Economics 3140, Spring 2019

Empirical Project

Due: Tuesday, May 7th

Read "Mortgage Lending in Boston: Interpreting HMDA Data" by Alicia H. Munnell, Geoffrey M.B. Tootell, Lynn E. Browne, and James McEneany, published in the *American Economic Review* 86 (1996), pp. 25-53. It is o.k. to merely glance over the more technical aspects of the paper, but you should be able to follow large parts of it (including everything up to page 30 and much of section VI).

As additional, recommended but not required reading, I provide an accessibly written survey paper, "Evidence on Discrimination in Mortgage Lending" by Helen F. Ladd, published in the *Journal of Economic Perspectives* 12 (1998), pp. 41-62. The paper recounts the discussion spawned by an earlier version of Munnell et al. One consequence of that discussion was to audit the data for mistakes. The empirical project is based on a subset (corresponding only to white and black applicants using a single-family residence as collateral) of the corrected data. This subsample selection also responds to a criticism of earlier versions of the paper, which pooled black and hispanic applicants as well as different types of residences.

1 Briefly summarize the *economic* substance of the paper (not the econometrics and not the details on data collection and quality) in your own words. What is the question posed? What type of data are being considered? What results existed previously in the literature, and why is their validity described as doubtful? What answer do the authors ultimetaly give to their question?

2 We will replicate this paper. Data and *codebook* (a file explaining variable definitions etc.) are uploaded.

A feature of real-world data is that often, some preparatory steps are needed to get the exact variables you want. For example, an indicator variable of a mortgage application being denied does not exist in the original data because the relevant outcome variable takes 5 values (denied, accepted but rejected by applicant, etc.). After consulting the codebook myself, I generated the *denied* variable as follows:

```
gen denied = 0
replace denied = 1 if s7 == 3.
```

Explain why these commands will generate the variable you want. Verify that the commands generate the *denied* variable that you found in the dta-file. (Of course, you may alternatively generate the variable in R, MATLAB,...)

Using similar commands and consulting the codebook, generate an indicator variable *black* that takes the value 1 for black applicants and the value 0 for white ones.

3 Provide a cross-tabulation of *denied* and *black*. What fraction of black applicants in the sample received a rejection, and what fraction of white applicants? Using your answer, provide estimators $(\hat{\beta}_0, \hat{\beta}_1)$ for the equation

$$denied = \beta_0 + \beta_1 black + u$$

without actually running the regression. Verify your results by running the regression.

4 Estimate the equation

$$denied = \beta_0 + \beta_1 PI + u,$$

where PI is the payment/income ratio. Give an economic interpretation for coefficient β_1. Do you prefer conventional or heteroskedasticity-robust standard errors? Is the coefficient significantly positive? Does the result make sense economically?

Attached to this exercise you find a scatterplot of *denied* versus PI with overlaid linear fit. Are your numbers consistent – to "eyeball accuracy" – with the plot?

5 What is the predicted value of denial if $PI = 20\%$? Given that denial is a binary variable, how do you interpret this predicted value?

What is the predicted value of denial if $PI = 10\%$? Does this value make sense? Explain and make a suggestion for how to modify the equation so that the problem cannot occur.

6 Implement your suggestion. That is, estimate your own modification of the equation. (There is more than one acceptable approach here.) Provide fitted values for the values of PI previously considered.

7 Use both OLS and your preferred other specification to estimate (an appropriate modification of)

$$denied = \beta_0 + \beta_1 PI + \beta_2 black + u.$$

Give an interpretation for β_2. Is the effect statistically significant? Is it large?

8 Explain why the estimator $\hat{\beta}_2$ from 7 should not be trusted to indicate a causal effect. Relate your answer to Munnell et al.'s criticism of some of the earlier literature.

9 We are now going to do something close to the paper. Use both OLS and your preferred other specification to regress *denial* on all of the following variables:

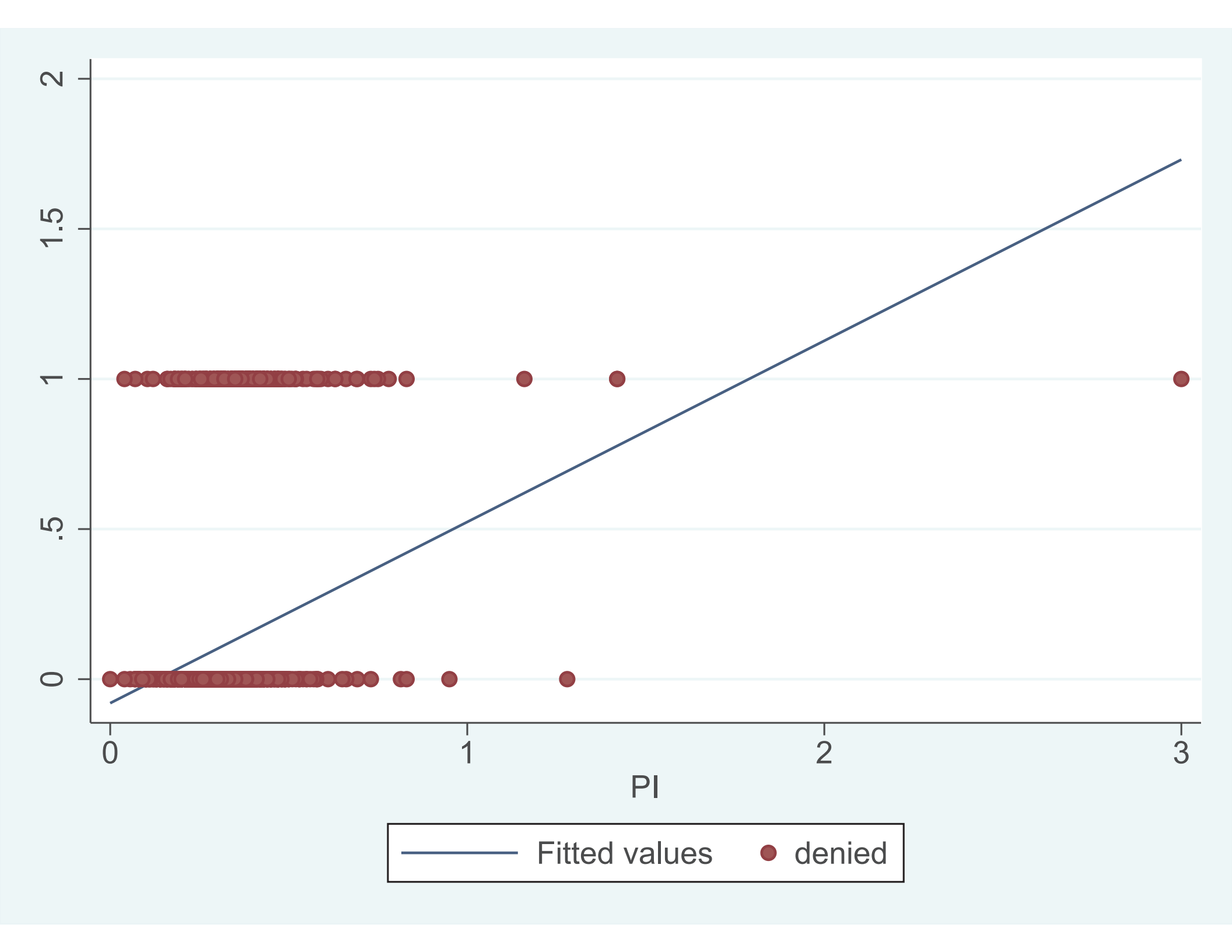
- $black$,
- PI,
- HI, which measures the ratio of housing expense to income,
- LV, which measures the ratio of loan to value,
- the square of LV,
- CCS, which is an ordinal indicator of the consumer credit score (values from 1 to 6 indicate increasingly strong problems),
- MCS, which is an ordinal indicator of the mortgage credit score (values from 1 to 4 indicate increasingly strong problems),
- $NoMI$, which is a dummy variable indicating whether mortgage insurance was sought after but denied,
- $self$, which is a dummy variable indicating self employment.

Comment on sign and significance of all coefficients. Ignoring the race indicator for the moment, do the other results make sense? Are results consistent across model specifications?

10 What is the predicted effect of race on the outcome variable in the OLS specification? Give a precise economic interpretation of this effect.

11 In your non-OLS specification, can you easily generate a number that is directly comparable to your answer to 10? Explain why or why not.

12 Compare your answer from part 10 to (i) the predicted effect of race across the two specifications for a hypothetical customer who is precisely average in the sample on all dimensions except race; (ii) the average predicted effect of race in the sample.


2
1.5
1
.5
0
0
1
2
3
PI
Fitted values
denied