General Linear Models

Logistic Regression

– O.Örsan Özener

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

- 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

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*

$$Yi = \beta_0 + \beta_1 Xi + ui$$

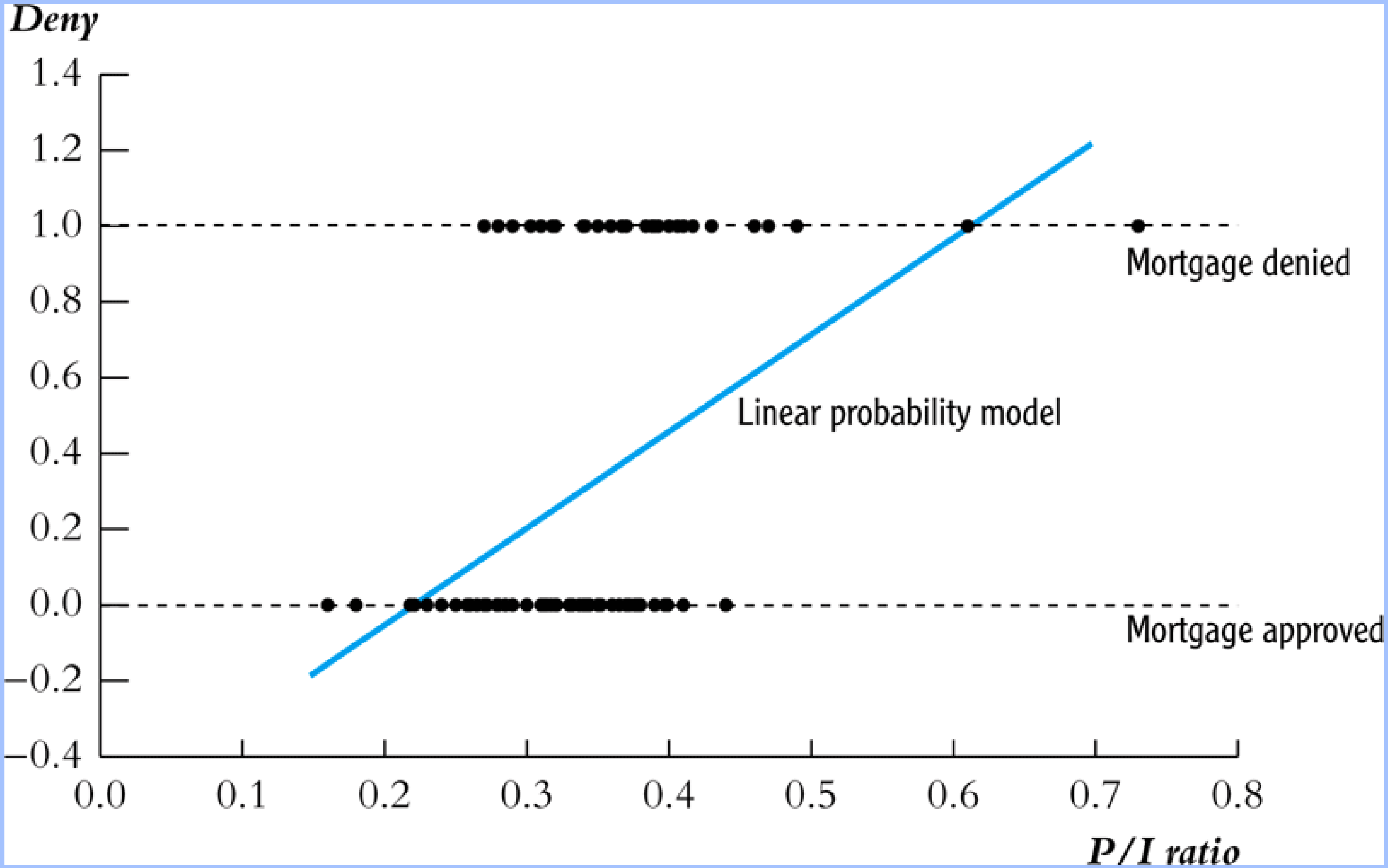
is called the ***linear probability model*** because

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

- 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

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

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

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

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

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

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*.

Linear probability model: HMDA data, ctd

Next include *black* as a regressor:

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

Predicted probability of denial:

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

  $\Pr(deny = 1) = -.091 + .559\times.3 + .177\times1 = .254$

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

  $\Pr(deny = 1) = -.091 + .559\times.3 + .177\times0 = .077$

- 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

    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 (means linear!)*, 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

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

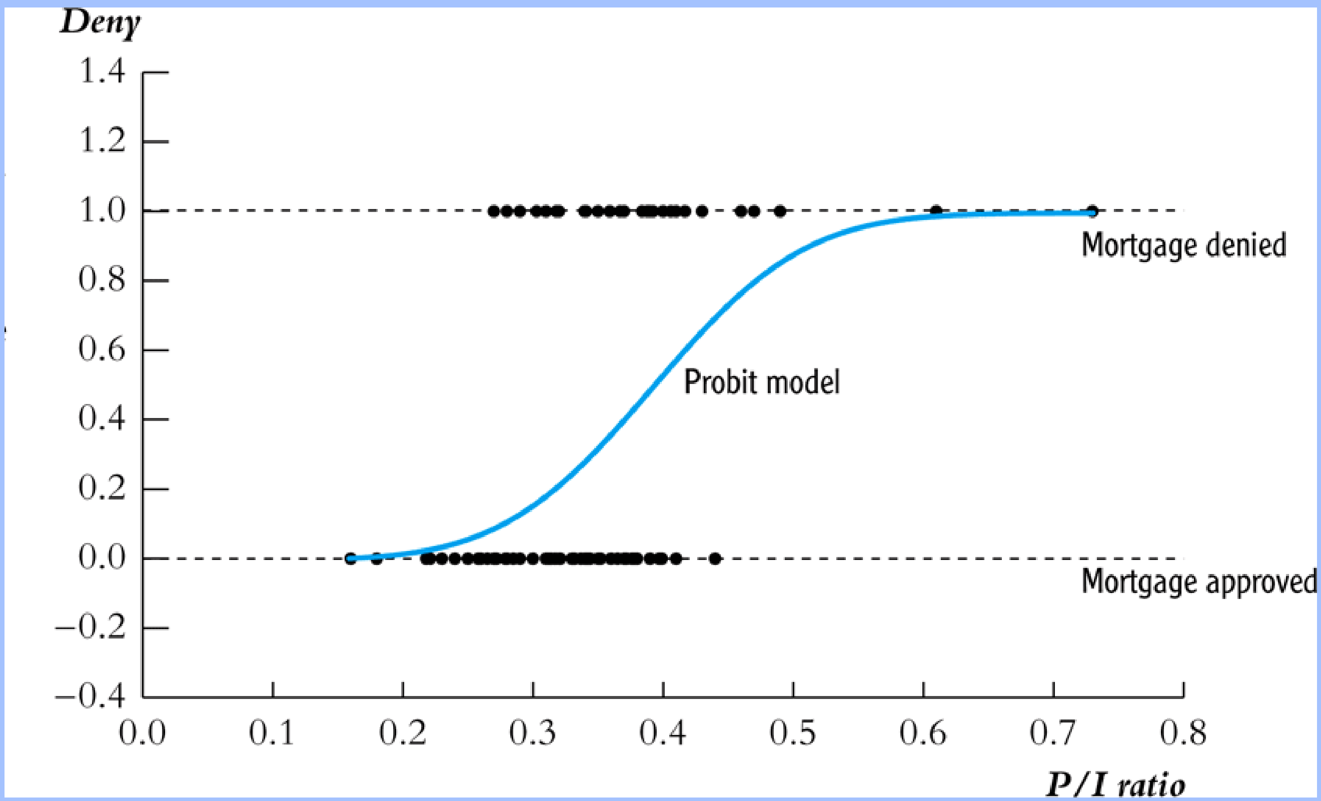
$$\Pr(Y = 1|X) = \beta_0 + \beta_1 X$$

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 β_1>0, and
  II. 0 ≤ Pr(Y = 1|X) ≤ 1 for all X

Probit regression models the probability that Y=1 using the cumulative standard normal distribution function, Φ(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…

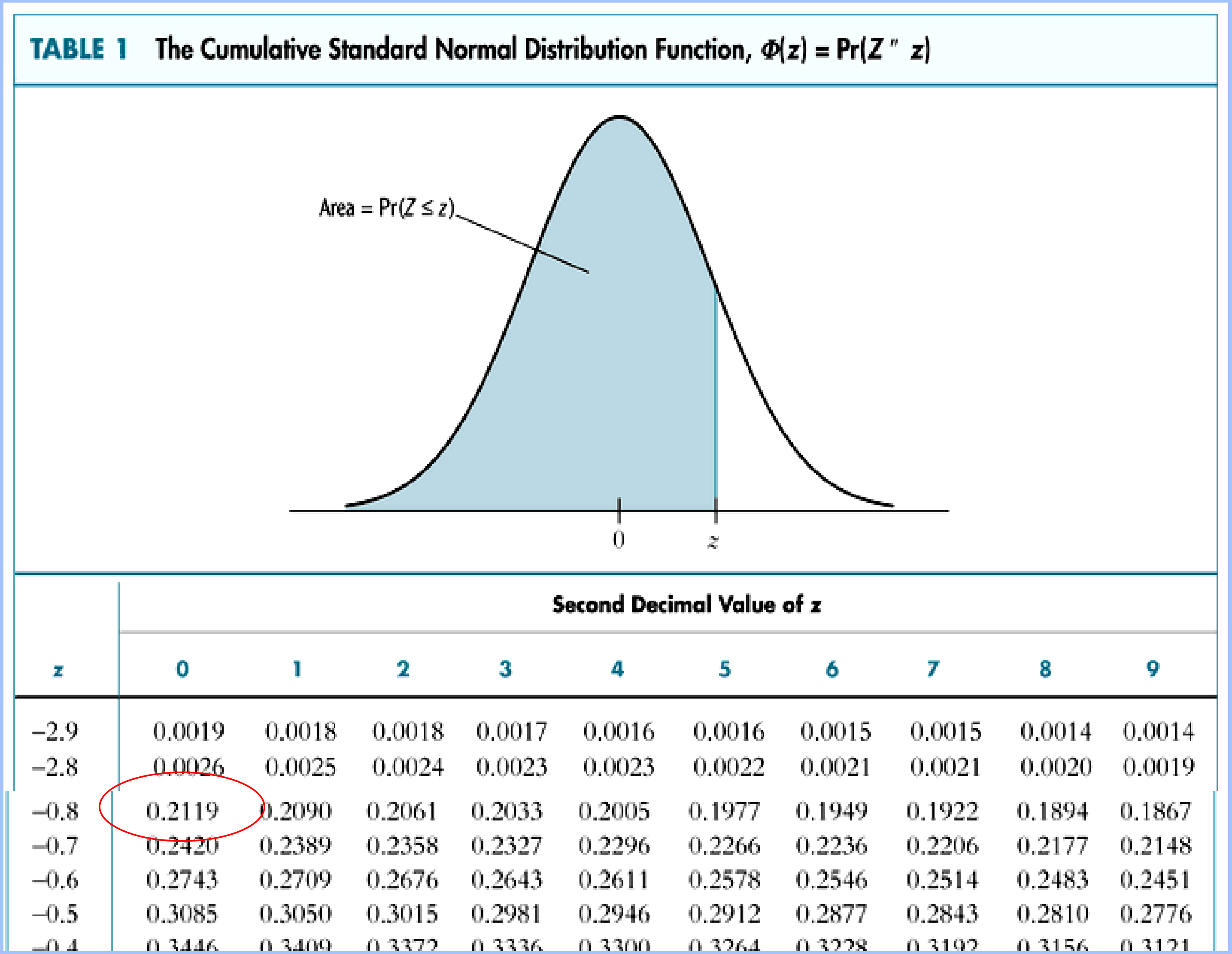


TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z) = \Pr(Z$ ″ $z)$

Area = Pr(Z ≤ z)
0
z

z	0	1	2	3	4	5	6	7	8	9
	Second Decimal Value of z									
−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:
  - $\Pr(Y = 1|X)$ is increasing in X for $\beta_1>0$
  - $0 \leq \Pr(Y = 1|X) \leq 1$ 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

HMDA data

```
. probit deny p_irat, r;
Iteration 0:   log likelihood =  -872.0853
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
------------------------------------------------------------------------------
```

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

HMDA data, ctd.

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

- Positive coefficient: *Does this make sense*?
- Predicted probabilities:

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

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

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

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

$$= \Phi(-1.00) = .159$$

- 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

HMDA 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
------------------------------------------------------------------------------
.  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
```

HMDA Example, ctd.

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

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

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

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

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

- 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 $z = \beta_0 + \beta_1 X$:

$$\mathrm{P(X)} = \Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

where F is the cumulative logistic distribution function:

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

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

Logit Regression

Odds:

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

Log of Odds (Logit):

$$\log(\frac{P(X)}{1-P(X)}) = \beta_0 + \beta_1 x$$

Logit regression, ctd.

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

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

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

so $\beta_0 + \beta_1 X = -3 + 2X.4 = -2.2$ 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

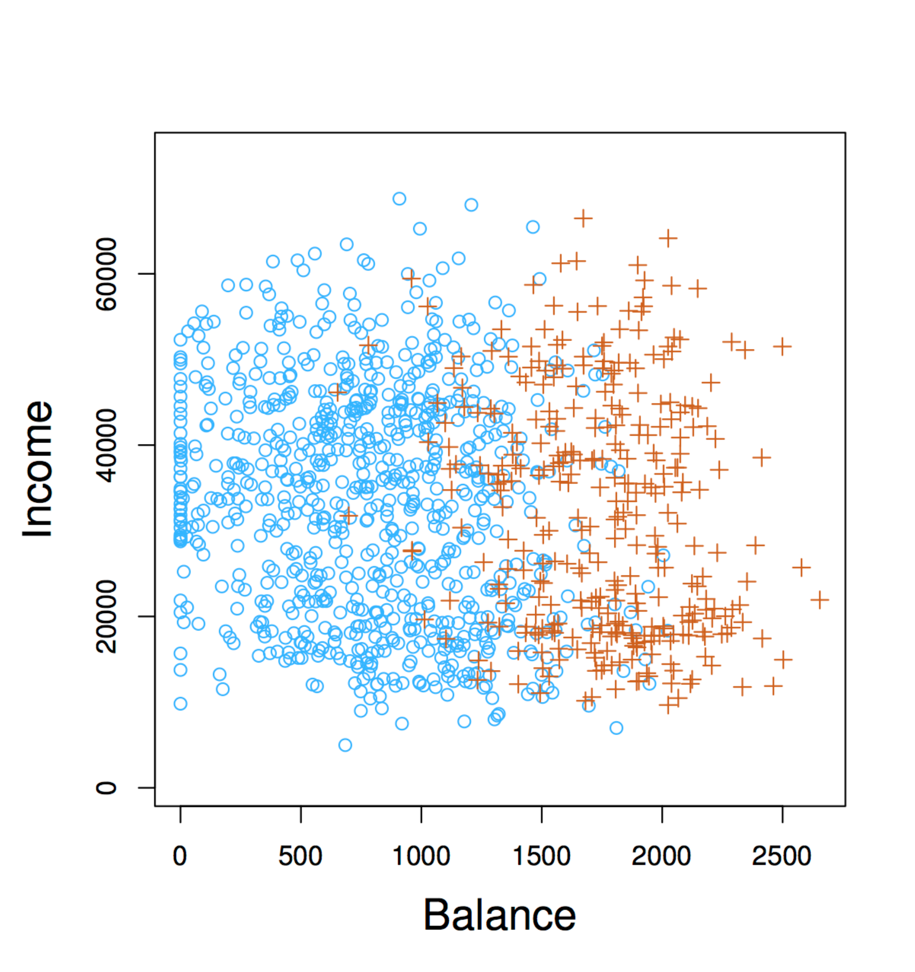
Credit Card Default Data

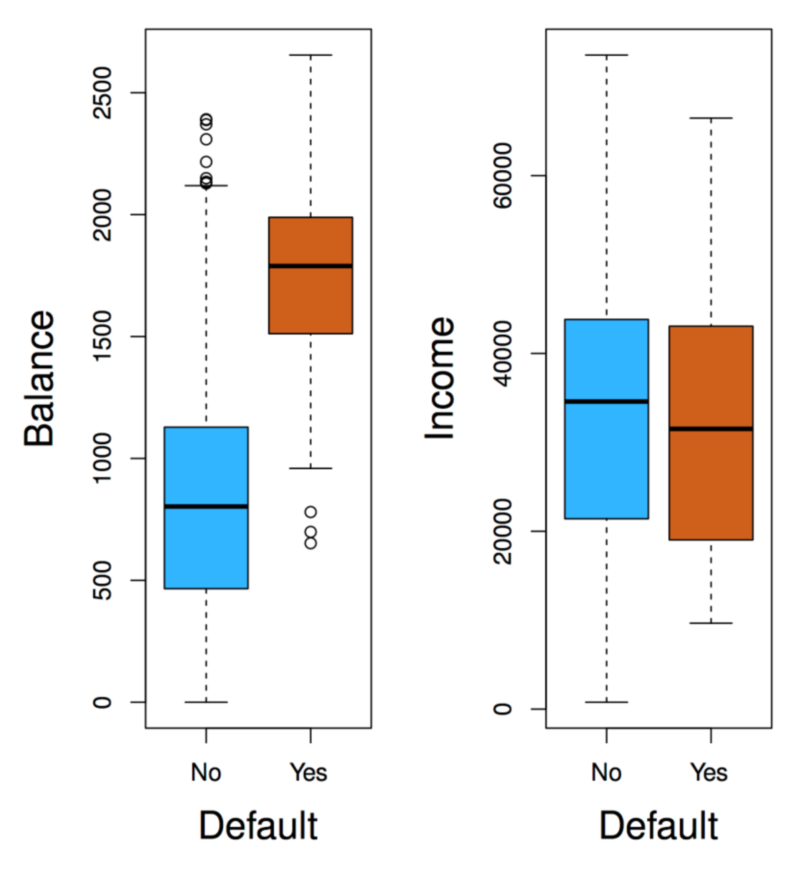
- We would like to be able to predict customers that are likely to default

- Possible X variables are:
  - Annual Income
  - Monthly credit card balance

- The Y variable (Default) is categorical: Yes or No

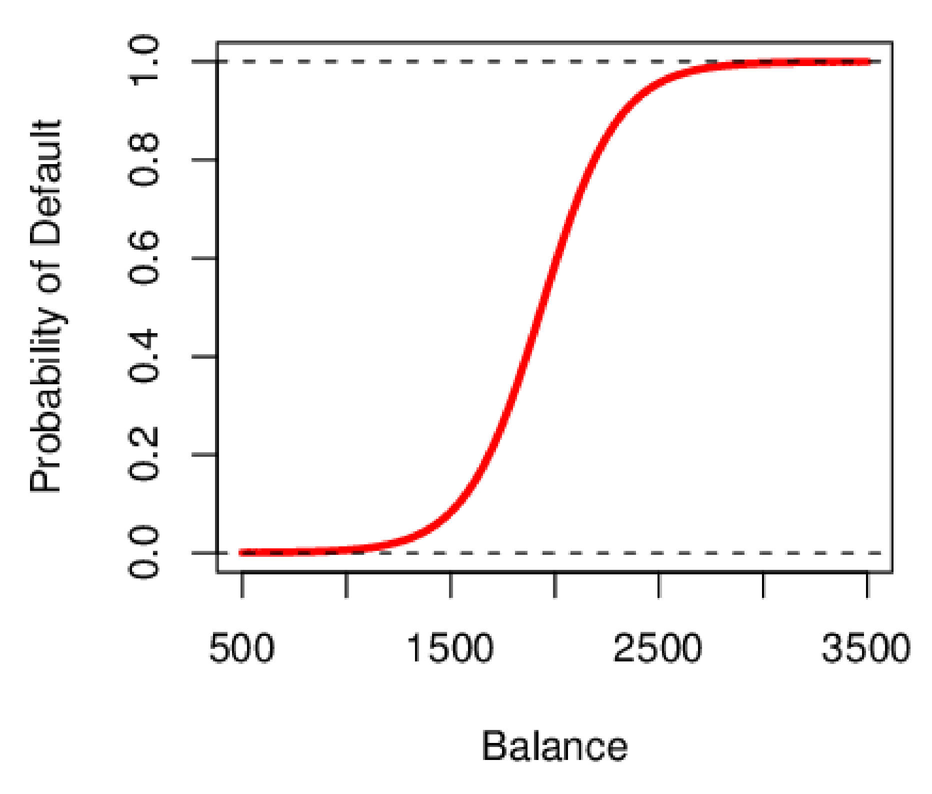
The Default Dataset





Logistic Function on Default Data

- Now the probability of default is close to, but not less than zero for low balances. And close to but not above 1 for high balances

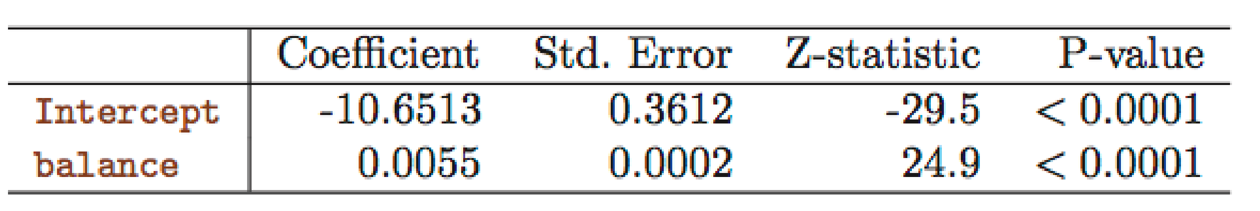


Interpreting β_1

- Interpreting what β_1 means is not very easy with logistic regression, simply because we are predicting P(Y) and not Y.
- If $\beta_1 = 0$, this means that there is no relationship between Y and X.
- If $\beta_1 > 0$, this means that when X gets larger so does the probability that Y = 1.
- If $\beta_1 < 0$, this means that when X gets larger, the probability that Y = 1 gets smaller.
- But how much bigger or smaller depends on where we are on the slope

Are the coefficients significant?

- We still want to perform a hypothesis test to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a Z test instead of a T test, but of course that doesn't change the way we interpret the p-value
- Here the p-value for balance is very small, and b_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Making Prediction

- Suppose an individual has an average balance of \$1000. What is their probability of default?

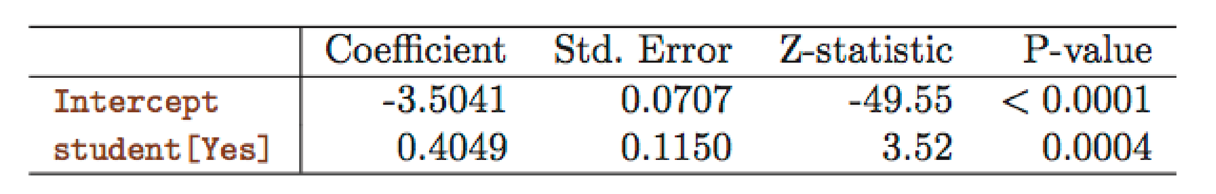
$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576$$

- The predicted probability of default for an individual with a balance of \$1000 is less than 1%.

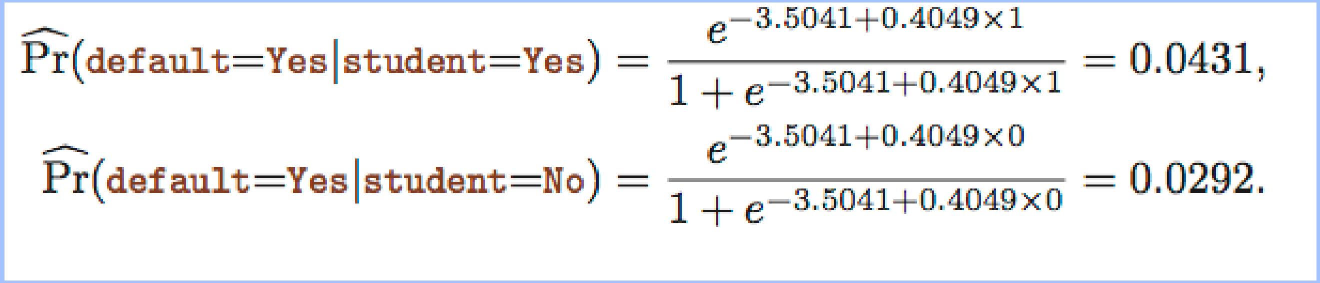
- For a balance of \$2000, the probability is much higher, and equals to 0.586 (58.6%).

Qualitative Predictors in Logistic Regression

- We can predict if an individual default by checking if she is a student or not. Thus we can use a qualitative variable “Student” coded as (Student = 1, Non-student =0).
- b_1 is positive: This indicates students tend to have higher default probabilities than non-students



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004



$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

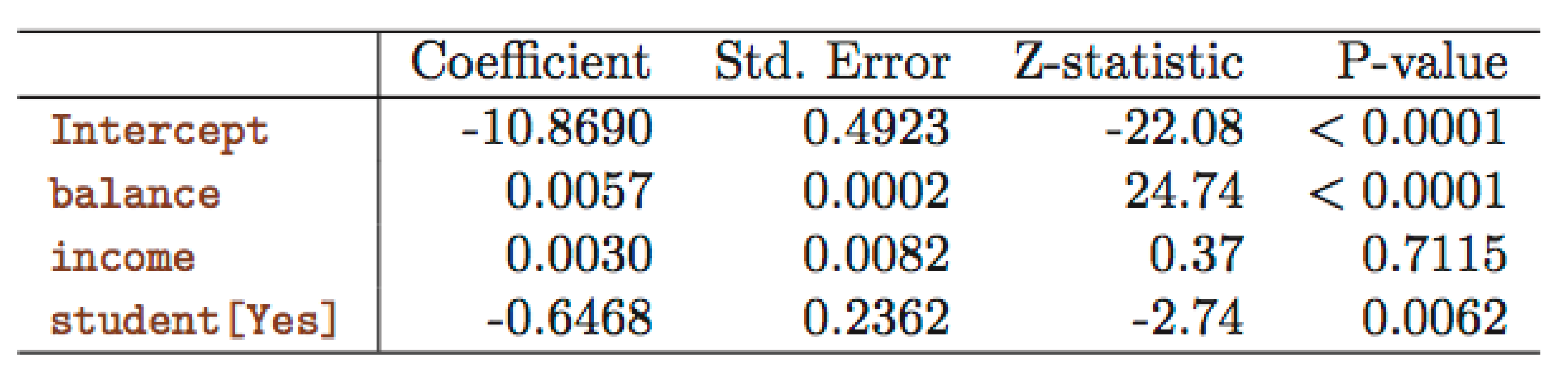
Multiple Logistic Regression

- We can fit multiple logistic just like regular regression

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Multiple Logistic Regression- Default Data

- Predict Default using:
  - Balance (quantitative)
  - Income (quantitative)
  - Student (qualitative)



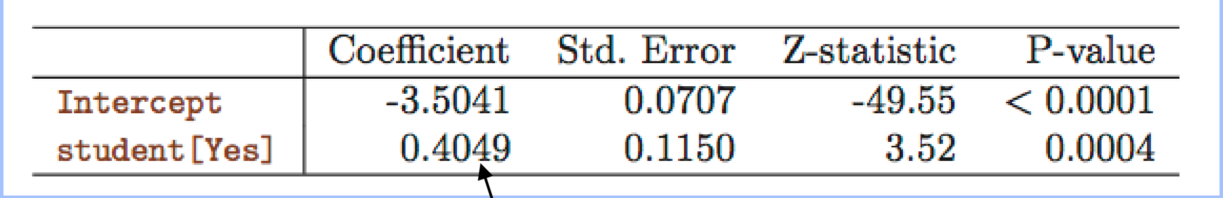
	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Predictions

- A student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default

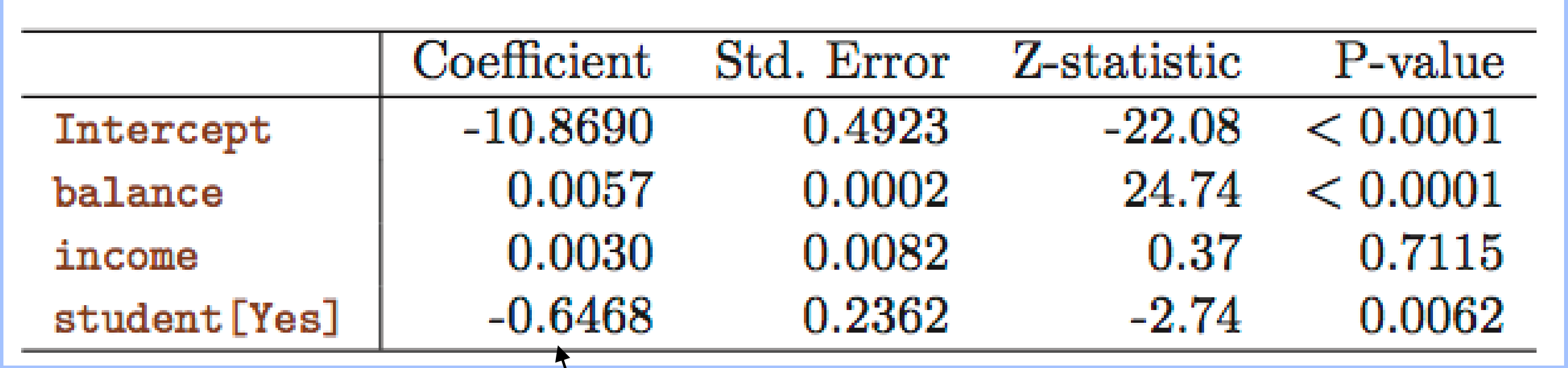
$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}} = 0.058.$$

An Apparent Contradiction!



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

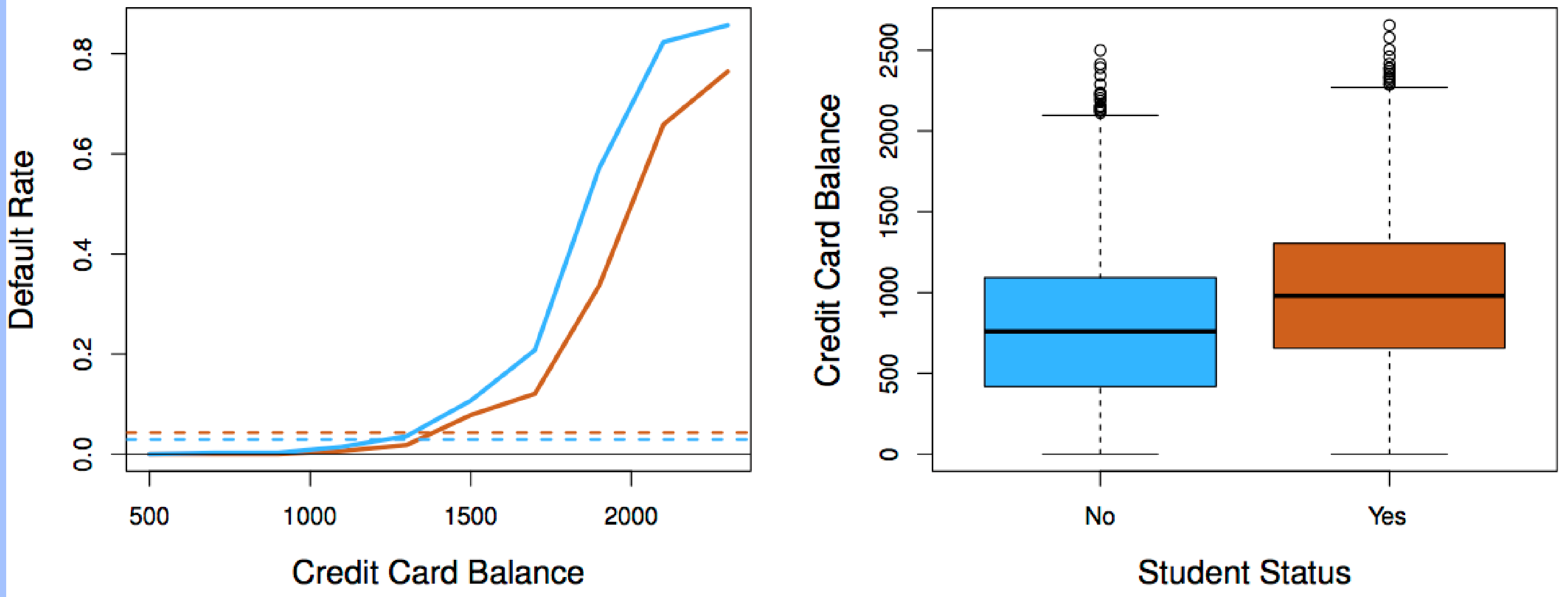
Positive



	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Negative

Students (Orange) vs. Non-students (Blue)



To whom should credit be offered?

- A student is risker than non students if no information about the credit card balance is available

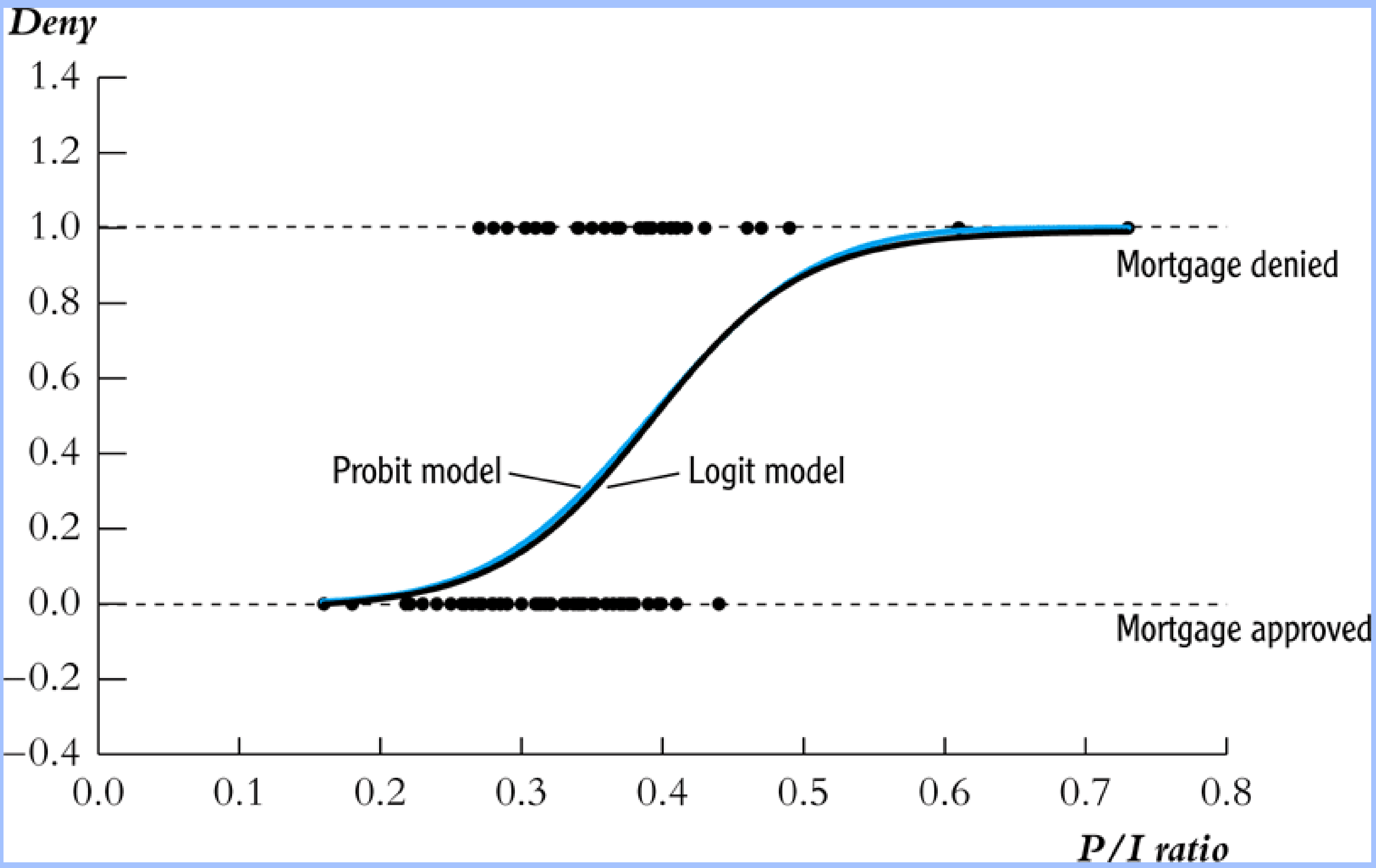
- However, that student is less risky than a non student with the same credit card balance!

HMDA Example

```
. 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
```

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.

Data set: See the article for details

Logit regression: 1 = died in Hezbollah military event

Table of logit results:

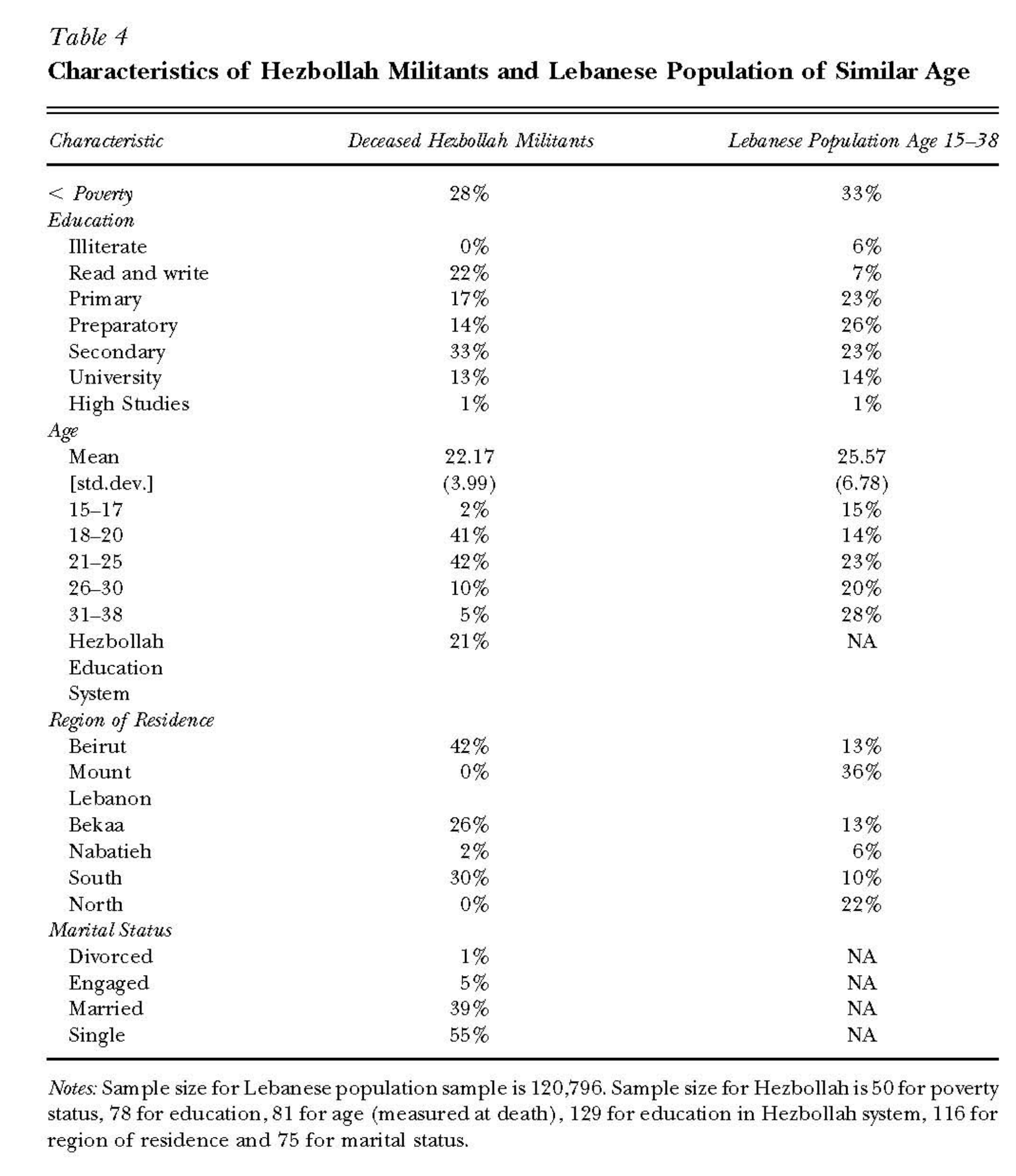


Table 4

Characteristics of Hezbollah Militants and Lebanese Population of Similar Age

Characteristic	*Deceased Hezbollah Militants*	*Lebanese Population Age 15–38*
< *Poverty*	28%	33%
Education		
Illiterate	0%	6%
Read and write	22%	7%
Primary	17%	23%
Preparatory	14%	26%
Secondary	33%	23%
University	13%	14%
High Studies	1%	1%
Age		
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
Region of Residence		
Beirut	42%	13%
Mount Lebanon	0%	36%
Bekaa	26%	13%
Nabatieh	2%	6%
South	30%	10%
North	0%	22%
Marital Status		
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.

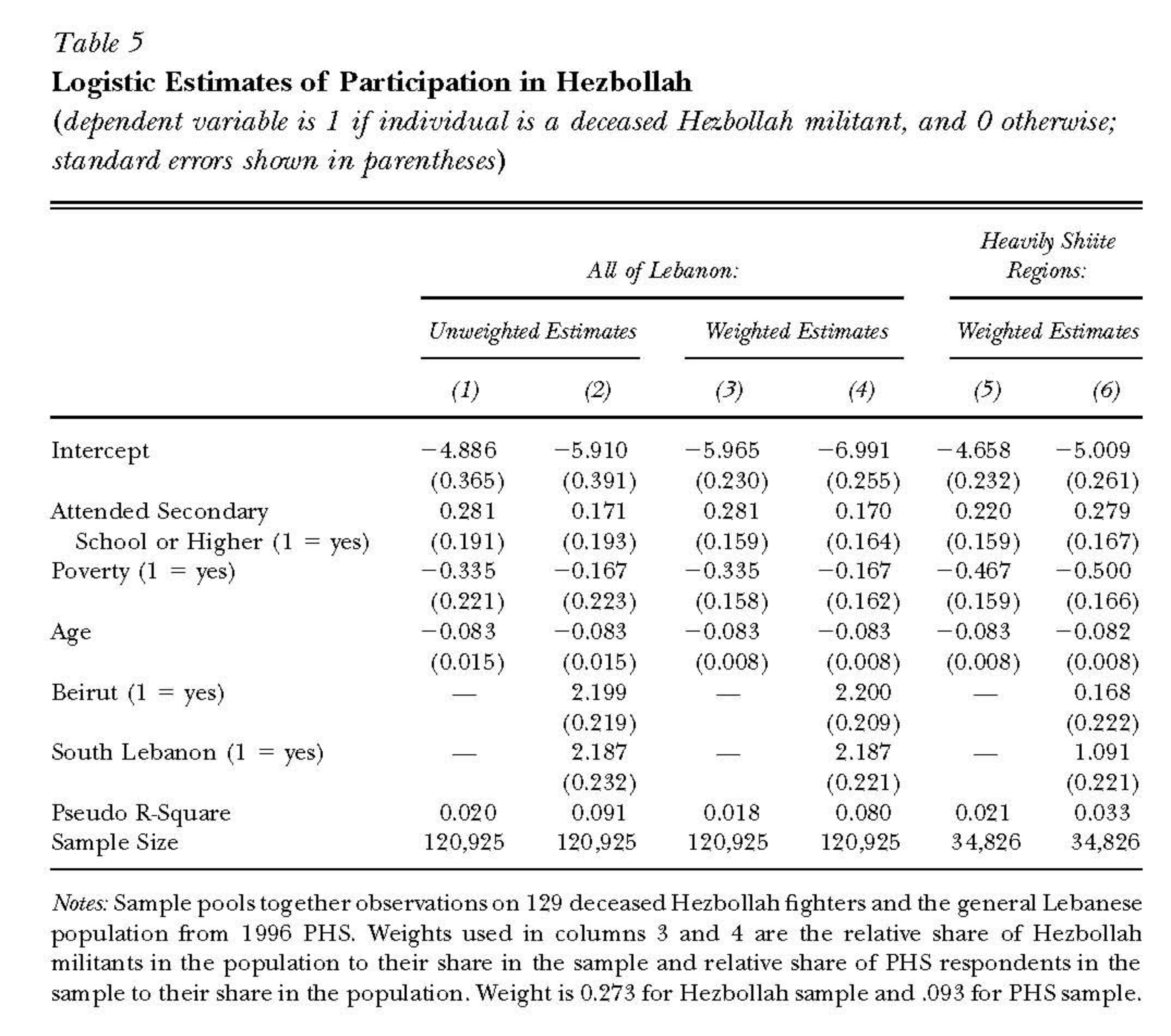


Table 5

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)*	*(6)*
Intercept	−4.886	−5.910	−5.965	−6.991	−4.658	−5.009
	(0.365)	(0.391)	(0.230)	(0.255)	(0.232)	(0.261)
Attended Secondary	0.281	0.171	0.281	0.170	0.220	0.279
School or Higher (1 = yes)	(0.191)	(0.193)	(0.159)	(0.164)	(0.159)	(0.167)
Poverty (1 = yes)	−0.335	−0.167	−0.335	−0.167	−0.467	−0.500
	(0.221)	(0.223)	(0.158)	(0.162)	(0.159)	(0.166)
Age	−0.083	−0.083	−0.083	−0.083	−0.083	−0.082
	(0.015)	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)
Beirut (1 = yes)	—	2.199	—	2.200	—	0.168
		(0.219)		(0.209)		(0.222)
South Lebanon (1 = yes)	—	2.187	—	2.187	—	1.091
		(0.232)		(0.221)		(0.221)
Pseudo R-Square	0.020	0.091	0.018	0.080	0.021	0.033
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):

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

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

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

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

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

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

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

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

Predicted change in probabilities:

Pr(*Y*=1|secondary = 1, poverty = 0, age = 20)

– Pr(*Y*=1|secondary = 1, poverty = 0, age = 20)

= .000646 – .000488 = .000158

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?*

Estimation and Inference in the Logit and Probit Models

We'll focus on the probit model:

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

- Estimation and inference
  - How can we estimate β_0 and β_1?
  - What is the sampling distribution of the estimators?
  - Why can we use the usual methods of inference?
- First motivate via nonlinear least squares
- 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 $\hat{\beta}_1$

Nonlinear least squares extends the idea of OLS to models in which the parameters enter nonlinearly:

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

How to solve this minimization problem?

- Calculus doesn't give and explicit solution.
- Solved *numerically* using the computer (specialized minimization algorithms)
- In practice, nonlinear least squares isn't used. A more efficient estimator (smaller variance) is…

The Maximum Likelihood Estimator of the Coefficients in the Probit Model

The ***likelihood function*** is the conditional density of $Y_1,\ldots,Y_n$ given $X_1,\ldots,X_n$, treated as a function of the unknown parameters β_0 and β_1.

- The maximum likelihood estimator (MLE) is the value of (β_0, β_1) that maximize the likelihood function.
- The MLE is the value of (β_0, β_1) that best describe the full distribution of the data.
- In large samples, the MLE is:
    - consistent
    - normally distributed
    - efficient (has the smallest variance of all estimators)

Special case: The probit MLE with no *X*

$Y = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1-p \end{cases}$ (Bernoulli distribution)

Data: $Y_1, \ldots, Y_n$, i.i.d.

Derivation of the likelihood starts with the density of Y_1:

$\Pr(Y_1 = 1) = p$ and $\Pr(Y_1 = 0) = 1-p$

so

$\Pr(Y_1 = y_1) = p^{y_1}(1-p)^{1-y_1}$ (*verify this for y_1=0, 1!*)

Joint density of (Y_1,Y_2): because Y_1 and Y_2 are independent,

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \Pr(Y_1 = y_1) \times \Pr(Y_2 = y_2)$$

$$= [\ p^{y_1}(1-p)^{1-y_1}\ ] \times [\ p^{y_2}(1-p)^{1-y_2}\ ]$$

$$= p^{(y_1+y_2)}(1-p)^{[2-(y_1+y_2)]}$$

Joint density of $(Y_1,..,Y_n)$:

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$$

$$= [\ p^{y_1}(1-p)^{1-y_1}\ ] \times [\ p^{y_2}(1-p)^{1-y_2}\ ] \times \ldots \times [\ p^{y_n}(1-p)^{1-y_n}\ ]$$

$$= p^{\sum_{i=1}^{n} y_i}(1-p)^{\left(n-\sum_{i=1}^{n} y_i\right)}$$

The likelihood is the joint density, treated as a function of the unknown parameters, which here is *p*:

$$f(p;Y_1,\ldots,Y_n) = p^{\sum_{i=1}^{n} Y_i}(1-p)^{\left(n-\sum_{i=1}^{n} Y_i\right)}$$

The MLE maximizes the likelihood. Its easier to work with the logarithm of the likelihood, $\ln[f(p;Y_1,\ldots,Y_n)]$:

$$\ln[f(p;Y_1,\ldots,Y_n)] = \left(\sum_{i=1}^{n} Y_i\right)\ln(p) + \left(n-\sum_{i=1}^{n} Y_i\right)\ln(1-p)$$

Maximize the likelihood by setting the derivative = 0:

$$\frac{d\ln f(p;Y_1,\ldots,Y_n)}{dp} = \left(\sum_{i=1}^{n} Y_i\right)\frac{1}{p} + \left(n-\sum_{i=1}^{n} Y_i\right)\left(\frac{-1}{1-p}\right) = 0$$

Solving for *p* yields the MLE; that is, $\hat{p}^{MLE}$ satisfies,

$$\left(\sum_{i=1}^{n} Y_i\right)\frac{1}{\hat{p}^{MLE}} + \left(n - \sum_{i=1}^{n} Y_i\right)\left(\frac{-1}{1-\hat{p}^{MLE}}\right) \quad = 0$$

or

$$\left(\sum_{i=1}^{n} Y_i\right)\frac{1}{\hat{p}^{MLE}} = \left(n - \sum_{i=1}^{n} Y_i\right)\frac{1}{1-\hat{p}^{MLE}}$$

or

$$\frac{\bar{Y}}{1-\bar{Y}} = \frac{\hat{p}^{MLE}}{1-\hat{p}^{MLE}}$$

or

$$\hat{p}^{MLE} = \bar{Y} = \text{fraction of 1's}$$

Whew… a lot of work to get back to the first thing you would think of using…but the nice thing is that this whole approach generalizes to more complicated models...

The MLE in the "No-X" Case (Bernoulli distribution), ctd.:

$$\hat{p}^{MLE} = \bar{Y} = \text{fraction of 1's}$$

- For Y_i i.i.d. Bernoulli, the MLE is the "natural" estimator of p, the fraction of 1's, which is $\bar{Y}$

- We already know the essentials of inference:
  - In large n, the sampling distribution of $\hat{p}^{MLE} = \bar{Y}$ is normally distributed
  - Thus inference is "as usual": hypothesis testing via *t*-statistic, confidence interval as ± 1.96*SE*

The MLE in the "No-X" Case (Bernoulli distribution), ctd:

- The theory of maximum likelihood estimation says that $\hat{p}^{MLE}$ is the ***most*** efficient estimator of *p* – of *all* possible estimators! – at least for large *n*. For this reason the MLE is primary estimator used for models that in which the parameters (coefficients) enter nonlinearly.

We are now ready to turn to the MLE of probit coefficients, in which the probability is conditional on *X*.

The Probit Likelihood with one *X*

The derivation starts with the density of Y_1, given X_1:

$\Pr(Y_1 = 1|X_1) = \Phi(\beta_0 + \beta_1 X_1)$

$\Pr(Y_1 = 0|X_1) = 1 – \Phi(\beta_0 + \beta_1 X_1)$

so

$\Pr(Y_1 = y_1|X_1) = \Phi(\beta_0 + \beta_1 X_1)^{y_1}[1 - \Phi(\beta_0 + \beta_1 X_1)]^{1-y_1}$

The probit likelihood function is the joint density of $Y_1,\ldots,Y_n$ given $X_1,\ldots,X_n$, treated as a function of β_0, β_1:

$$f(\beta_0, \beta_1;\ Y_1,\ldots,Y_n|X_1,\ldots,X_n)$$

$$= \{\ \Phi(\beta_0 + \beta_1 X_1)^{Y_1}[1 - \Phi(\beta_0 + \beta_1 X_1)]^{1-Y_1}\ \} \times$$

$$\ldots \times \{\Phi(\beta_0 + \beta_1 X_n)^{Y_n}[1 - \Phi(\beta_0 + \beta_1 X_n)]^{1-Y_n}\ \}$$

The probit likelihood function:

$f(\beta_0,\beta_1;\ Y_1,\ldots,Y_n|X_1,\ldots,X_n)$

$$= \{\ \Phi(\beta_0+\beta_1X_1)^{Y_1}[1-\Phi(\beta_0+\beta_1X_1)]^{1-Y_1}\ \}\times$$

$$\ldots\times\{\Phi(\beta_0+\beta_1X_n)^{Y_n}[1-\Phi(\beta_0+\beta_1X_n)]^{1-Y_n}\ \}$$

- $\hat\beta_0^{MLE}$, $\hat\beta_1^{MLE}$ maximize this likelihood function.
- But we can’t solve for the maximum explicitly! So the MLE must be maximized using numerical methods
- As in the case of no *X*, in large samples:
  - $\hat\beta_0^{MLE}$, $\hat\beta_1^{MLE}$ are consistent
  - $\hat\beta_0^{MLE}$, $\hat\beta_1^{MLE}$ are normally distributed
  - $\hat\beta_0^{MLE}$, $\hat\beta_1^{MLE}$ are asymptotically efficient – among all estimators (assuming the probit model is the correct model)

The Logit Likelihood with one *X*

- The only difference between probit and logit is the functional form used for the probability: Φ is replaced by the cumulative logistic function.
- Otherwise, the likelihood is similar;

- As with probit,
  - $\hat{\beta}_0^{MLE}$, $\hat{\beta}_1^{MLE}$ are consistent
  - $\hat{\beta}_0^{MLE}$, $\hat{\beta}_1^{MLE}$ are normally distributed
  - Their standard errors can be computed
  - Testing confidence intervals proceeds as usual

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. The pseudo-R^2 simplifies to the R^2 in the linear model with normally distributed errors.

Application to the Boston HMDA Data

- 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

$\Pr(deny=1|black$, other X's) = …

- 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…

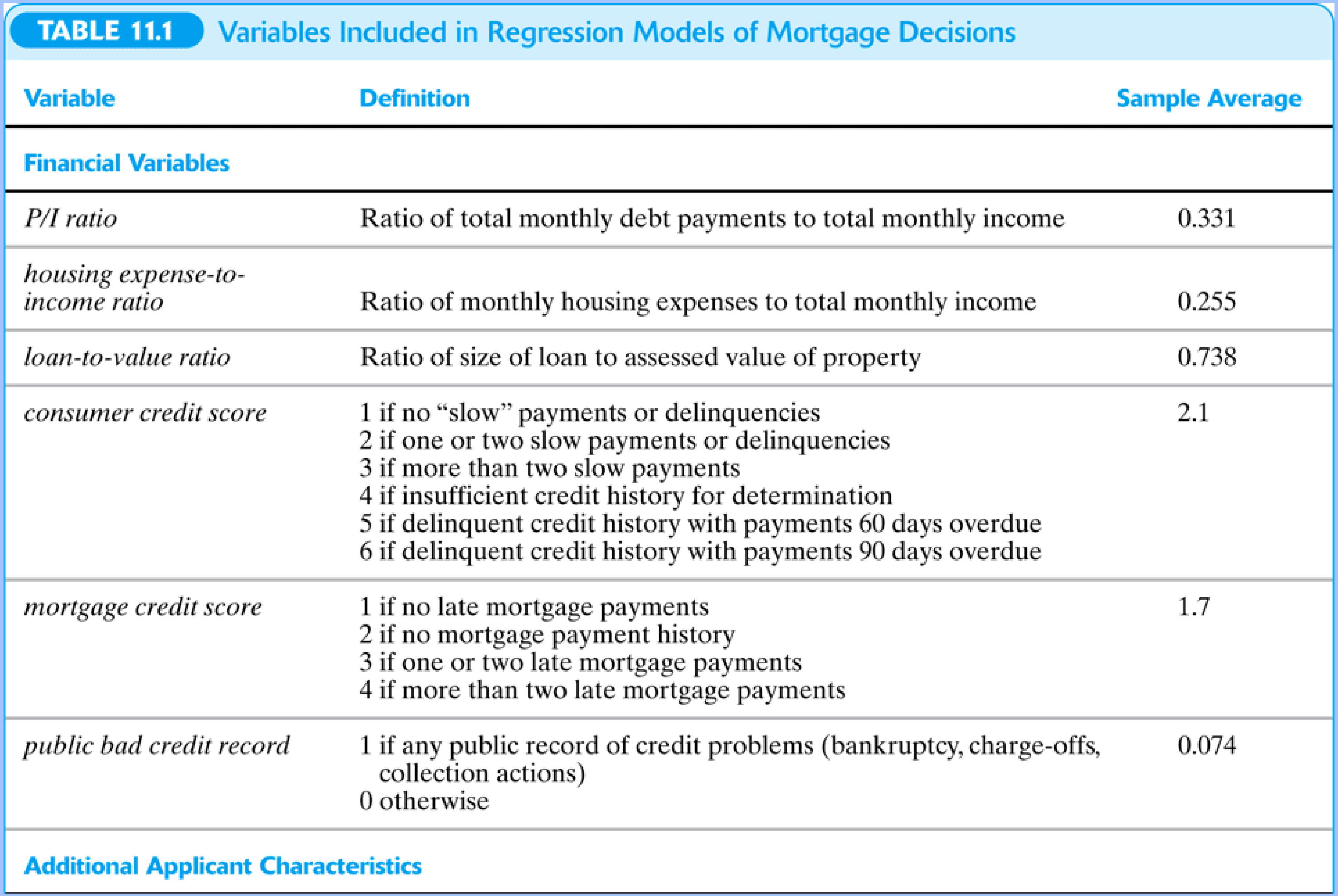
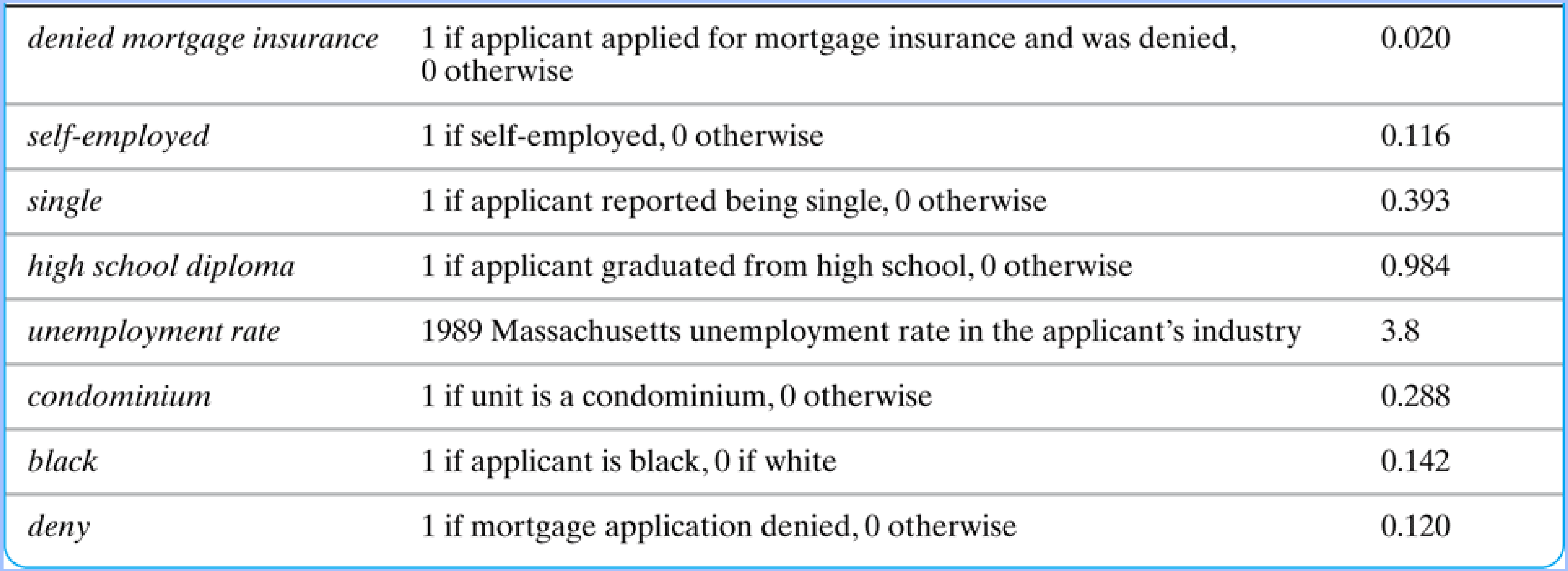


TABLE 11.1 Variables Included in Regression Models of Mortgage Decisions

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		

Table 11.1 *(cont.)*



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

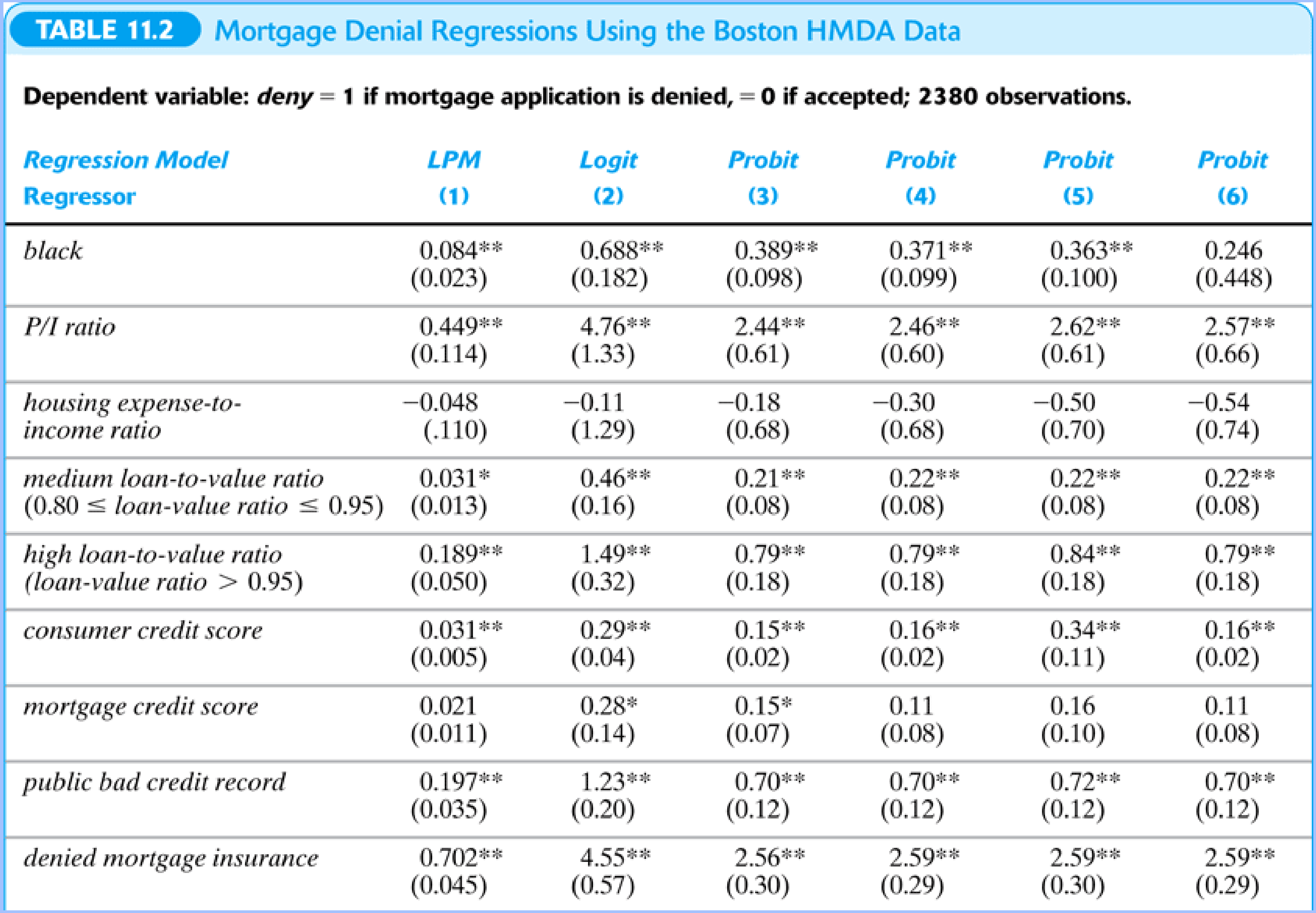


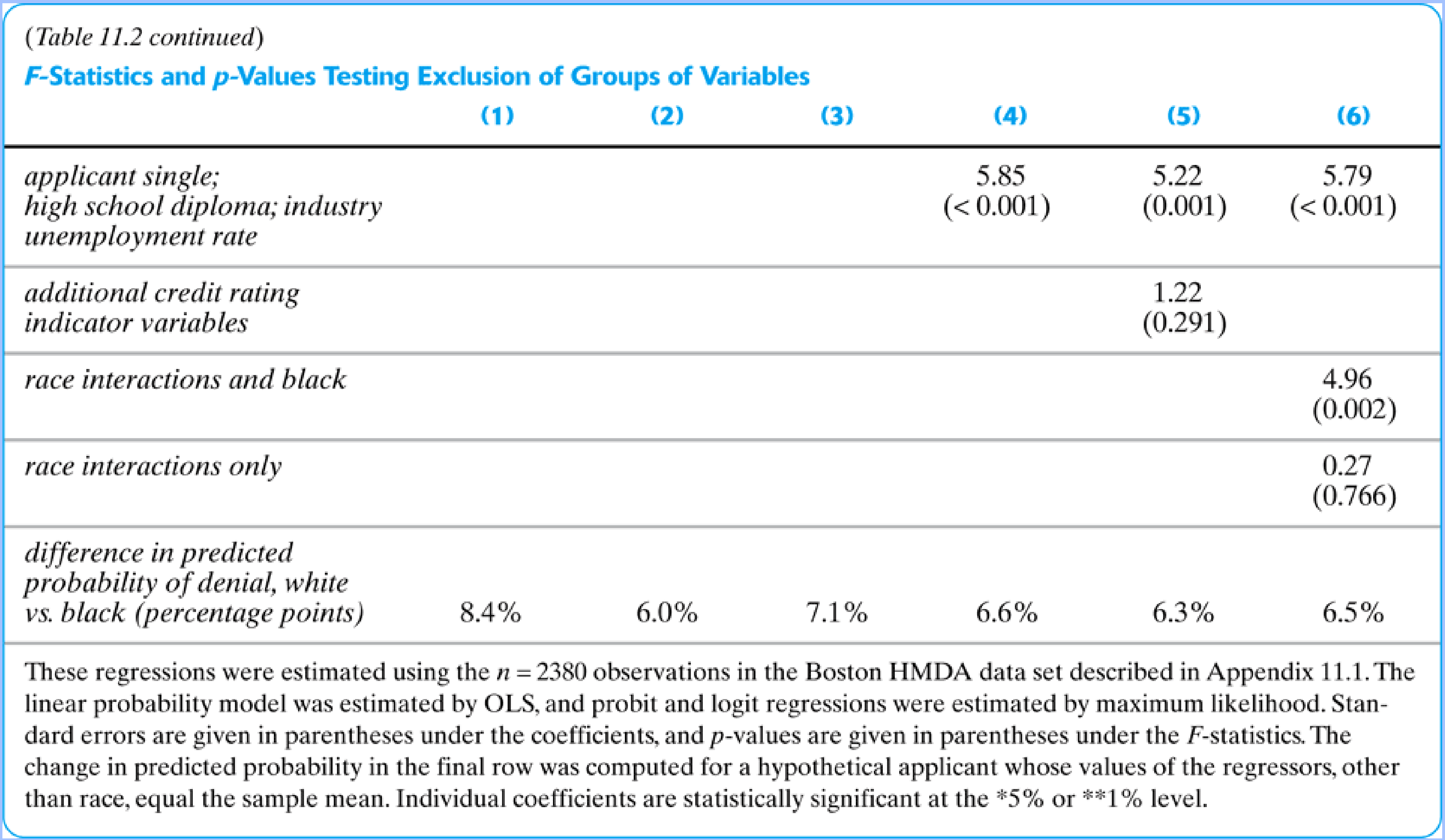
TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: *deny* = 1 if mortgage application is denied, = 0 if accepted; 2380 observations.

Regression Model Regressor	*LPM* (1)	*Logit* (2)	*Probit* (3)	*Probit* (4)	*Probit* (5)	*Probit* (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 ≤ *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)

(continued)



(*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
- 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.

Conclusion

- 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

Why Linear? Why Discriminant?

- LDA involves the determination of linear equation (just like linear regression) that will predict which group the case belongs to.

$$\boxed{D = v_1 X_1 + v_2 X_2 + \ldots + v_i X_i + a}$$

  - D: discriminant function
  - v: discriminant coefficient or weight for the variable
  - X: variable
  - a: constant

Purpose of LDA

- Choose the v’s in a way to maximize the distance between the means of different categories

- Good predictors tend to have large v’s (weight)

- We want to discriminate between the different categories

Assumptions of LDA

- The observations are a random sample

- Each predictor variable is normally distributed

Why not Logistic Regression?

- Logistic regression is unstable when the classes are well separated

- In the case where n is small, and the distribution of predictors X is approximately normal, then LDA is more stable than Logistic Regression

- LDA is more popular when we have more than two response classes

Estimating Bayes' Classifier

- With Logistic Regression we modeled the probability of Y being from the k^{th} class as

$$p(X) = \Pr(Y = k|X = x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

- However, Bayes' Theorem states

$$p(X) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

π_k: Probability of coming from class k (prior probability)

$f_k(x)$: Density function for X given that X is an observation from class k

Estimate π_k and $f_k(X)$

- We can estimate π_k and $f_k(x)$ to compute $p(X)$
- The most common model for $f_k(x)$ is the Normal Density

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2 \right)$$

- Using the density, we only need to estimate three quantities to compute $p(X)$

$$\mu_k \qquad \sigma_k^2 \qquad \pi_k$$

Use Training Data set for Estimation

- The mean μ_k could be estimated by the average of all training observations from the kth class.
- The variance σ_k^2 could be estimated as the weighted average of variances of all k classes.
- And, π_k is estimated as the proportion of the training observations that belong to the kth class.

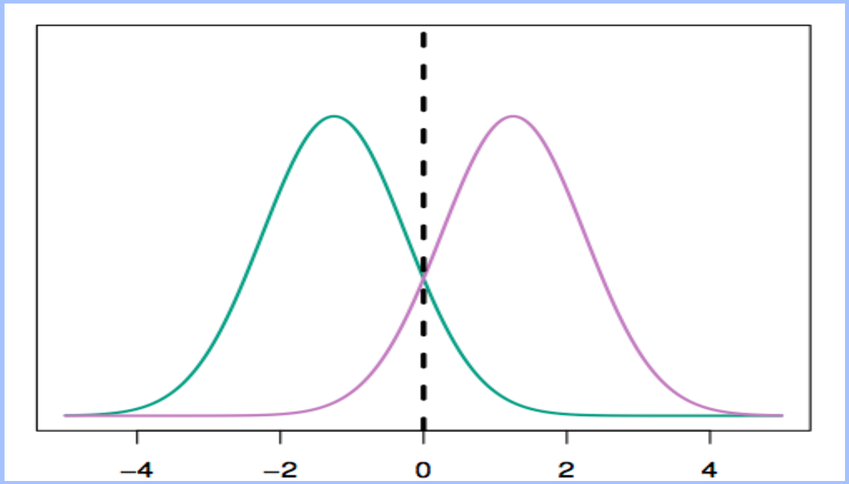
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i = k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i:y_i = k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = n_k / n.$$

A Simple Example with One Predictor (p =1)

- Suppose we have only one predictor (p = 1)
- Two normal density function $f_1(x)$ and $f_2(x)$, represent two distinct classes
- The two density functions overlap, so there is some uncertainty about the class to which an observation with an unknown class belongs
- The dashed vertical line represents Bayes' decision boundary



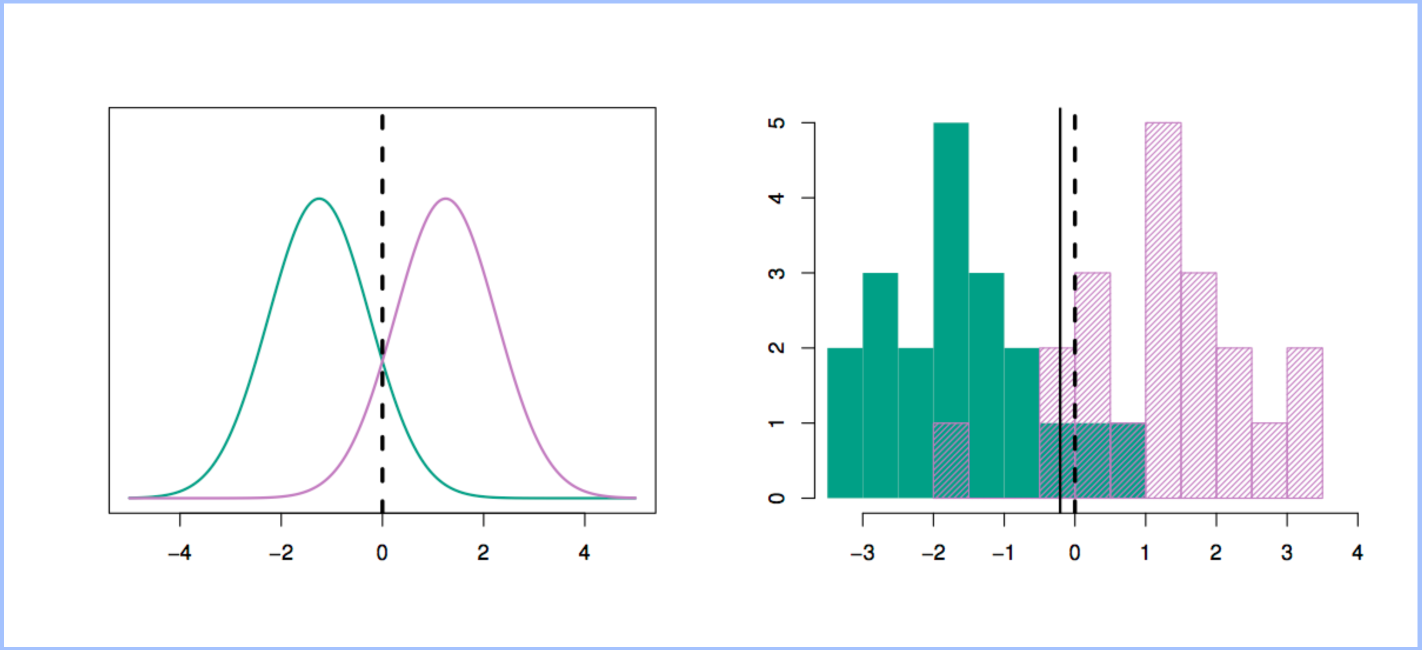
Apply LDA

- LDA starts by assuming that each class has a normal distribution with a common variance

- The mean and the variance are estimated

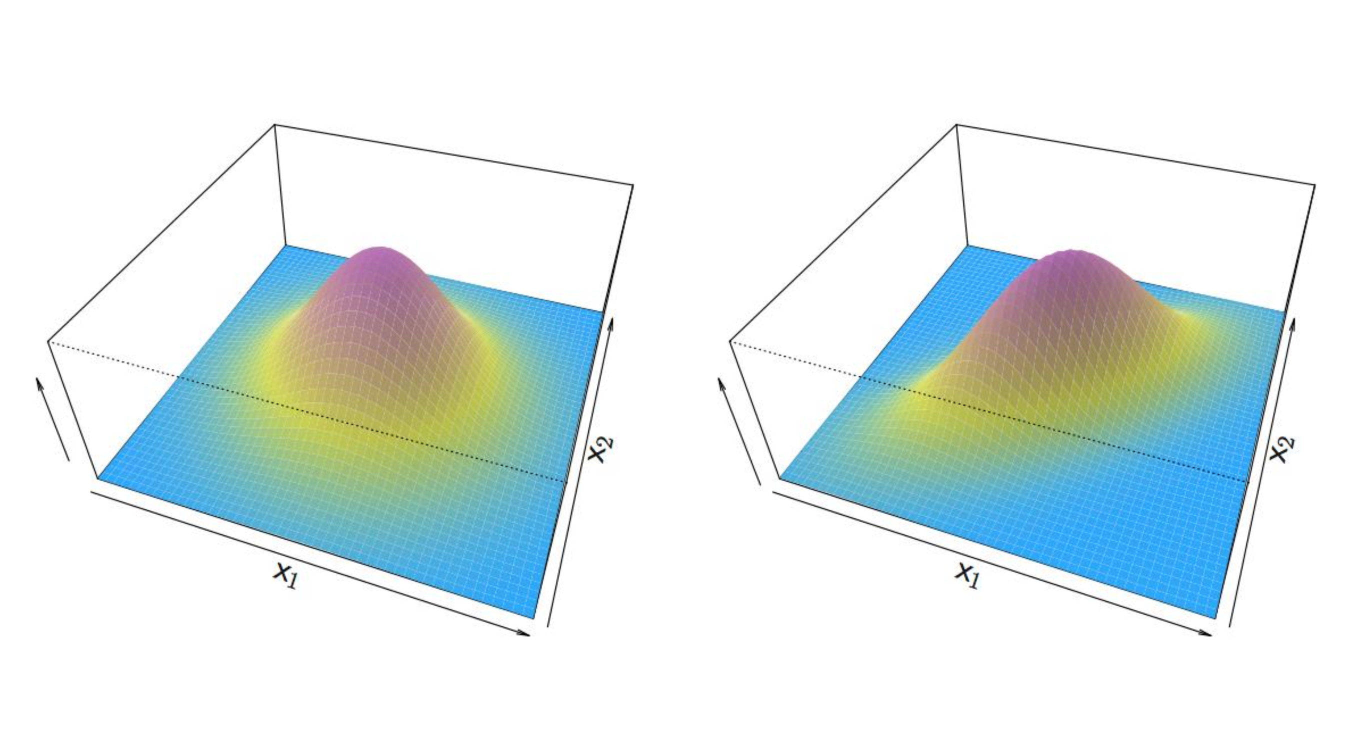
- Finally, Bayes' theorem is used to compute p_k and the observation is assigned to the class with the maximum probability among all k probabilities

- 20 observations were drawn from each of the two classes
- The dashed line is the Bayes' decision boundary
- The solid line is the LDA decision boundary
  - Bayes' error rate: 10.6%
  - LDA error rate: 11.1%
- Thus, LDA is performing pretty well!

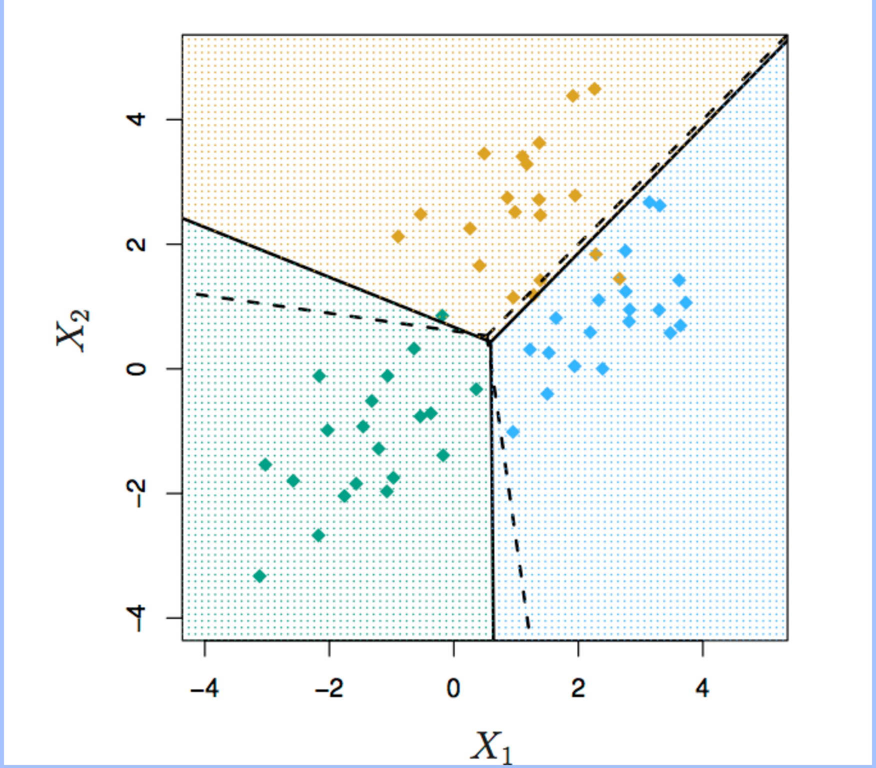


An Example When p > 1

- If X is multidimensional (p > 1), we use exactly the same approach except the density function *f(x)* is modeled using the multivariate normal density



- We have two predictors (p =2)
- Three classes
- 20 observations were generated from each class
- The solid lines are Bayes’ boundaries
- The dashed lines are LDA boundaries



Quadratic Discriminant Analysis (QDA)

- LDA assumed that every class has the same variance/ covariance
- However, LDA may perform poorly if this assumption is far from true
- QDA works identically as LDA except that it estimates separate variances/ covariance for each class

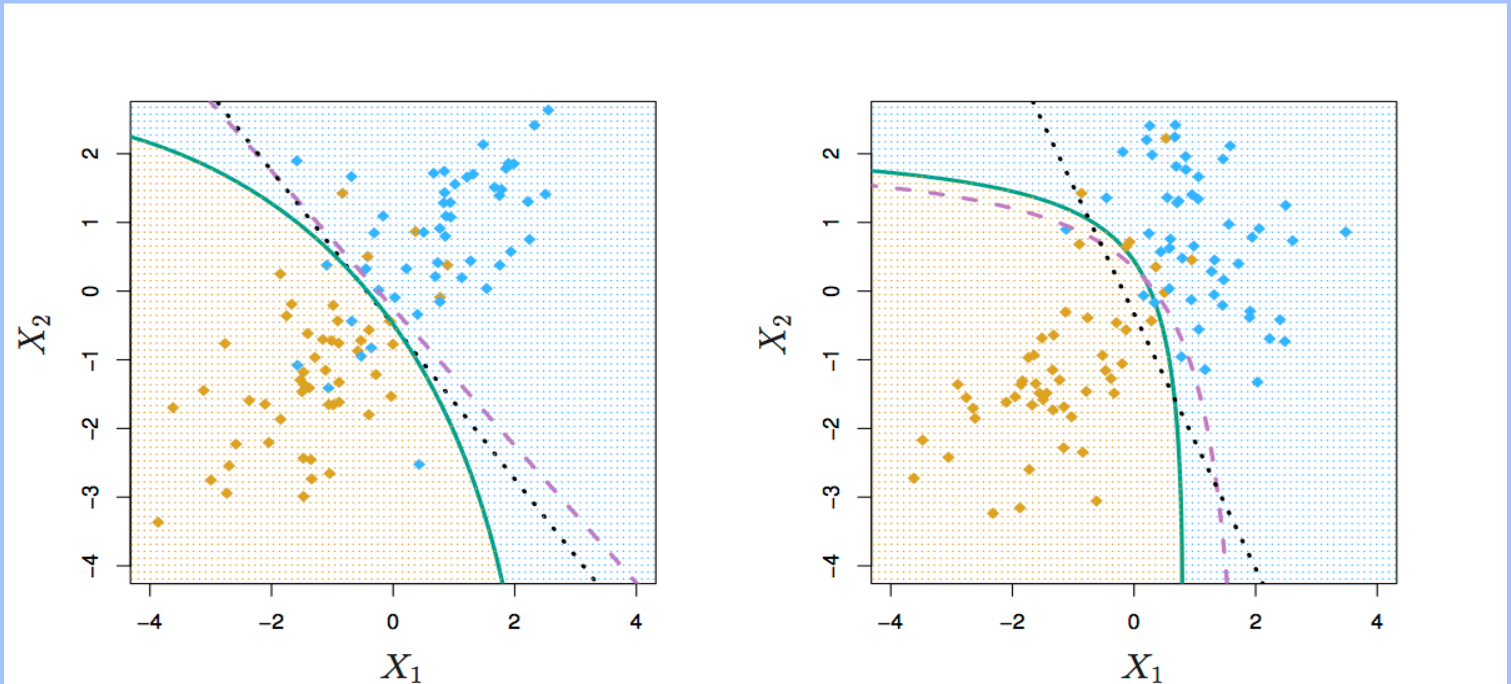
Which is better? LDA or QDA?

- Since QDA allows for different variances among classes, the resulting boundaries become quadratic

- Which approach is better: LDA or QDA?
  - QDA will work best when the variances are very different between classes and we have enough observations to accurately estimate the variances
  - LDA will work best when the variances are similar among classes or we don’t have enough data to accurately estimate the variances

Comparing LDA to QDA

- Black dotted: LDA boundary
- Purple dashed: Bayes’ boundary
- Green solid: QDA boundary
- Left: variances of the classes are equal (LDA is better fit)
- Right: variances of the classes are not equal (QDA is better fit)



Logistic Regression vs. LDA

- Similarity: Both Logistic Regression and LDA produce linear boundaries
- Difference: LDA assumes that the observations are drawn from the normal distribution with common variance in each class, while logistic regression does not have this assumption. LDA would do better than Logistic Regression if the assumption of normality hold, otherwise logistic regression can outperform LDA

QDA vs. (LDA, Logistic Regression)

- If the true decision boundary is:
  - Linear: LDA and Logistic outperforms
  - Non-linear: QDA outperforms

Polytomous (Multinomial) Logistic Regression

- In logistic regression, the response is a binary variable with 'success' and 'failure' being only two categories. But logistic regression can be extended to handle responses, *Y*, that are *polytomous*

- When r = 2, Y is dichotomous and we can model log of odds that an event occurs or does not occur. For binary logistic regression there is only 1 logit that we can form.
  - $\text{logit}(\pi)=\log(\pi/1- \pi)$ where $\pi = E[Yi]$

- When $r > 2$, we have a multi-category or polytomous response variable. There are $r\,(r - 1)/2$ logits (odds) that we can form, but only $(r - 1)$ are non-redundant.

Polytomous (Multinomial) Logistic Regression

- **Goal**: Give a simultaneous representation (summary) of the odds of being in one category relative to being in a designated category, called the baseline category, for all pairs of categories.
- This is an extension of binary logistic regression model, where we consider $r - 1$ non-redundant logits.
- A ***response variable Y***,
- A set of ***explanatory variables*** $X = (X_1, X_2, \ldots, X_k)$ can be discrete, continuous, or a combination of both.
- The regression coefficients mean the increase in log-odds of falling into category j versus category j^* resulting from a one-unit increase in the k^{th} predictor term, holding the other terms constant.

Polytomous (Multinomial) Logistic Regression

- **Goal**: Find the type of the glass

Polytomous (Multinomial) Logistic Regression

- **Goal**: Determine school and employment decisions for young men

Polytomous (Multinomial) Logistic Regression

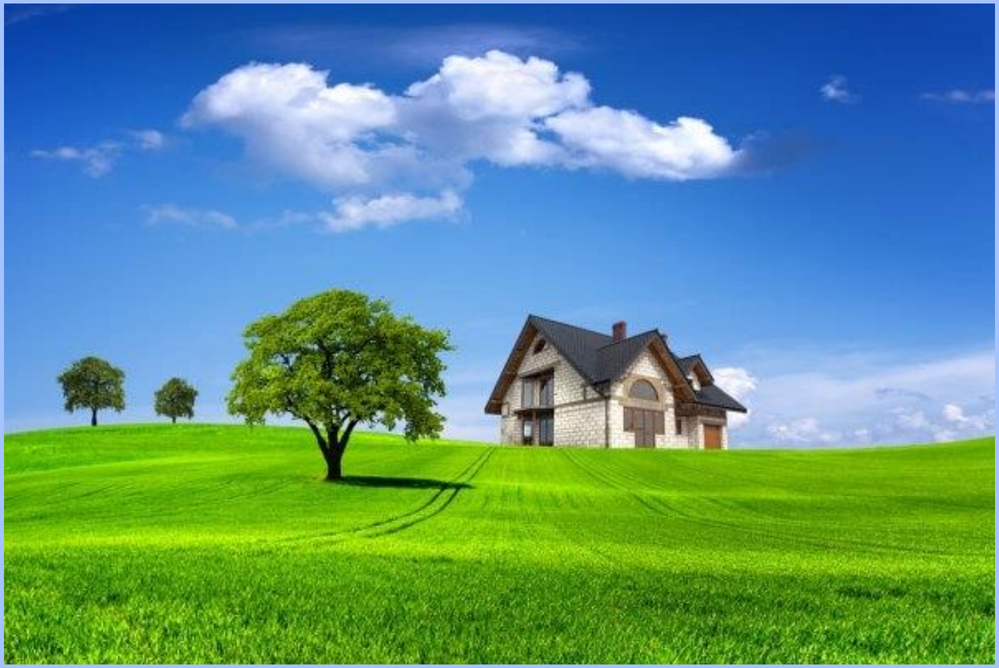
- **Relative prob's:** Coefficient of black working = 0.311. Exp(0.311) = 1.37. Thus, the relative probability of working rather school is 37% higher for blacks
- A common mistake is to interpret this as the probability of working is higher for blacks (*relative* probability of work over school).
- The coefficient of black in the home equation is 0.813. Exp(0.813) = 2.25
- Thus, the relative probability home - school for blacks is more than double.
- In short, black is associated with an increase in the relative probability of work over school, but also a much large increase in the relative probability of home over school. What happens with the actual probability of working depends on how these two effects balance out.

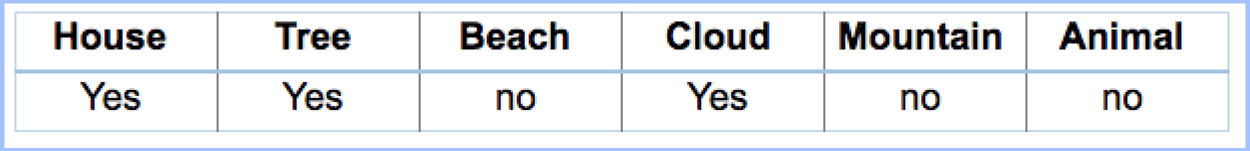
Polytomous (Multinomial) Logistic Regression

- **Marginal Effects:** We find that the average marginal effect of black on work is actually negative: -0.0406. This means that the probability of working is on average about four percentage points lower for blacks than for non-blacks with the same education and experience.

Multi-Label Classification

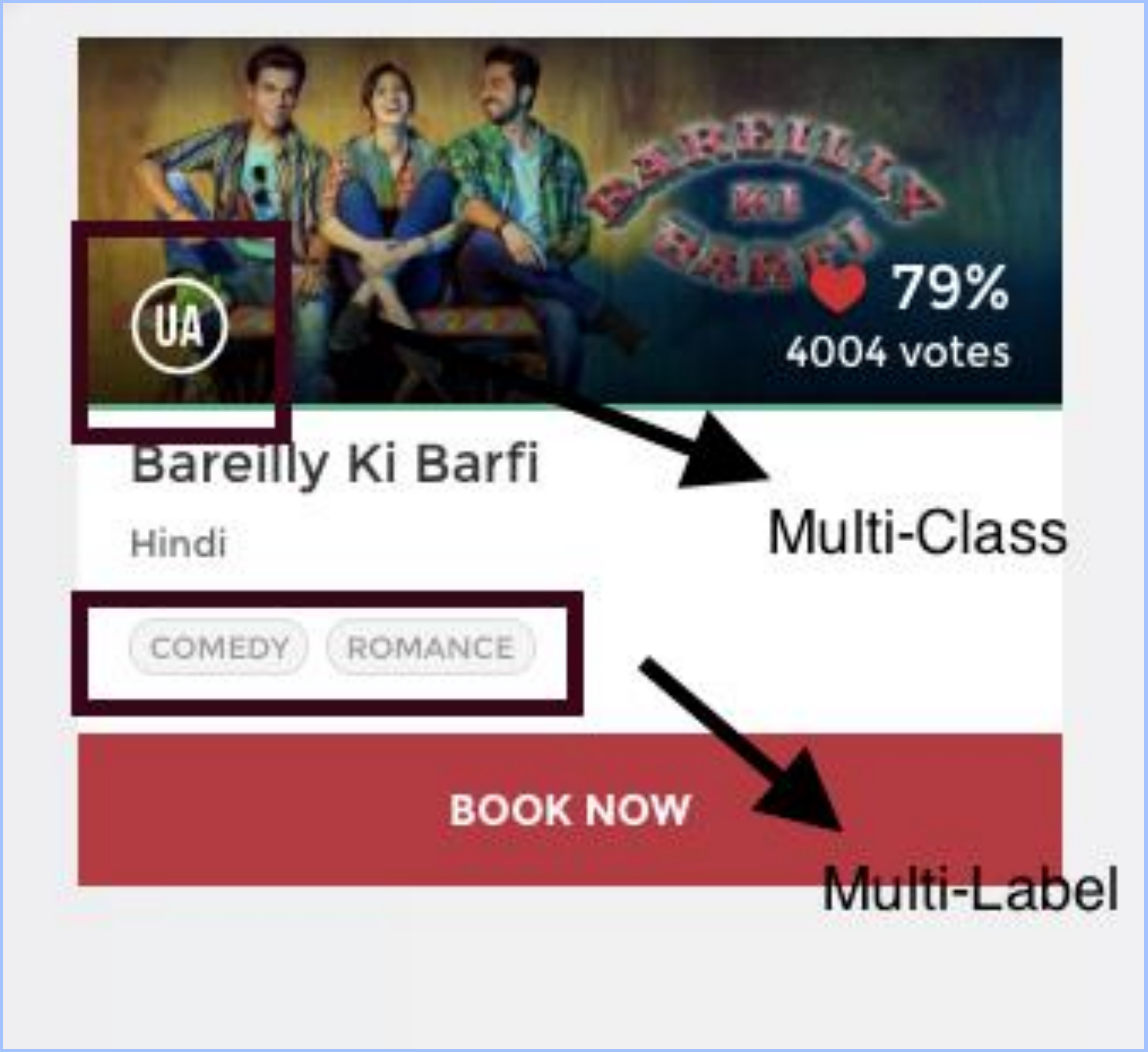
- In multi-label classification multiple labels are to be predicted for each instance.





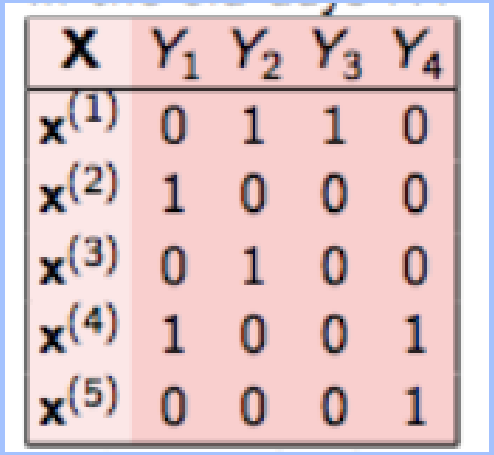
House	Tree	Beach	Cloud	Mountain	Animal
Yes	Yes	no	Yes	no	no

Multi-Label vs Multi-Class

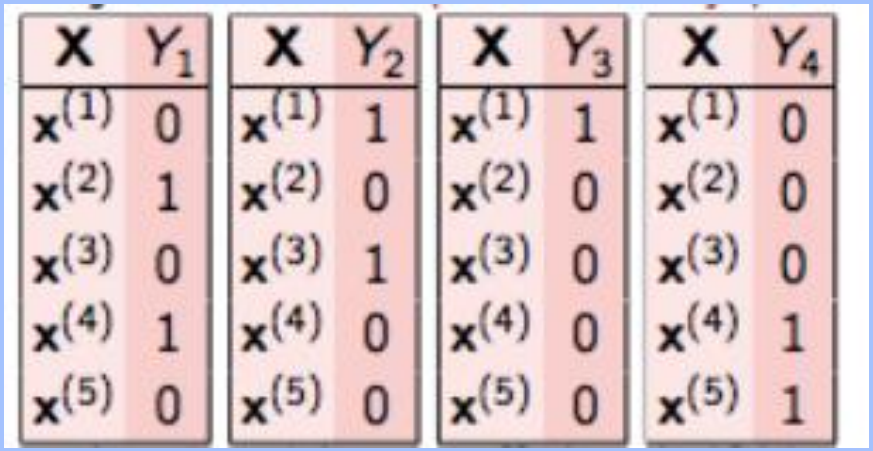


Binary Relevance

- This is the simplest technique, which basically treats each label as a separate single class classification problem.
- For example, let us consider a case as shown below. We have the data set like this, where X is the independent feature and Y's are the target variable.



X	Y_1	Y_2	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1



X	Y_1
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0

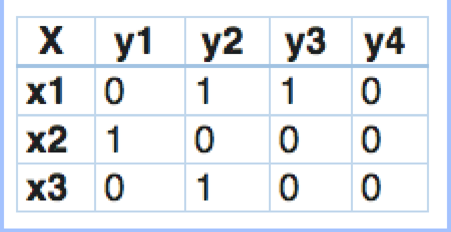
X	Y_2
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	1
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

X	Y_3
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

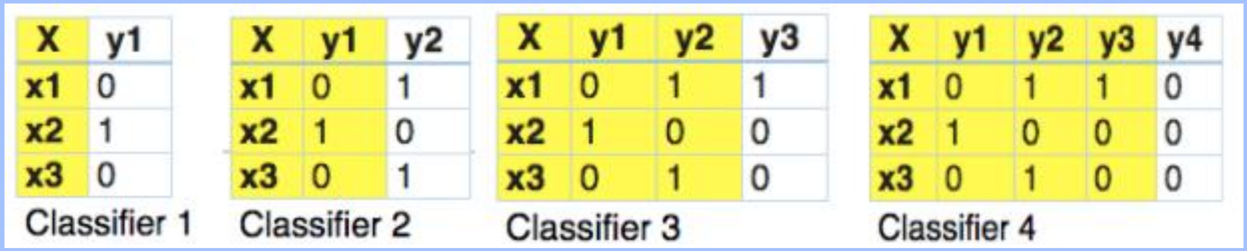
X	Y_4
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	1

Classifier Chains

- In this, the first classifier is trained just on the input data and then each next classifier is trained on the input space and all the previous classifiers in the chain.
- Let's try to this understand this by an example. In the dataset given below, we have X as the input space and Y's as the labels.



X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0



X	y1
x1	0
x2	1
x3	0

Classifier 1

X	y1	y2
x1	0	1
x2	1	0
x3	0	1

Classifier 2

X	y1	y2	y3
x1	0	1	1
x2	1	0	0
x3	0	1	0

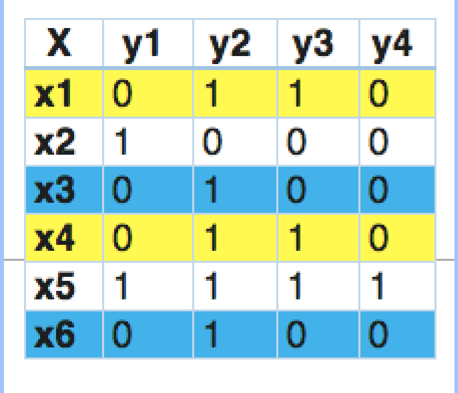
Classifier 3

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0

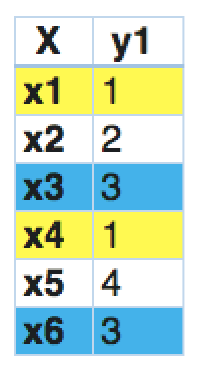
Classifier 4

Label Powerset

- In this, we transform the problem into a multi-class problem with one multi-class classifier is trained on all unique label combinations found in the training data.



X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0



X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3