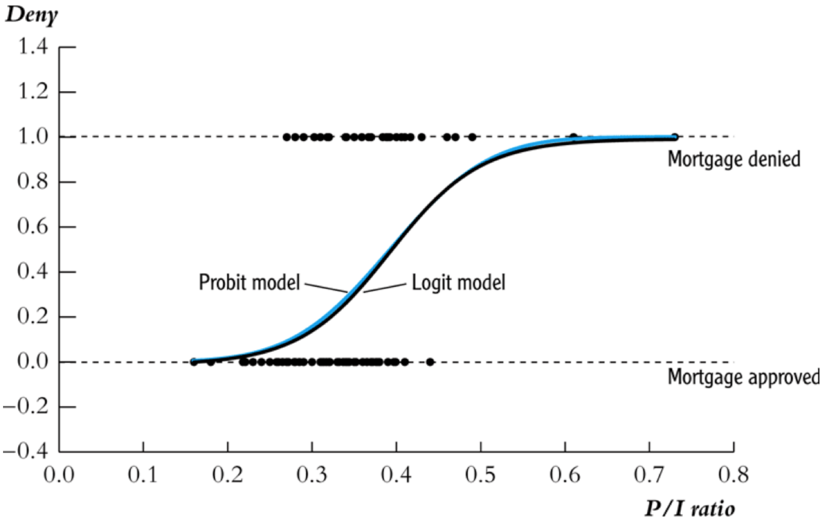
STATA Example: logit

```
. 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
------------------------------------------------------------------------------
             |               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
```

Comparison: probit vs logit

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



Estimation

- We obtain maximum likelihood estimator (MLE) for logit/probit .
- For probit, we have

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$
$$\Pr(Y = 0|X) = 1 - \Phi(\beta_0 + \beta_1 X)$$

- Then, the probability mass function (PMF) of Y can be written as

$$\Phi(\beta_0 + \beta_1 X)^Y (1 - \Phi(\beta_0 + \beta_1 X))^{1-Y}$$

- When $(Y_1, X_1), \ldots, (Y_n, X_n)$ are independently distributed, the joint PMF of $(Y_1, \ldots, Y_n)$ conditional on $(X_1, \ldots, X_n)$ is

$$\prod_{i=1}^{n} \Phi(\beta_0 + \beta_1 X_i)^{Y_i} (1 - \Phi(\beta_0 + \beta_1 X_i))^{1-Y_i}$$

which can be viewed as the likelihood function of (β_0, β_1).

Estimation: MLE

- Idea is to find (β_0, β_1) under which $(Y_1, X_1), \ldots, (Y_n, X_n)$ is the most likely. That is, the MLE $(\widehat{\beta}_0^{ML}, \widehat{\beta}_1^{ML})$ solves

$$\max_{\beta_0, \beta_1} \prod_{i=1}^{n} \Phi(\beta_0 + \beta_1 X_i)^{Y_i} (1 - \Phi(\beta_0 + \beta_1 X_i))^{1 - Y_i}$$

- We cannot solve this problem by hand. Use Stata or other softwares.
- It turns out that
  - $(\widehat{\beta}_0^{ML}, \widehat{\beta}_1^{ML})$ are consistent and asymptotically normally distributed.
  - Using the asymptotic distribution, we can construct the standard errors.
  - Testing and confidence intervals proceed as usual
- Logit is the same but uses the logit CDF F instead of Φ

Measure of Fit

- Usual R^2 and $\bar{R}^2$ do not make sense here. We use two alternative measures of fit.
- The **fraction correctly predicted**:
  - For each $i = 1, \ldots, n$, let $I_i = 1$ if $\Phi(\widehat{\beta}_0^{ML} + \widehat{\beta}_1^{ML} X_i) \geq 0.5$ for $Y_i = 1$ or $\Phi(\widehat{\beta}_0^{ML} + \widehat{\beta}_1^{ML} X_i) < 0.5$ for $Y_i = 0$. Then, the fraction correctly predicted = $\sum_i I_i / n$.
  - Drawback: both 0.51 and 0.99 are counted in the same way. (quality of prediction)
- The **pseudo** R^2: uses the maximized (log) likelihood taking into account the number of regressors.
  - The pseudo-R^2 measures the quality of fit of a probit/logit model by comparing values of the maximized likelihood function with all the regressors to the value of the likelihood with none.
  - Pseudo-$R^2 = 1 - \frac{LL_{ur}}{LL_0}$.

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?
- 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

The Loan Officer's Decision

- 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.
- 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

- Estimate $Pr(deny = 1|black, otherXs)$ by 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		
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** (0.023)	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 ($0.80 \leq$ *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)

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*)

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
applicant single; high school 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.

Summary of Empirical Results

- Coefficients on the financial variables make sense.
- Black is statistically significant in all specifications
- Race-financial variable interactions are not 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.
- Finally, we should carefully think about ‘internal validity’ and ‘external validity’ of the empirical findings.

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