EC 485 **Project 3** **Fall 2023**

This exercise follows the textbook example of modelling the decision to deny an applicant a mortgage and estimates the extent of racial discrimination in the decision.

Readings:
1) Chapter 11 of the textbook
2) Article 1: "Mortgage Lending in Boston: Interpreting HMDA Data" by Munnell, et. al., *American Economic Review*, March 1996, on Canvas . Pages 25-35.
3) Article 2: "The Role of Race in Mortgage Application Denials" by Lim and Ky, 2022, Federal Reserve Bank of Minneapolis.

PART 1: Summary of Munnell

a) *Carefully read sections I and II of the Munnell paper. Write a brief summary of the authors' research agenda as it relates to previous research. Make sure to address the issue of omitted variable bias.*

b) *The authors are estimating the probability that a loan is denied by the lender. The lender will deny a loan for a number of reasons, most important being when the lender expects that the applicant will default. Explain some of the variables the authors use to measure the probability of default. (Explain in your own words, do not pull verbatim from the article.*

c) *In addition to the probability of default, the lender is also concerned with the cost of the default because, when a borrower defaults, the lender takes possession of the property. What are some of the variables the authors used to capture the cost of default?*

PART 2: Analyze the data. Use the data set hmda_f22.csv for the analysis.

a) *What percentage of all loan applications were denied? What percentage of them were accepted?*

b) *Using the dummy variable RACE, separate the data into two groups: minority applications and white applicants.*

- *What percentage of the white applicants were denied?*
- *What percentage of minority applicants were denied?*

These percentages can be used as a crude predictor that a person in each respective group will be denied a mortgage. The difference in these probabilities brought attention to the issue and motivated the study. A multivariate model allows the authors to control for other determinants of the decision to deny a mortgage and thus remove some of the omitted variable bias to the estimate of discrimination in lending.

c) *Repeat the analysis in b) for the dummy variable MALE.*

- *What percentage of male applicants were denied?*
- *What percentage of female applicants were denied?*
- *Does there appear to be differences in loan denial based on gender?*

PART 3: Estimate a Linear Probability Model (OLS):

$$DENIED = \beta_0 + \beta_1 PI + \beta_2 RACE + \beta_3 CCS + \beta_4 LV + \beta_5 PBCR + \beta_6 URIA + \beta_7 SE + \beta_8 CONDO + \beta_9 SFAM + \beta_{10} MALE + \beta_{11} IRAJ + u_i$$

Munnell reports linear probability model results labeled "OLS" in Table 2

a) Is the variable PI statistically significant? Provide support.

b) What is the predicted effect on the probability of being denied a loan when the applicant's PI is 10 percentage points higher than a comparable loan applicant? (Pay attention to units)

c) What is the difference in predicted probability of denial for someone who has "bad credit" compared to someone who doesn't have bad credit, after controlling for other factors? Is the effect statistically significant?

d) What is the impact on the probability of being denied a loan if the applicant is applying for an adjustable rate mortgage? Is the effect statistically significant? Can you make sense of the sign on the estimated coefficient?

e) What is the difference in the predicted probability of denial for the two racial groups after controlling for other factors? Is the effect statistically significant? How does it compare to the value reported by Munnell in table 2?

f) Examine your predicted values for this regression. The fitted values are your predicted probabilities that a person is denied a mortgage. You will notice some of the predicted values for DENIED are outside of the [0,1] range.

- *Describe the person who has the lowest probability of being denied a loan (the most negative predicted probability) and the person who has the highest probability of being denied a loan.*
- *What are the predicted probabilities for these two individuals?*
- *Which factors do you think had the biggest impact on these two predicted probabilities? You should compare some of the relevant data values for two applicants to the sample averages.*

PART 4 Read sections 1, 2 and skim sections 3 and 4 of the **Lim and Ky** paper.

a) *What time period do these authors study?*
b) *Does their data set appear to be much larger or smaller than the HMDA set that we are using?*
c) *What percentage of their sample was denied a mortgage?*
d) *They estimate a linear probability model; results are in Table 4. Summarize the differences between their main results and your/Munnell results for the linear probability model concerning the effect of race on mortgage denial. How has the estimate of bias changed over time?*

PART 5 Estimate the Probit Model:

Estimate this Probit model:

$$Pr(DENIED=1) = \Phi(\beta_0 + \beta_1 PI + \beta_2 RACE + \beta_3 CCS + \beta 4\, LV + \beta_5 PBCR + \beta_6 URIA + \beta_7 SE + \beta_8 CONDO + \beta_9 SFAM + \beta_{10} MALE + \beta_{11} IRAJ)$$

Where $\Phi(.)$ is the cumulative standard normal probability distribution.

a) Comment on the statistical significance of the coefficient estimates.
Are the same variables significant here as were significant in the linear probability model?

b) Use the code to construct DENIED_hat values; that is, if the predicted probability is greater than 0.5, the observation gets classified as "predicted to be denied" so DENIED_hat=1; otherwise DENIED_hat=0.

What percentage of the observations did the model generate correct predictions for being denied a mortgage ? Is this better than the flip of a coin?

PART 6 Marginal Effects

A: Continuous X variables:

In class we discussed how to calculate the marginal effects for continuous explanatory variables:

$$\frac{\partial \Pr(DENIED_i = 1)}{\partial PI_i} = \Phi'\left(\beta_o + \beta_1 PI_i + \beta_2 RACE_i + ...\right) \cdot \beta_1 = \phi\left(\beta_o + \beta_1 PI_i + \beta_2 RACE_i + ...\right) \cdot \beta_1$$

Where $\varphi(.)$ is the normal density function, which is the derivative of the cumulative normal distribution function. Although β_1 is a constant, $\varphi(\beta_o + \beta_1 PI_i + \beta_2 RACE_i + ...)$ is not a constant. It depends on the values of PI_i and all other independent variables. Technically, this marginal effect of the variable PI_i on the probability of Y is different for each observation in the data set. As discussed in class, we usually evaluate $\varphi(\beta_o + \beta_1 PI_i + \beta_2 RACE_i + ...)$ for each observation, take the average of these values (average of the `phi_z` values in the class code example) and then multiply by the estimate of β_1. This gives us the average change in the probability of denial from a one unit change in X. Since PI is measured in percentage points, let's ask what effect a 10 percentage point increase in PI would have on the probability of denial. This would be similar to above, except we allow $\Delta PI = 10$:

$$\frac{\partial \Pr(DENIED_i = 1)}{\partial PI_i} = \overline{\phi\left(\beta_o + \beta_1 PI_i + \beta_2 RACE_i + \beta_3 CCS + ...\right)} \cdot \beta_1 * \Delta PI$$

a.1 Calculate and report the marginal effect of a 10 percentage point increase in the PI ratio. Compare it to the similar marginal effect for the linear probability model. Are the two effects in the same ballpark? Show your work.

a.2 Calculate and report the effect of one more "slow pay" on your credit account, that is. $\Delta CCS = 1$. *Show your work.*

a.3 Calculate and report the effect of a 10 percentage point increase in the loan to value ratio. Show your work.

B. Dichotomous X Variables (dummy variables)

Marginal effects for dummy explanatory variables are not calculated in the same way. In order to correctly predict the difference in the probability of denial for the two groups, we want to evaluate the cumulative normal distribution two times: once when RACE=1 and once when RACE=0. Note, that this cumulative normal distribution also has the other independent variables as arguments. The typical procedure is to plug in the average values for each X. (*These are the averages over the entire sample, not over the two groups.*) We then take the difference of these two predicted probabilities. This difference is our estimate of the effect of RACE on the probability of loan denial, holding all other independent variables constant.

$$\Phi\left(\hat{\beta}_o + \hat{\beta}_1 \overline{PI} + \hat{\beta}_2 1 + \hat{\beta}_3 \overline{CCS} + ...\right) - \Phi\left(\hat{\beta}_o + \hat{\beta}_1 \overline{PI} + \hat{\beta}_2 0 + \hat{\beta}_3 \overline{CCS} + ...\right)$$

This is the difference in predicted probabilities due to race: $\hat{P}^{RACE=1} - \hat{P}^{RACE=0}$ where the predictions are made at the sample means for all independent variables other than RACE.

You can code this in R or compute the effects in EXCEL using the function =NORMSDIST(z) where we evaluate the function twice, using:

$$z_1 = \left(\hat{\beta}_o + \hat{\beta}_1 \overline{PI} + \hat{\beta}_2 1 + \hat{\beta}_3 \overline{CCS} + ...\right) \quad \text{and} \quad z_0 = \left(\hat{\beta}_o + \hat{\beta}_1 \overline{PI} + \hat{\beta}_2 0 + \hat{\beta}_3 \overline{CCS} + ...\right)$$

where the z_1 and z_o values are calculated in EXCEL using the estimated coefficients and the appropriate sample means. For variables other than RACE, you will change the calculation of the z1 and zo values accordingly.

b.1 *Calculate and report this difference in probability of denial for RACE. Show your calculations. Is this value similar to the value found by using the linear probability model? Is there statistical evidence of discrimination in mortgage lending? How does your answer compare to the logit result reported by Munnell in table 2?*

b.2 *The variable IRAJ is also a dummy variable. Calculate and report the difference in the probability of denial for when the applicant is applying for an adjustable rate mortgage. Is this value similar to the value found by using the linear probability model?*

b.3 *The variable SE is also a dummy variable. It equals one for applicants who are self-employed. Calculate and report the difference in the probability of denial for self-employed applicants.*

CODE: Use the class example on labor force participation for the relevant R code for the analysis.

Hand-in: CODE that produces your output, your output, and answers, including any calculations that were done to arrive at your answers (such as Excel calculations for part 6). When providing an answer, use complete sentences. When making a claim of statistical significance/insignificance, provide support by stating a t-statistic or a p-value. When doing a calculation, show your work.

The data set contains the following variables for 2,835 loan applicants.

denied	= 1 if the applicant was denied a loan, = 0 if accepted.
pi	Debt-to-income ratio(%) (100*total monthly obligations)/(monthly income) which includes the mortgage payment should the loan be accepted.
race	= 1 if the applicant is African-American or Hispanic, = 0 if white
ccs	credit history. This variable can take on values 1, 2, 3, 4, 5, 6, depending on the number of delinquent or "slow pay" accounts.
pbcr	=1 if the applicant has bad credit, with bankruptcies and/or collection actions
uria	= probability the applicant will become unemployed; depends on industry in which the applicant works.
se	= 1 if the applicant is self-employed, 0 otherwise
lv	= loan to value ratio calculated as the amount of the loan divided by the appraised value of the property.
sfam	=1 if the property is a single family home; = 0 otherwise
multi	= 1 if the property was a multi-family unit; = 0 otherwise
condo	=1 if the property was a condominium unit; = 0 otherwise
male	= 1 if applicant is a male, =0 if female
iraj	=1 if adjustable rate mortage, = 0 if fixed rate mortgate