Econ 423 – Lecture Notes

(These notes are slightly modified versions of lecture notes provided by Stock and Watson, 2007. They are for instructional purposes only and are not to be distributed outside of the classroom.)
Regression with a Binary Dependent Variable

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; \(X = \) years of education
- \(Y = \) person smokes, or not; \(X = \) income
- \(Y = \) mortgage application is accepted, or not; \(X = \) income, house characteristics, marital status, race
Example: Mortgage denial and race
The Boston Fed HMDA data set

- 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
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 \( \beta_1 \) 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.

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]

Recall assumption #1: \( E(u_i|X_i) = 0 \), so

\[ E(Y_i|X_i) = E(\beta_0 + \beta_1 X_i + u_i|X_i) = \beta_0 + \beta_1 X_i \]

When \( Y \) is binary,

\[ E(Y) = 1 \times \Pr(Y=1) + 0 \times \Pr(Y=0) = \Pr(Y=1) \]

so

\[ E(Y|X) = \Pr(Y=1|X) \]
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**.

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

- \( \beta_1 \) = change in probability that \( Y = 1 \) for a given \( \Delta x \):
\[
\beta_1 = \frac{\Pr(Y = 1 | X = x + \Delta x) - \Pr(Y = 1 | X = x)}{\Delta x}
\]
**Example**: linear probability model, HMDA data

Mortgage denial v. ratio of debt payments to income (P/I ratio) in the HMDA data set (subset)
Linear probability model: HMDA data, ctd.

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

• What is the predicted value for \( P/I\ ratio = .3 \)?

\[ \Pr(\text{deny} = 1 | P/I\ ratio = .3) = -.080 + .604 \times .3 = .151 \]

• Calculating “effects:” increase \( P/I\ ratio \) from .3 to .4:

\[ \Pr(\text{deny} = 1 | P/I\ ratio = .4) = -.080 + .604 \times .4 = .212 \]

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 (what?).
Linear probability model: HMDA data, ctd

Next include black as a regressor:

\[
\hat{deny} = -0.091 + 0.559 \times P/I\text{ ratio} + 0.177 \times black
\]

\[(0.032) \quad (0.098) \quad (0.025)\]

Predicted probability of denial:

• for black applicant with \( P/I\text{ ratio} = .3\):

\[
Pr(\hat{deny} = 1) = -0.091 + 0.559 \times 0.3 + 0.177 \times 1 = 0.254
\]

• for white applicant, \( P/I\text{ ratio} = .3\):

\[
Pr(\hat{deny} = 1) = -0.091 + 0.559 \times 0.3 + 0.177 \times 0 = 0.077
\]

• difference = 0.177 = 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

• 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:
  ○ Does it make sense that the probability should be linear in $X$?
  ○ 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:

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

This requires a *nonlinear* functional form for the probability. How about an “S-curve”…
The probit model satisfies these conditions:

- $0 \leq \Pr(Y = 1|X) \leq 1$ for all $X$
- $\Pr(Y = 1|X)$ to be increasing in $X$ (for $\beta_1 > 0$)
**Probit regression** models the probability that $Y=1$ using the cumulative standard normal distribution function, evaluated at $z = \beta_0 + \beta_1 X$:

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

- $\Phi$ is the cumulative normal distribution function.
- $z = \beta_0 + \beta_1 X$ 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, \( \Phi(z) = \Pr(Z \leq z) \)

<table>
<thead>
<tr>
<th>Second Decimal Value of ( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>(-2.9)</td>
</tr>
<tr>
<td>(-2.8)</td>
</tr>
</tbody>
</table>
Probit regression, ctd.

Why use the cumulative normal probability distribution?

- The “S-shape” gives us what we want:
  - $0 \leq \Pr(Y = 1|X) \leq 1$ for all $X$
  - $\Pr(Y = 1|X)$ to be increasing in $X$ (for $\beta_1 > 0$)

- Easy to use – the probabilities are tabulated in the cumulative normal tables

- Relatively straightforward interpretation:
  - $z$-value = $\beta_0 + \beta_1 X$
  - $\hat{\beta}_0 + \hat{\beta}_1 X$ is the predicted $z$-value, given $X$
  - $\beta_1$ is the change in the $z$-value for a unit change in $X$
Enter your feedback or question here
**STATA Example: HMDA data, ctd.**

\[
\Pr(\text{deny} = 1 \mid P / I\text{ratio}) = \Phi(-2.19 + 2.97 \times P/I \text{ratio})
\]

\[
(16) \quad (47)
\]

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

\[
\Pr(\text{deny} = 1 \mid P / I\text{ratio} = .3) = \Phi(-2.19 + 2.97 \times .3)
\]

\[
= \Phi(-1.30) = .097
\]

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

\[
\Pr(\text{deny} = 1 \mid P / I\text{ratio} = .4) = \Phi(-2.19 + 2.97 \times .4) = .159
\]

Predicted probability of denial rises from .097 to .159
Probit regression with multiple regressors

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

- $\Phi$ 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.
- $\beta_1$ 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

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

Log likelihood = -797.13604

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

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
Pseudo R2       =     0.0859

Log likelihood = -797.13604

------------------------------------------------------------------------------
|               Robust         |            |                      | z   |      P>|z|     |                     |                          |
|-------------|--------------|----------------------|-----|---------|-----------------|-------------------------|
|               | 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, ctd.**

\[
\Pr(\text{deny} = 1 | P/I, black)
\]

\[
= \Phi(-2.26 + 2.74 \times P/I \text{ ratio} + .71 \times black)
\]

\[
(0.16) \quad (0.44) \quad (0.08)
\]

- Is the coefficient on \textit{black} statistically significant?

- Estimated effect of race for \textit{P/I ratio} = .3:

\[
\Pr(\text{deny} = 1 | .3,1) = \Phi(-2.26+2.74 \times 0.3+.71 \times 1) = .233
\]

\[
\Pr(\text{deny} = 1 | .3,0) = \Phi(-2.26+2.74 \times 0.3+.71 \times 0) = .075
\]

- Difference in rejection probabilities = .158 (15.8 percentage points)

- \textit{Still plenty of room still for omitted variable bias}...
Logit Regression

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

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

$F$ is the cumulative logistic distribution function:

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

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

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

**Example:** \( \beta_0 = -3, \beta_1 = 2, X = .4, \)

so \( \beta_0 + \beta_1 X = -3 + 2 \times .4 = -2.2 \) so

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

Why bother with logit if we have probit?

- Historically, logit is more convenient computationally
- In practice, logit and probit are very similar
**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

```
|               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 (as usual) very close in this application.
**Example for class discussion:**

Characterizing the Background of Hezbollah Militants


Logit regression: $1 = $died in Hezbollah military event$

Table of logit results:
### Table 4
Characteristics of Hezbollah Militants and Lebanese Population of Similar Age

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Deceased Hezbollah Militants</th>
<th>Lebanese Population Age 15–38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty</td>
<td>28%</td>
<td>33%</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illiterate</td>
<td>9%</td>
<td>6%</td>
</tr>
<tr>
<td>Read and write</td>
<td>22%</td>
<td>7%</td>
</tr>
<tr>
<td>Primary</td>
<td>17%</td>
<td>23%</td>
</tr>
<tr>
<td>Preparatory</td>
<td>14%</td>
<td>26%</td>
</tr>
<tr>
<td>Secondary</td>
<td>33%</td>
<td>23%</td>
</tr>
<tr>
<td>University</td>
<td>13%</td>
<td>14%</td>
</tr>
<tr>
<td>High Studies</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>22.17</td>
<td>25.57</td>
</tr>
<tr>
<td>[std.dev.]</td>
<td>(3.99)</td>
<td>(6.78)</td>
</tr>
<tr>
<td>15–17</td>
<td>2%</td>
<td>15%</td>
</tr>
<tr>
<td>18–20</td>
<td>41%</td>
<td>14%</td>
</tr>
<tr>
<td>21–25</td>
<td>42%</td>
<td>23%</td>
</tr>
<tr>
<td>26–30</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>31–38</td>
<td>5%</td>
<td>28%</td>
</tr>
<tr>
<td>Hezbollah</td>
<td>21%</td>
<td>NA</td>
</tr>
<tr>
<td>Region of Residence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beirut</td>
<td>42%</td>
<td>13%</td>
</tr>
<tr>
<td>Mount</td>
<td>0%</td>
<td>36%</td>
</tr>
<tr>
<td>Lebanon</td>
<td>26%</td>
<td>13%</td>
</tr>
<tr>
<td>Bekaa</td>
<td>2%</td>
<td>6%</td>
</tr>
<tr>
<td>Nabatiyeh</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>South</td>
<td>10%</td>
<td>19%</td>
</tr>
<tr>
<td>North</td>
<td>0%</td>
<td>22%</td>
</tr>
<tr>
<td>Marital Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divorced</td>
<td>1%</td>
<td>NA</td>
</tr>
<tr>
<td>Engaged</td>
<td>5%</td>
<td>NA</td>
</tr>
<tr>
<td>Married</td>
<td>39%</td>
<td>NA</td>
</tr>
<tr>
<td>Single</td>
<td>55%</td>
<td>NA</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>All of Lebanon:</th>
<th>Heavily Shiite Regions:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unweighted Estimates</td>
<td>Weighted Estimates</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-4.886</td>
<td>-5.910</td>
</tr>
<tr>
<td></td>
<td>(0.365)</td>
<td>(0.391)</td>
</tr>
<tr>
<td>Attended Secondary School or Higher (1 = yes)</td>
<td>0.281</td>
<td>0.171</td>
</tr>
<tr>
<td>Poverty (1 = yes)</td>
<td>-0.335</td>
<td>-0.167</td>
</tr>
<tr>
<td></td>
<td>(0.221)</td>
<td>(0.223)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.083</td>
<td>-0.083</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Beirut (1 = yes)</td>
<td></td>
<td>2.199</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.219)</td>
</tr>
<tr>
<td>South Lebanon (1 = yes)</td>
<td></td>
<td>2.187</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.232)</td>
</tr>
<tr>
<td>Pseudo R-Square</td>
<td>0.020</td>
<td>0.091</td>
</tr>
<tr>
<td>Sample Size</td>
<td>120,925</td>
<td>120,925</td>
</tr>
</tbody>
</table>

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):

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

\[
\Pr(Y=1|\text{secondary} = 1, \text{poverty} = 0, \text{age} = 20) = \frac{1}{1 + e^{-(-5.965 + .281 \times 1 - .335 \times 0 - .083 \times 20)}} = \frac{1}{1 + e^{7.344}} = .000646 \text{ does this make sense?}
\]

\[
\Pr(Y=1|\text{secondary} = 0, \text{poverty} = 0, \text{age} = 20) = \frac{1}{1 + e^{-(-5.965 + .281 \times 0 - .335 \times 0 - .083 \times 20)}} = \frac{1}{1 + e^{7.625}} = .000488 \text{ does this make sense?}
\]
Predicted change in probabilities:

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

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

\[= .000646 - .000488 = .000158 \]

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 Probit (and Logit) Models

Probit model:
\[ Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X) \]

- Estimation and inference
  - How can we estimate \( \beta_0 \) and \( \beta_1 \)?
  - 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 estimator of probit coefficients:

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

How to solve this minimization problem?

- Calculus doesn’t give an explicit solution.
- Solved *numerically* using the computer (specialized minimization algorithms)
- In practice, nonlinear least squares isn’t used because it isn’t efficient – an estimator with a smaller variance is…
Probit estimation by maximum likelihood

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

- The maximum likelihood estimator (MLE) is the value of $(\beta_0, \beta_1)$ that maximize the likelihood function.
- The MLE is the value of $(\beta_0, \beta_1)$ that best describe the full distribution of the data.
- In large samples, the MLE is:
  - consistent
  - normally distributed
  - efficient (has the smallest variance of all estimators)
Special case: the probit MLE with no $X$

\[ Y = \begin{cases} 
1 & \text{with probability } p \\
0 & \text{with probability } 1 - p 
\end{cases} \quad \text{(Bernoulli distribution)}
\]

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

Derivation of the likelihood starts with the density of $Y_1$:

\[ \Pr(Y_1 = 1) = p \text{ and } \Pr(Y_1 = 0) = 1 - p \]

so

\[ \Pr(Y_1 = y_1) = p^{y_1} (1 - p)^{1-y_1} \quad \text{(verify this for } y_1=0, 1!) \]
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 \((Y_1,\ldots,Y_n)\):

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

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

\[
= p^{\sum_{i=1}^{n} y_i} (1 - p)^{\left(n - \sum_{i=1}^{n} y_i\right)}
\]
The likelihood is the joint density, treated as a function of the unknown parameters, which here is $p$:

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

The MLE maximizes the likelihood. It's 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,
\[
\left( \sum_{i=1}^{n} Y_i \right) \frac{1}{\hat{p}^{MLE}} + \left( n - \sum_{i=1}^{n} Y_i \right) \left( \frac{-1}{1 - \hat{p}^{MLE}} \right) = 0
\]

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

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

\textit{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 \( \pm 1.96SE \)
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 $n$. (Much stronger than the Gauss-Markov theorem). This is why people use the MLE.

• STATA note: to emphasize requirement of large-$n$, the printout calls the $t$-statistic the $z$-statistic; instead of the $F$-statistic, the *chi-squared* statistic ($= q \times F$).

• Now we extend this to probit – in which the probability is conditional on $X$ – the MLE of the probit coefficients.
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 $\beta_0, \beta_1$:
\[
f(\beta_0, \beta_1; Y_1, \ldots, Y_n|X_1, \ldots, X_n) \\
= \{ \Phi(\beta_0 + \beta_1 X_1)^{y_1} [1 - \Phi(\beta_0 + \beta_1 X_1)]^{1-y_1} \} \times \\
\ldots \times \{ \Phi(\beta_0 + \beta_1 X_n)^{y_n} [1 - \Phi(\beta_0 + \beta_1 X_n)]^{1-y_n} \}
\]
The probit likelihood function:

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

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

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

- Can’t solve for the maximum explicitly
- Must maximize 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 Probit MLE, ctd.

- Standard errors of $\hat{\beta}_0^{MLE}$, $\hat{\beta}_1^{MLE}$ are computed automatically…
- Testing, confidence intervals proceeds as usual
- For multiple $X$’s, see SW App. 11.2
The logit likelihood with one $X$

- The only difference between probit and logit is the functional form used for the probability: $\Phi$ is replaced by the cumulative logistic function.
- Otherwise, the likelihood is similar; for details see SW App. 11.2
- 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 predicted probability is $>50\%$ (if $Y_i=1$) or is $<50\%$ (if $Y_i=0$).

2. The *pseudo-$R^2$* measure the fit using the likelihood function: measures the improvement in the value of the log likelihood, relative to having no $X$’s (see SW App. 11.2). This simplifies to the $R^2$ in the linear model with normally distributed errors.
• Mortgages (home loans) are an essential part of buying a home.
• Is there differential access to home loans by race?
• If two otherwise identical individuals, one white and one black, applied for a home loan, is there a difference in the probability of denial?
The HMDA Data Set

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

• Loan officer uses key financial variables:
  o \textit{P/I ratio}
  o housing expense-to-income ratio
  o loan-to-value ratio
  o personal credit history

• The decision rule is nonlinear:
  o loan-to-value ratio > 80%
  o loan-to-value ratio > 95% (what happens in default?)
  o credit score
Regression specifications

\[ \text{Pr}(\text{deny}=1|\text{black}, \text{other } X\text{'s}) = \ldots \]

- linear probability model
- probit

Main problem with the regressions so far: potential omitted variable bias. All these (i) enter the loan officer decision function, all (ii) are or could be correlated with race:
  - wealth, type of employment
  - credit history
  - family status

The HMDA data set is very rich…
<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Sample Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial Variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P/I ratio</td>
<td>Ratio of total monthly debt payments to total monthly income</td>
<td>0.331</td>
</tr>
<tr>
<td>housing expense-to-income ratio</td>
<td>Ratio of monthly housing expenses to total monthly income</td>
<td>0.255</td>
</tr>
<tr>
<td>loan-to-value ratio</td>
<td>Ratio of size of loan to assessed value of property</td>
<td>0.738</td>
</tr>
<tr>
<td>consumer credit score</td>
<td>1 if no “slow” payments or delinquencies</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>2 if one or two slow payments or delinquencies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 if more than two slow payments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 if insufficient credit history for determination</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 if delinquent credit history with payments 60 days overdue</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 if delinquent credit history with payments 90 days overdue</td>
<td></td>
</tr>
<tr>
<td>mortgage credit score</td>
<td>1 if no late mortgage payments</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>2 if no mortgage payment history</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 if one or two late mortgage payments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 if more than two late mortgage payments</td>
<td></td>
</tr>
<tr>
<td>public bad credit record</td>
<td>1 if any public record of credit problems (bankruptcy, charge-offs, collection actions)</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>0 otherwise</td>
<td></td>
</tr>
</tbody>
</table>
**Additional Applicant Characteristics**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>denied mortgage insurance</td>
<td>1 if applicant applied for mortgage insurance and was denied, 0 otherwise</td>
<td>0.020</td>
</tr>
<tr>
<td>self-employed</td>
<td>1 if self-employed, 0 otherwise</td>
<td>0.116</td>
</tr>
<tr>
<td>single</td>
<td>1 if applicant reported being single, 0 otherwise</td>
<td>0.393</td>
</tr>
<tr>
<td>high school diploma</td>
<td>1 if applicant graduated from high school, 0 otherwise</td>
<td>0.984</td>
</tr>
<tr>
<td>unemployment rate</td>
<td>1989 Massachusetts unemployment rate in the applicant’s industry</td>
<td>3.8</td>
</tr>
<tr>
<td>condominium</td>
<td>1 if unit is a condominium, 0 otherwise</td>
<td>0.288</td>
</tr>
<tr>
<td>black</td>
<td>1 if applicant is black, 0 if white</td>
<td>0.142</td>
</tr>
<tr>
<td>deny</td>
<td>1 if mortgage application denied, 0 otherwise</td>
<td>0.120</td>
</tr>
</tbody>
</table>
### Table 11.2  Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: $\text{deny} = 1$ If mortgage application is denied, $= 0$ if accepted; 2380 observations.

<table>
<thead>
<tr>
<th>Regression Model Regressor</th>
<th>LPM (1)</th>
<th>Logit (2)</th>
<th>Probit (3)</th>
<th>Probit (4)</th>
<th>Probit (5)</th>
<th>Probit (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>black</strong></td>
<td>0.084** (0.023)</td>
<td>0.688** (0.182)</td>
<td>0.389** (0.098)</td>
<td>0.371** (0.099)</td>
<td>0.363** (0.100)</td>
<td>0.246 (0.448)</td>
</tr>
<tr>
<td><strong>P/I ratio</strong></td>
<td>0.449** (0.114)</td>
<td>4.76** (1.33)</td>
<td>2.44** (0.61)</td>
<td>2.46** (0.60)</td>
<td>2.62** (0.61)</td>
<td>2.57** (0.66)</td>
</tr>
<tr>
<td><strong>housing expense-to-income ratio</strong></td>
<td>−0.048 (0.110)</td>
<td>−0.11 (1.29)</td>
<td>−0.18 (0.68)</td>
<td>−0.30 (0.68)</td>
<td>−0.50 (0.70)</td>
<td>−0.54 (0.74)</td>
</tr>
<tr>
<td><strong>medium loan-to-value ratio</strong> (0.80 ≤ loan-value ratio ≤ 0.95)</td>
<td>0.031* (0.013)</td>
<td>0.46** (0.16)</td>
<td>0.21** (0.08)</td>
<td>0.22** (0.08)</td>
<td>0.22** (0.08)</td>
<td>0.22** (0.08)</td>
</tr>
<tr>
<td><strong>high loan-to-value ratio</strong> (loan-value ratio ≥ 0.95)</td>
<td>0.189** (0.050)</td>
<td>1.49** (0.32)</td>
<td>0.79** (0.18)</td>
<td>0.79** (0.18)</td>
<td>0.84** (0.18)</td>
<td>0.79** (0.18)</td>
</tr>
<tr>
<td><strong>consumer credit score</strong></td>
<td>0.031** (0.005)</td>
<td>0.29** (0.04)</td>
<td>0.15** (0.02)</td>
<td>0.16** (0.02)</td>
<td>0.34** (0.11)</td>
<td>0.16** (0.02)</td>
</tr>
<tr>
<td><strong>mortgage credit score</strong></td>
<td>0.021 (0.011)</td>
<td>0.28* (0.14)</td>
<td>0.15* (0.07)</td>
<td>0.11 (0.08)</td>
<td>0.16 (0.10)</td>
<td>0.11 (0.08)</td>
</tr>
<tr>
<td><strong>public bad credit record</strong></td>
<td>0.197** (0.035)</td>
<td>1.23** (0.20)</td>
<td>0.70** (0.12)</td>
<td>0.70** (0.12)</td>
<td>0.72** (0.12)</td>
<td>0.70** (0.12)</td>
</tr>
<tr>
<td><strong>denied mortgage insurance</strong></td>
<td>0.702** (0.045)</td>
<td>4.55** (0.57)</td>
<td>2.56** (0.30)</td>
<td>2.59** (0.29)</td>
<td>2.59** (0.30)</td>
<td>2.59** (0.29)</td>
</tr>
</tbody>
</table>
### Table 11.2, ctd.

<table>
<thead>
<tr>
<th></th>
<th>0.060*** (0.021)</th>
<th>0.67*** (0.21)</th>
<th>0.36*** (0.11)</th>
<th>0.35*** (0.11)</th>
<th>0.34*** (0.11)</th>
<th>0.35*** (0.11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>self-employed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>single</td>
<td>0.23** (0.08)</td>
<td>0.23** (0.08)</td>
<td>0.23** (0.08)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high school diploma</td>
<td></td>
<td></td>
<td></td>
<td>−0.61*** (0.23)</td>
<td>−0.60* (0.24)</td>
<td>−0.62** (0.23)</td>
</tr>
<tr>
<td>unemployment rate</td>
<td>0.03 (0.02)</td>
<td>0.03 (0.02)</td>
<td>0.03 (0.02)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>condominium</td>
<td></td>
<td></td>
<td></td>
<td>−0.05 (0.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black × P/I ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>−0.58 (1.47)</td>
<td></td>
</tr>
<tr>
<td>black × housing expense-to-income ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.23 (1.69)</td>
</tr>
<tr>
<td>Additional credit rating indicator variables</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>constant</td>
<td>−0.183*** (0.028)</td>
<td>−5.71*** (0.48)</td>
<td>−3.04*** (0.23)</td>
<td>−2.57** (0.34)</td>
<td>−2.90** (0.39)</td>
<td>−2.54** (0.35)</td>
</tr>
</tbody>
</table>

(Table 11.2 continued)
Table 11.2, ctd.

(Table 11.2 continued)

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

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
  • what else is learned in the in-person interviews?
  2. functional form misspecification (no…)
  3. measurement error (originally, yes; now, no…)
  4. selection
    • random sample of loan applications
    • define population to be loan applicants
  5. simultaneous causality (no)

• External validity
  This is for Boston in 1990-91. What about today?
Summary
(SW Section 11.5)

• If \( Y_i \) is binary, then \( E(Y|X) = \Pr(Y=1|X) \)

• Three models:
  o linear probability model (linear multiple regression)
  o probit (cumulative standard normal distribution)
  o logit (cumulative standard logistic distribution)

• LPM, probit, logit all produce predicted probabilities

• Effect of \( \Delta X \) is change in conditional probability that \( Y=1 \). For logit and probit, this depends on the initial \( X \)

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