Regression with a Binary Dependent Variable

(SW Chapter 11)

<u>Part I: The Linear Probability Model</u>

Binary Dependent Variables: What's Different?

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

- district-wide average test score
- traffic fatality rate
- income

What if Y is binary?

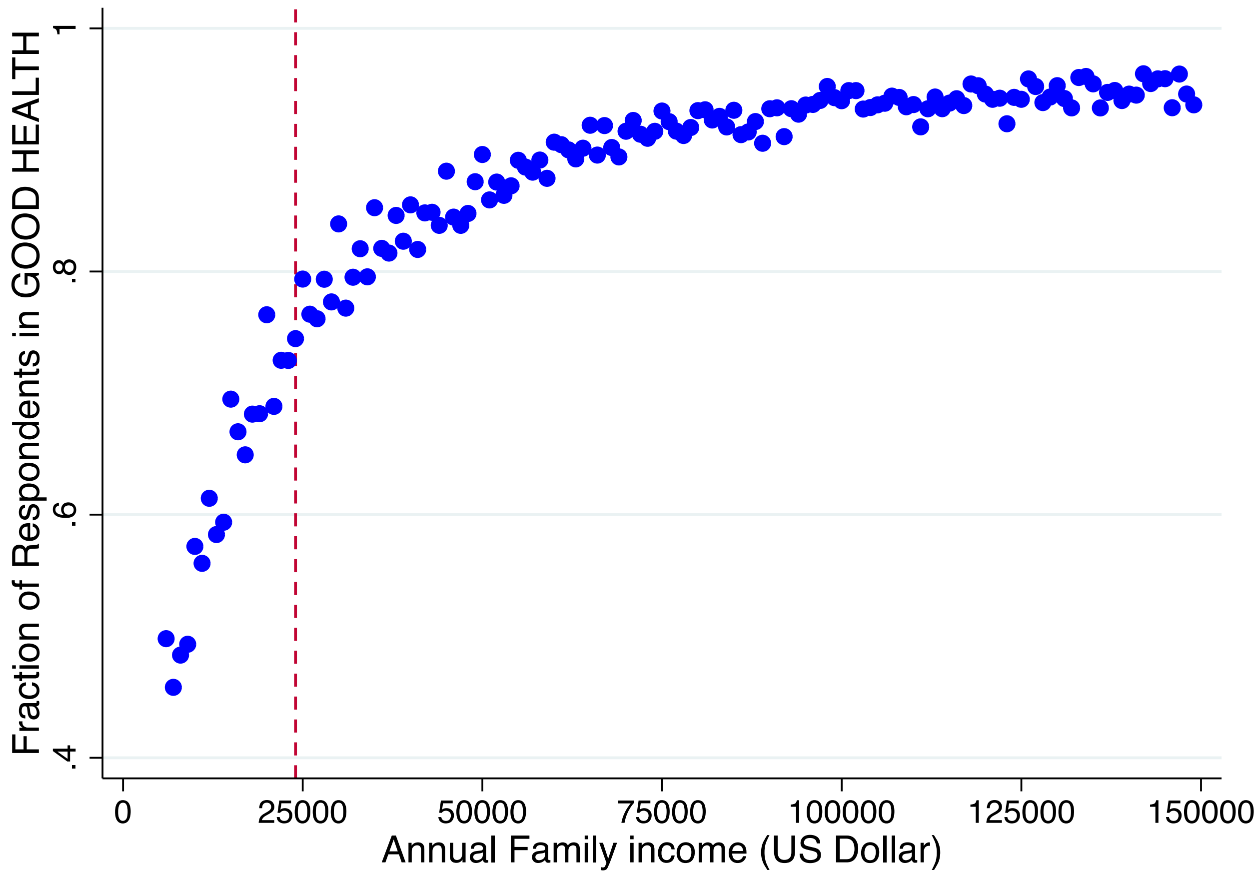
- Y = get into college, or not; X = high school grades, SAT scores, demographic variables
- Y = person smokes, or not; X = cigarette tax rate, income, demographic variables
- Y = respondent is in good health, or not; X = income

Influenza Pregnancy Infections and Birth Outcomes

Dependent variable	Gestation length (wks) (1)	Prematurity (<37 wks) (2)	Birth weight (gr) (3)	Low birth weight (<2500 gr) (4)
A. Baseline controls (no mother FEs)				
Influenza during pregnancy	-0.529***	0.059***	-150.322***	0.061***
	[0.056]	[0.008]	[16.169]	[0.007]
B. Baseline controls + mother FEs				
Influenza during pregnancy	-0.319***	0.045***	-84.483***	0.035***
	[0.091]	[0.010]	[22.238]	[0.011]
N	460,618	460,618	459,987	459,987
Mean dep. var.	39.7	0.042	3,461	0.039

Example: Income and Health

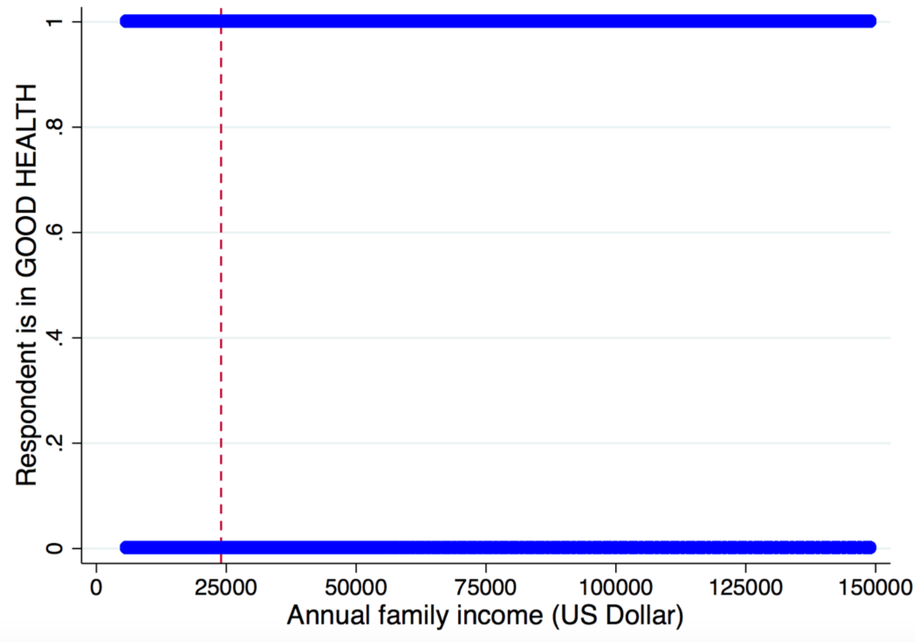
- You have seen this figure in the first lecture



Example: Income and Health, continued

- Being in good health is a binary variable: The respondent is either in good health or not.
- Why does it look like a continues variable in the scatter plot?

Example: Income and Health, continued

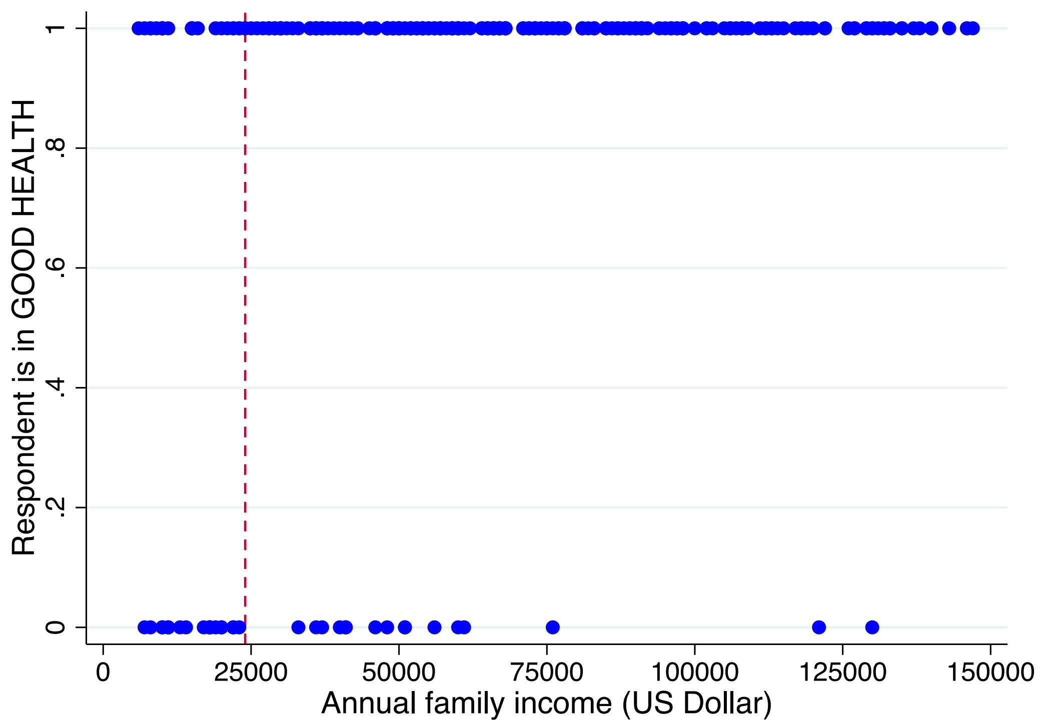


N= 219,338

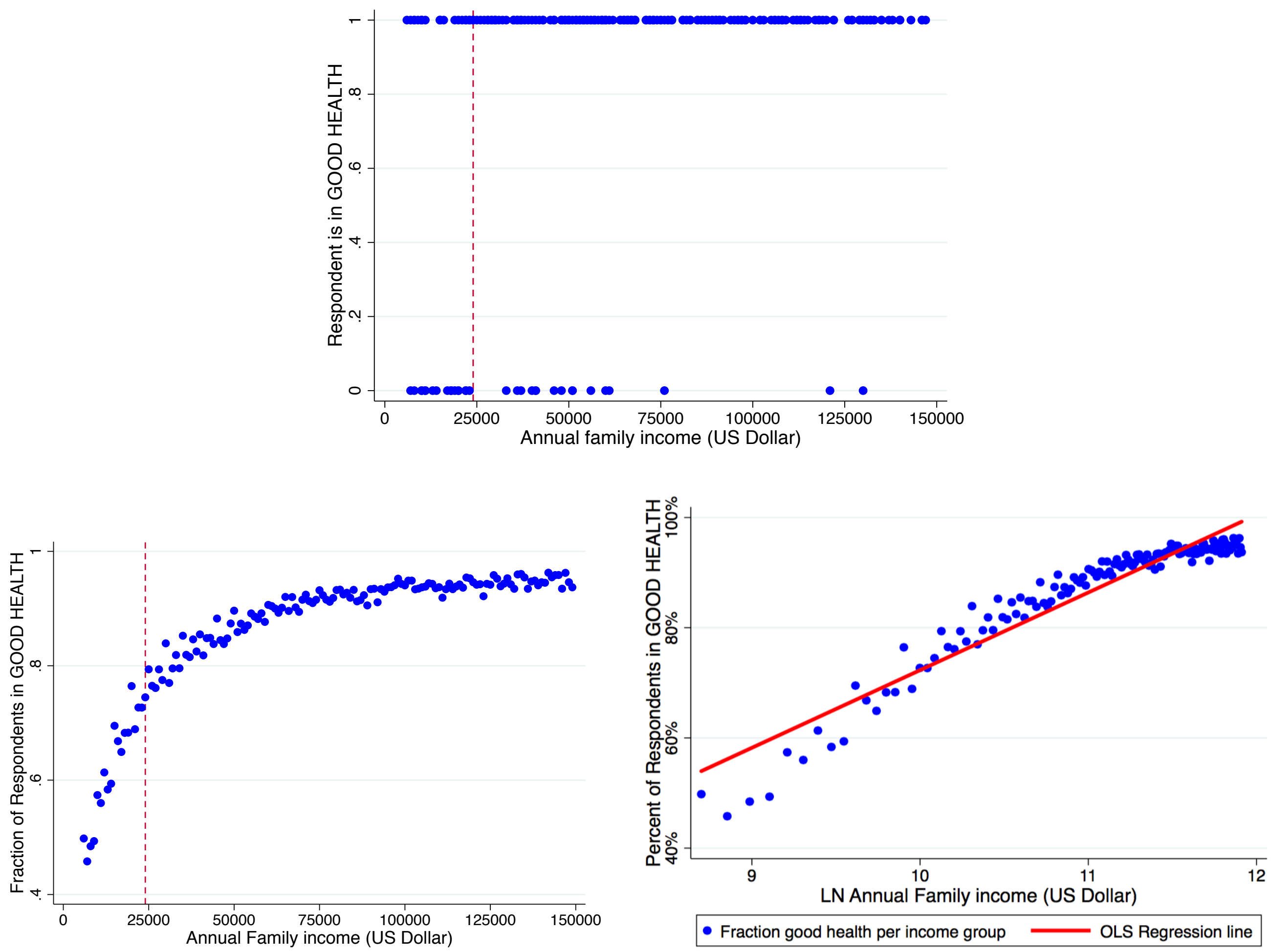
- Wait, I don't see anything! Why?

Example: Income and Health, continued

- Let's take a random subsample of 200 respondents:



Example: Income and Health, continued



How to interpret the effect of income in this figure?

Example: Mortgage Denial and Race
The Boston Fed HMDA Dataset

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

Variables

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

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

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

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

But:

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

The linear probability model, ctd.

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

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

When Y is binary,

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

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

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

so

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

The linear probability model, ctd.

When Y is binary, the linear regression model

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

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

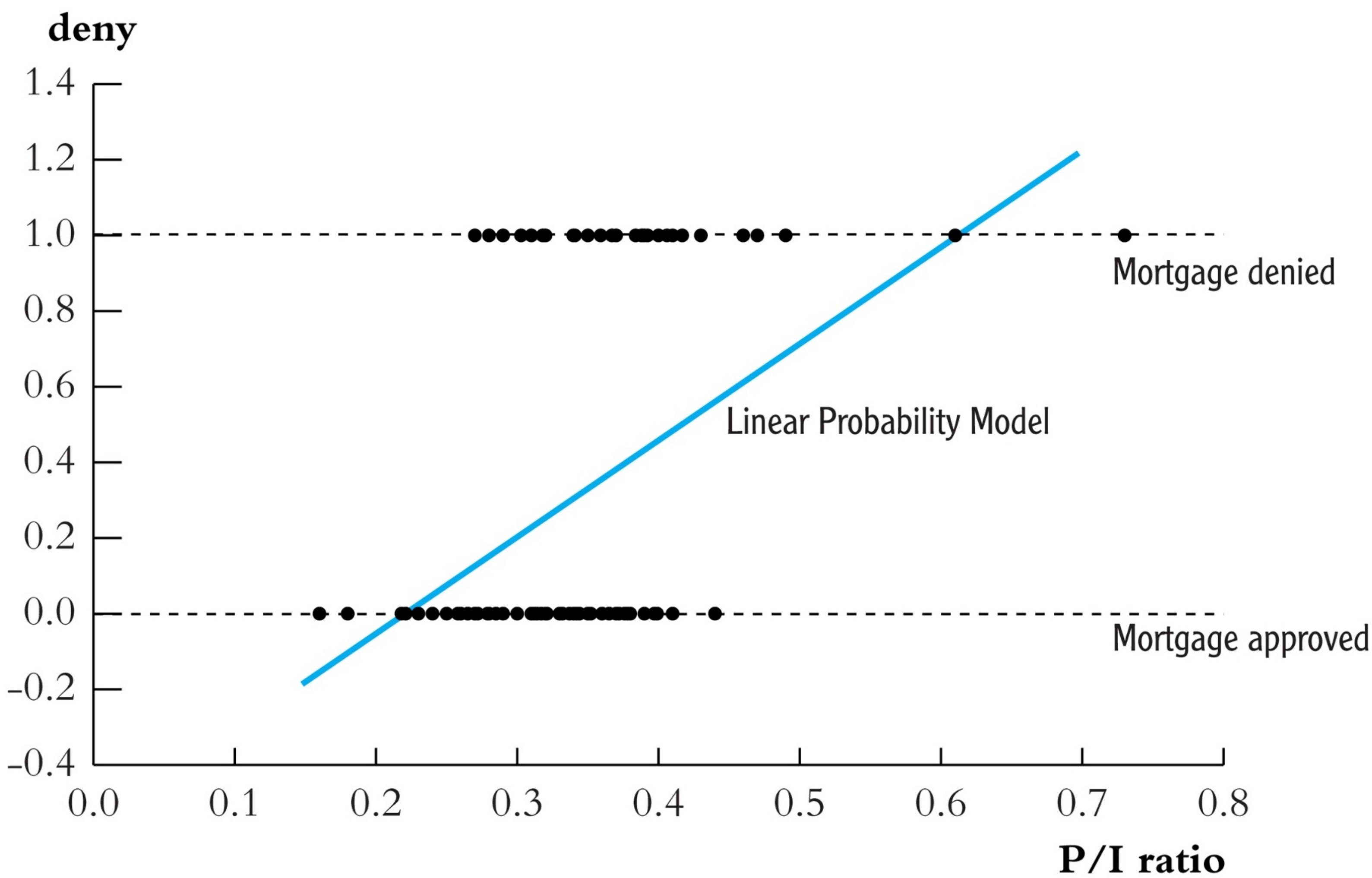
$$\Pr(Y=1|X) = \beta_0 + \beta_1 X_i$$

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

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

Example: linear probability model, HMDA data

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



Linear probability model: full HMDA data set

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

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

  $\Pr(deny = 1 \mid P/Iratio = .3) = -.080 + .604\times.3 = \mathbf{.1012}$

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

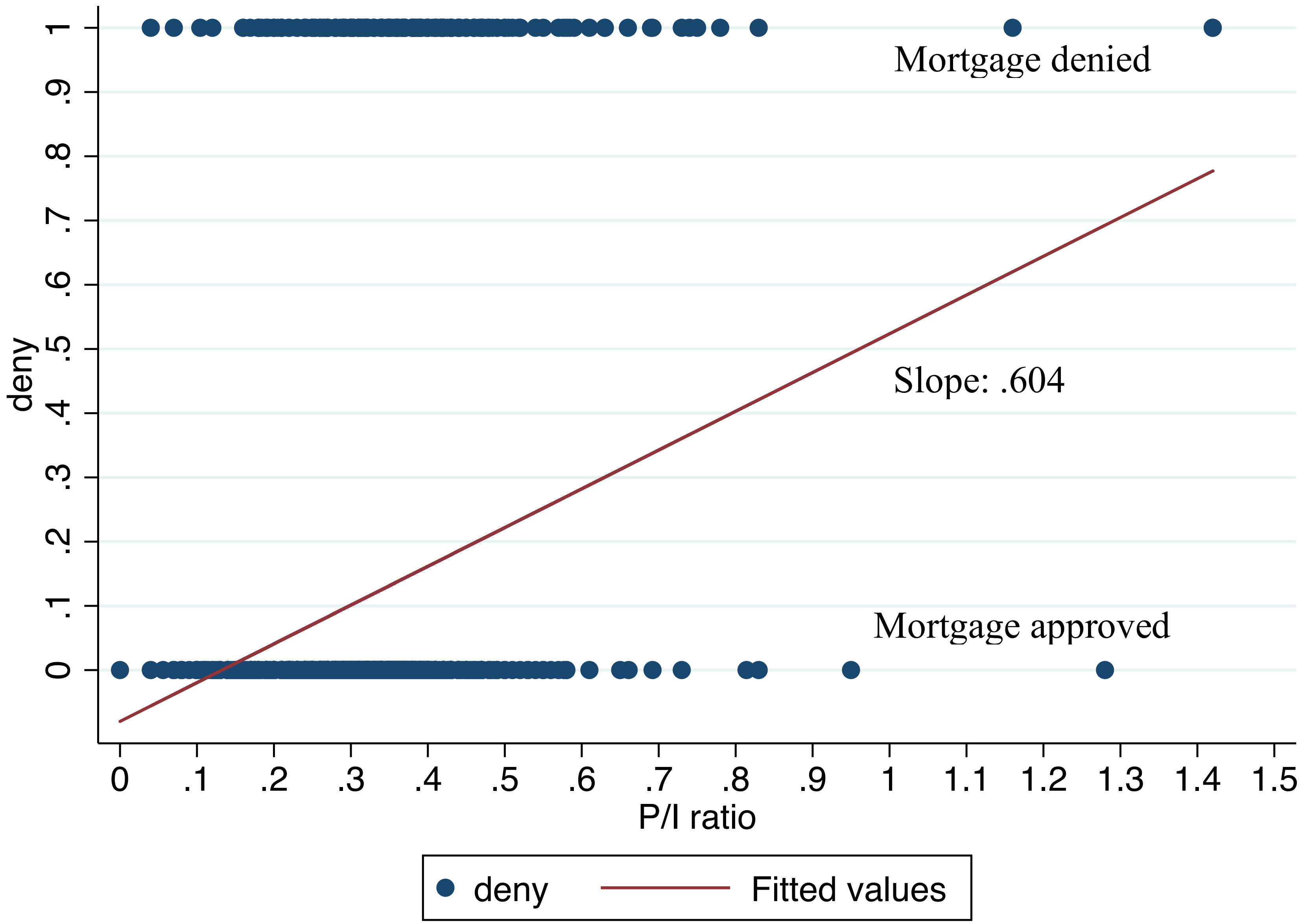
  $\Pr(deny = 1 \mid P/Iratio = .4) = -.080 + .604\times.4 = \mathbf{.1616}$

  The effect on the probability of denial of an increase in *P/I ratio* from .3 to .4 is to increase the probability by **.0604**, that is, by 6 *percentage points*.

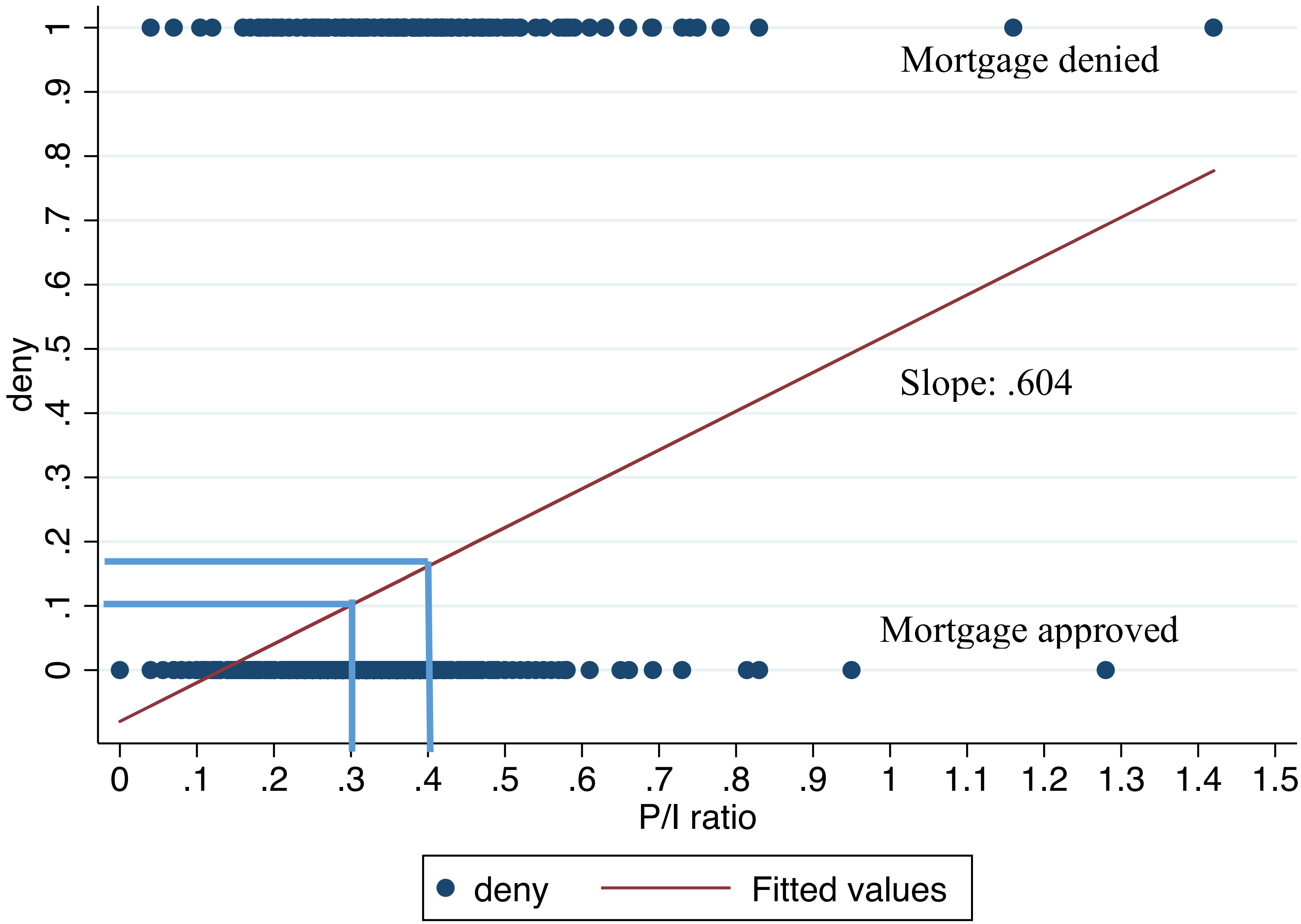
Figure with full HMDA data set

- The figure on slide 7 showed the linear probability model only for a subset of the HMDA data set (so that the individual data points can be better seen).
- But the regression line in the subsample does not correspond to the regression in the full sample
- Let's repeat the figure for the overall sample.
- Stata code:

```
use http://fmwww.bc.edu/ec-p/data/stockwatson/hmda_sw, clear

gen deny = (s7==3)
gen pi_rat = s46/100
gen black = (s13==3)

regress deny pi_rat, r
predict d_hat

twoway (scatter deny pi_rat if d_<=1) (line d_hat pi_rat if d_<=1),
ylabel(0(.1)1) xlabel(0(.1)1.5) graphregion(color(white)) ///
ytitle("deny") xtitle("P/I ratio")
```



Mortgage denied
Slope: .604
Mortgage approved
deny
P/I ratio
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1 1.1 1.2 1.3 1.4 1.5
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1
deny
Fitted values


Mortgage denied
Slope: .604
Mortgage approved
deny
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
1.1
1.2
1.3
1.4
1.5
P/I ratio
deny
Fitted values

Percentage point vs percent

- The probability to be denied increases from 10.12% to 16.16%
  → That's an increase by 6.04 percentage points
- The average probability to be denied in the sample is 11.97%.
- Compared to that average, the probability to be denied increases by 6.04/11.97 = 50.5% !

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:

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

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

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

- difference = .177 = 17.7 percentage points
- *What's that in percentage term (or "in relative terms")?*

The linear probability model: Summary

- The linear probability model models $\Pr(Y=1|X)$ as a linear function of X
- Advantages:
  - simple to estimate and to interpret
  - inference is the same as for multiple regression (need heteroskedasticity-robust standard errors)
- Disadvantages:
  - A LPM says that the change in the predicted probability for a given change in X is the same for all values of X (but the impact on acceptance might depend on X).
  - Also, LPM predicted probabilities can be <0 or >1!
- These disadvantages can be solved by using a *nonlinear* probability model: probit and logit regression