Machine Learning for Public Policy - Mini-Project 1 1

The University of Chicago - Harris School of Public Policy
PPHA 30545 - Professors Clapp and Levy
Winter 2025

This assignment must be handed in via Gradescope on Canvas by **11:45pm Central Time on Thursday, January 30rd.** There will be separate Gradescope assignments for R and Python students. Please be sure to submit to the version that matches the coding language of the lab section you are enrolled in.

You are welcome (and encouraged!) to form study groups (of no more than 2 students) to work on the problem sets and mini-projects together. But you must write your own code and your own solutions. Please be sure to include the names of those in your group on your submission.

You should format your submission in one of two ways:

1. As a single PDF containing BOTH a write-up of your solutions that directly integrates any relevant supporting output from your code (e.g., estimates, tables, figures) AND your code appended to the end of your write up. You may type your answers or write them out by hand and scan them (as long as they are legible). Your original code may be a R (*.rmd) or Python (*.py) file converted to PDF format. OR

2. As a single PDF of an R Markdown (*.rmd), Jupyter Notebook (*.ipynb), or Quarto (*.qmd) document with your your solutions and explanations written in Markdown.[1]

Regardless of how you format your answers, be sure to make it clear what question you are answering by labeling the sections of your write up well and assigning your answers to the appropriate question in Gradescope. Also, be sure that it is immediately obvious what supporting output from your code (e.g., estimates, tables, figures) you are referring to in your answers. In addition, your answers should be direct and concise. Points will be taken off for including extraneous information, even if part of your answer is correct. You may use bullet points if they are beneficial. Finally, for your code, please also be sure to practice the good coding practices you learned in Data and Programming and comment your code, cite any sources you consult, etc.

You are allowed to consult the textbook authors' website, Python documentation, and websites like StackOverflow for general coding questions. If you get help from a large language model (LLM) or other AI tool (e.g., ChatGPT), you must provide in the query string you used and an explanation of how you used the AI tool's response as part of your answer. You are not allowed to consult material from other classes (e.g., old problem sets, exams, answer keys) or websites that post solutions to the textbook questions.

[1] Converting a Jupyter Notebook to PDF is not always straightforward (e.g., some methods don't wrap text properly). Please ensure that your PDF is legible! We will deduct points if we cannot read your PDF (even if you have the correct answers in your Notebook).

1 Overview

After graduating from Harris, you are quickly hired to work for the President's Council of Economic Advisors (CEA).[2] The CEA is an agency within the Executive Branch that provides the President with objective advice to inform both domestic and international policy. According to its webpage, the "[CEA] bases its recommendations and analysis on economic research and empirical evidence, using the best data available to support the President in setting our nation's economic policy."

Your boss has asked you to conduct research using data from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) to predict the returns to education and inform policy. Your analysis will help shape your office's recommendations to the President and help set her education agenda.[3] The project has three parts: (1) obtaining data from the Internet, (2) cleaning that data, and (3) performing data analysis and answering questions.

2 Obtaining the Data

1. First, navigate to the IPUMS USA website: https://usa.ipums.org/usa/index.shtml.[4]
2. Choose "Browse and Select Data" from the menu on the left.
3. Choose "Select Samples" by clicking the light blue box.
4. Select the most current year of ACS data only. Do not include the 3 and 5-year versions of the data.[5] Then "Submit sample selections."
5. Now you get to go shopping for data.[6] Under "Select Harmonized Variables" ->
   (a) "Person" -> "Demographics," add the following to your cart
      i. AGE
      ii. SEX
      iii. MARST
   (b) "Person" -> "Family Interrelationship," add the following to your cart

[2] Your family is very proud and all of your friends are jealous of your great gig. You tell them you're so glad that you took Machine Learning, as it really helped you land the job.

[3] The ACS contains information similar to the Decennial Census Long Form Questionnaire that it replaced after the 2000 Census. It is an annual sample of one in 40 households in the country. For reference, every decade the Long Form sampled one in 6 households.

[4] Census Bureau datasets are notoriously difficult to download in usable forms. In order make the data more accessible, the wonderful people at the Institute for Social Research and Data Innovation at the University of Minnesota created the Integrated Public Use Microdata Series (IPUMS) which is an awesomely streamlined way to get your hands on the data you want. Note that that they make many additional datasets available for download via (for example) IPUMS International, IPUMS Global Health, and IPUMS Time Use, among others.

[5] In order to ensure large enough sample sizes to maintain confidentiality, the Census pools data over multiple years for geographic units with fewer people.

[6] This is like opening birthday presents for a data scientist!

i. NCHILD

(c) "Person" -> "Race, Ethnicity and Nativity," add the following to your cart

i. RACE
ii. HISPAN

(d) "Person" -> "Education," add the following to your cart

i. EDUC

(e) "Person" -> "Work," add the following to your cart

i. EMPSTAT

(f) "Person" -> "Income," add the following to your cart

i. INCWAGE

(g) "Person" -> "Veteran Status," add the following to your cart

i. VETSTAT

6. Click on the "View Cart" button. Check to make sure you got everything. Click on the "Create data extract" button.

7. Click on "Customize sample sizes."

   (a) Since we're dealing with a large sample from the national population, we have far more observations than we can easily process. Under "Households," enter "10" so the dataset you create has 10,000 households. This makes working with the data easier, but will still give us a "big data" dataset.[7]

   (b) Click "Submit."

8. Click on "Select cases."

   (a) Since we're looking at wages as a function of education, we're only going to keep those involved in the labor force. Select EMPSTAT. Just to be safe, let's also restrict our sample by age. Select AGE.

   (b) Click "Submit."

   (c) Check "Include only those persons meeting case selection criteria."

      i. Under EMPSTAT, check the box for "Employed" workers.
      ii. For AGE, select ages from 18 to 65.

   (d) Click "Submit."

9. To the right of "Data Format," click on "Change."

   (a) Select "Comma delimited (.csv)" or whatever your preferred format is.

[7] When you "Customize sample sizes," IPUMS will randomly draw 10,000 observations for you. Since this is a random process and the "Select cases" occurs after the random draw of observations, don't be worried if a study partner has a slightly (up to a few hundred) different number of observations.

(b) Select "Rectangular, person (default)."

(c) Click "Submit."

10. Give your extract a brief description.

11. Click on the box that says "Submit extract."

12. Request an account or sign in.

13. Finally, hit "Submit extract."[8]

14. Once your extract has been created, navigate to the IPUMS download page: https://usa.ipums.org/usa-action/data_requests/download.

    (a) Click on the "Download CSV" link (in the first column). Save the file to your hard drive.

    (b) Right-click on the "Basic" codebook file and save the *.cbk (text) file to your hard drive.

    (c) Unzip the data file and load the data in Python. For help unzipping a *.gz file (Unix's version of *.zip), check out "Step 2: Decompress the data file" here: https://usa.ipums.org/usa/extract_instructions.shtml (just note that the instructions are for the *.dat (text) file and you want the *.csv file).

3 Preparing the Data

1. First, take a few minutes to become familiar with the data.

2. For our analysis, we'll need to use the codebook we saved to clean and create a few variables.

    (a) Education - We have a categorical measurement of education (*educd*). For some of our analysis, we need a continuous variable. Use the *educd* variable to create a continuous measure of education called *educdc* using the crosswalk at the end of this document. A *.csv version of the crosswalk is available on Canvas.

    (b) Dummy Variables - Create the following dummy variables:

        i. A dummy, *hsdip*, equal to 1 if the individual has a high school diploma (but not a bachelors or higher degree).[9]

[8] It will take the IPUMS system a little while to create your extract, so go take a break or work on something else. The IPUMS system will email you once your extract has been created. Try to contain your excitement over the fun data that you'll soon get to play with, lest friends and family think you're weird.

[9] Note: in the US, students traditionally graduate high school after completing the 12th grade. Some students who do not graduate from high school take equivalency exams and can earn a credential similar to a diploma that is known as a GED. In general, how one codes individuals with a GED or associates degree is a decision the researcher has to make based on the context of his/her research question. To keep things standard for the project, code these individuals as having a high school diploma.

ii. A dummy, $coldip$, equal to 1 if the individual has a four-year college diploma (i.e, a bachelor's or a higher degree that required earning a college diploma first).

iii. A dummy, $white$, equal to 1 if the individual is white.

iv. A dummy, $black$, equal to 1 if the individual is black.

v. A dummy, $hispanic$, equal to 1 if the individual is of Hispanic origin.

vi. A dummy, $married$, equal to 1 if the individual is married.

vii. A dummy, $female$, equal to 1 if the individual is female.

viii. A dummy, vet, equal to 1 if the individual is a veteran.

(c) Interaction Terms - Create an interaction between each of the education dummy variables ($hsdip$ and $coldip$) and the continuous measure of education ($educdc$).

(d) Created Variables - Create the following

i. Age squared.

ii. The natural log of $incwage$.[10]

4 Data Analysis Questions

Now that the data is ready for analysis, please answer the following questions. Note that only your responses in this section will be graded directly, but they'll make it clear whether you obtained and prepared the data correctly.

1. Compute descriptive (summary) statistics for the following variables: $year$, $incwage$, $lnincwage$, $educdc$, $female$, age, age^2, $white$, $black$, $hispanic$, $married$, $nchild$, vet, $hsdip$, $coldip$, and the interaction terms. In other words, compute sample means, standard deviations, etc.

2. Scatter plot $\ln(incwage)$ and education (the continuous measure). Include a linear fit line. Be sure to label all axes and include an informative title.

3. Estimate the following model:

$$\begin{aligned}\ln(incwage) &= \beta_0 + \beta_1 educdc + \beta_2 female + \beta_3 age + \beta_4 age^2 \\ &+ \beta_5 white + \beta_6 black + \beta_7 hispanic \\ &+ \beta_8 married + \beta_9 nchild + \beta_{10} vet + \varepsilon,\end{aligned}$$

and report your results.

(a) What fraction of the variation in log wages does the model explain?

[10] Note that despite selecting on EMPSTAT = "Employed" and working age individuals, we may still observe some people who report having no income. Since we cannot take the natural log of a number less than or equal to zero (why not?), we have two options. The first is to just drop these these individuals. Another is to make an arbitrary adjustment to our log formula such as $\ln(incwage+1)$. Both are common techniques, and both have drawbacks. We'll go with the former, so you should drop any observations where $incwage \leq 0$ before taking the natural log.

(b) What is the return to an additional year of education?[11] Is this statistically significant? Is it practically significant? Briefly explain.

(c) At what age does the model predict an individual will achieve the highest wage?

(d) Does the model predict that men or women will have higher wages, all else equal? Briefly explain why we might observe this pattern in the data.

(e) Interpret the coefficients on the *white* and *black* variables and their significance.[12]

4. Graph $\ln(incwage)$ and education. Include three distinct linear fit lines specific to individuals with no high school diploma, a high school diploma, and a college degree. Be sure to label all axis and include an informative title.

5. The President asks you to determine a tool that can be used to predict wages for those considering a college degree so that future constituents can make informed decisions.

   (a) Write down a differential intercept and/or differential slope model of log wages that will allow the returns to education to vary by degree acquired (use the three categories in the previous question).[13] Be sure to include the controls from Question 3. Using theory, intuition, and/or common sense, explain/justify why you think your model is the best possible representation of the way the world works (in other words, why you think you are correctly modeling $f(X)$ but not over-fitting ε).

   (b) Estimate the model you proposed in the previous question and report your results.

   (c) Given your model estimates from the previous question, predict the wages of an 22 year old, female individual (who is neither white, black, nor Hispanic, is not married, has no children, and is not a veteran) with a high school diploma and an all else equal individual with a college diploma. Assume that it takes someone 12 years to graduate high school and 16 years to graduate college.

   (d) The President wants to know, given your results from the previous question, do individuals with college degrees have higher predicted wages than those without? By how much? Briefly explain.

   (e) The President gets excited by your results and is now considering legislation that will expand access to college education (for instance, by increasing student loan subsidies). Given the evidence provided by your model, would you advise the President to pursue this legislation?

   (f) What fraction of the variation in log wages does the model explain? How does this compare to the model you estimated in Question 3?

   (g) The President is concerned that citizens will be harmed (and voters unhappy) if the predictions from your model turn out to be wrong. She wants to know how confident you are in your predictions. Briefly explain.

[11] Hint: note that your answer should be in terms of wages, not log wages.

[12] Hints: you will probably need to look at how the RACE variable is coded on the IPUMS website to answer this question. Also, note that race is different from origin in the data, so "people of Hispanic origin may be of any race."

[13] These are known as "sheepskin" effects.

6. You remember that splines may be useful when generating predictions when you have non-linear relationships, such as with age. You hypothesize that there may be an increase in predictive power if we account for life milestones.

   (a) Estimate a model, keeping the other predictors, with a cubic spline in age and two knots: one at age 18 and another at age 65. Report the adjusted R^2.

   (b) Compare this adjusted R^2 to the analog from the regression in Question 3. Why are these different? Briefly explain.

   (c) We used theory to motivate where we placed our knots. (Not very machine learning of us.) To practice tuning model parameters (that are determined prior to fitting and affect model flexibility), experiment by estimating two models one with knots at 24 and 55, and another of your own choosing. Report the adjusted R^2 for each. Explain which model you prefer and why.

   (d) Splines use our data differently than traditional, more-linear regressions. Using the first model from Question 6c (with knots at 24 and 55), generate two predicted values for a female individual (who is neither white, black, nor Hispanic, is not married, has no children, and is not a veteran) with a college diploma: one prediction for the individual at age 17 and the other at age 50. Explain why these values are different.

Table 1: Crosswalk

educd	*educdc*
2	0
10	0
11	2
12	0
13	2.5
14	1
15	2
16	3
17	4
20	6.5
21	5.5
22	5
23	6
24	7.5
25	7
26	8
30	9
40	10
50	11
61	12
62	12
63	12
64	12
65	13
70	13
71	14
80	14
81	14
82	14
83	14
90	15
100	16
101	16
110	17
111	18
112	19
113	20
114	18
115	18
116	22