Technical Tutorial:
Creating a Proxy Means Test

Merritt Smith

Motivation for this Technical Tutorial

- This session will give you hands-on experience implementing basic machine learning methods on (simulated) household survey data
- Apply the tools you've heard us talk about
- Use the skills you already possess, or use this opportunity to learn a new language

Background: Proxy Means Test

Q: What is a proxy means test (PMT?)

A: A method to estimate the wealth of a household without directly asking about income. It instead uses observable characteristics of a household like what their roof is made out of or how many TVs they own. In this sense it's a "proxy" for actual income data. We use models to determine which questions to ask, and to map the relationship between answers to those questions and approximate income / consumption.v

Example:

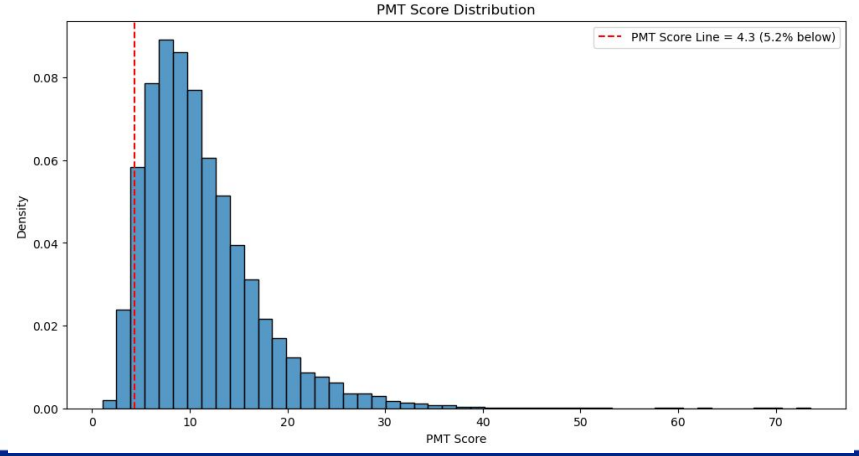
PMT Score = 2.1 + 0.3***number of rooms* + 0.8***electricity* + 0.5***refrigerator*+0.4***motorcycle*+0.6***household head years of education*

If PMT Score < 4.2, a household is eligible for aid.

Background: Proxy Means Test

Q: Why are they important?

A: They are commonly used to identify which families should receive government assistance like cash transfers. They help determine who qualifies as “poor enough” to receive aid. They help us allocate scarce aid to those who most need it, while being cheaper and harder to deceive than a detailed income survey.



Background: Proxy Means Test

Q: How do you construct a PMT?

A: Historically, they were created using expert knowledge and intuition to select variables, or simple regression models based on household surveys. Now, we use modern machine learning tools like train test splits and non-linear regression models.

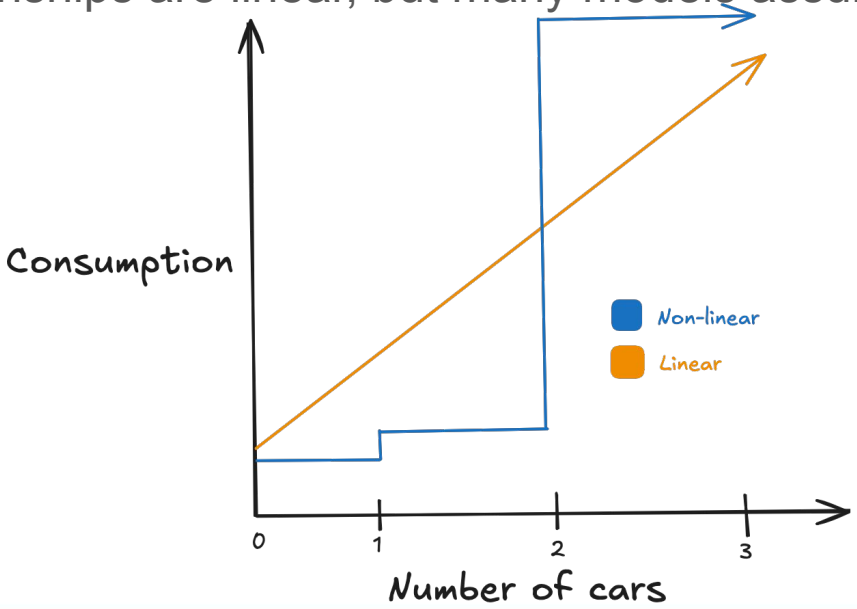
Q: Why is it important to use a train/test split?

A: In short, to prevent overfitting. If we use the full data set when we choose variables to include in our PMT, we risk the model working very well on the data we have, but not on real world data. Training on the train set and evaluating on the test set allows us to check how well our model can “generalize” to new, unseen data.

Background: Proxy Means Test

Q: Why are non-linear models important?

A: Not all relationships are linear, but many models assume they are



Background: Proxy Means Test

Q: What are other machine learning best practices that are important when creating PMTs?

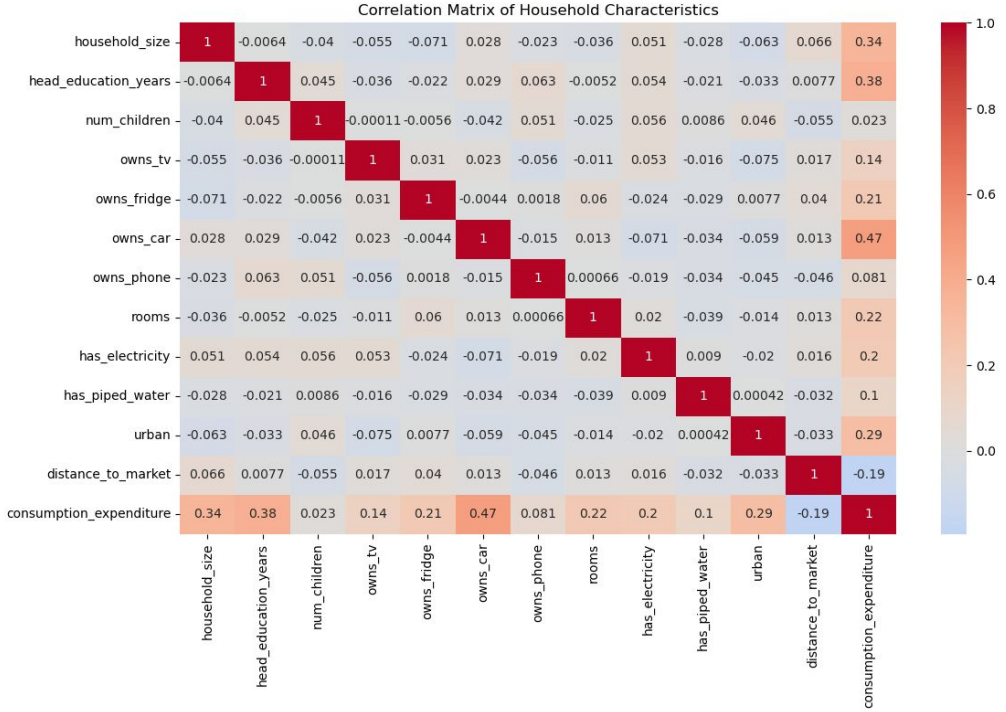
A:

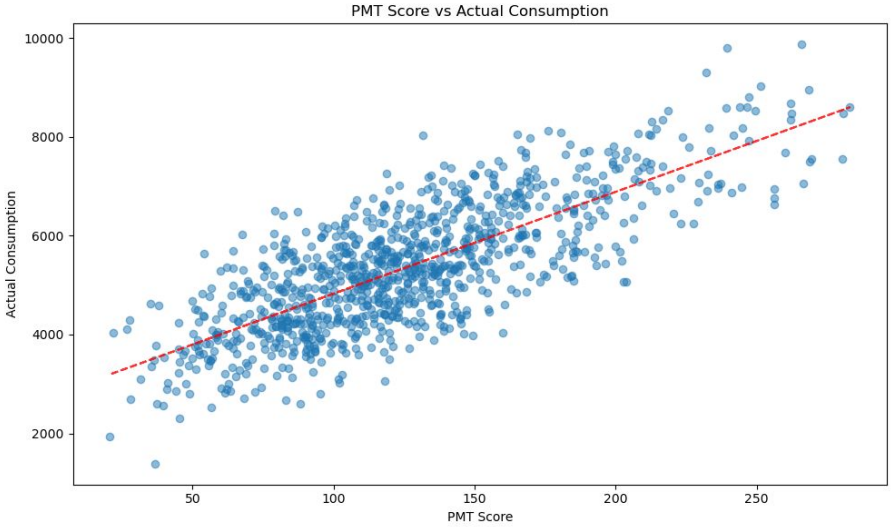
- Cross-validation, which offers the same benefits as a train/test split but allows us to use all our data
- Data pre-processing e.g. normalization, featurization
  - Normalization allows us to view all features on the same scale meaning that variables like land ownership in square feet don't necessarily dominate number of rooms
  - Featurization can help us make variables more interpretable or useful (as we will see in the CDR tutorial)
- Model updating, meaning that we should periodically re-train and re-evaluate our models because of changes in the real world
- Machine learning metrics beyond "accuracy". If 99% of my sample is rich, and I predict "rich" every time, I'm 99% accurate but not learning the underlying pattern

Overview of this Technical Tutorial

1. Synthetic Data Generation
2. Exploratory Data Analysis
    a. What columns do we have available?
    b. What is the target variable we are trying to predict?
    c. What are the distributions of each variable?
3. Model Development
    a. Is our train/test split “balanced”?
    b. How well can we predict consumption with all variables?
    c. What are the most important variables that we should include in our PMT?
4. PMT Scoring System
    a. How well can we predict consumption with our PMT variables?

A Sample of What We'll Achieve:





Let's Get Started!

- Together we'll create a PMT using simulated data
- Download the R markdown file "pmt_creation.Rmd" from the link below in the "PMT Activity" folder

https://shorturl.at/apCQx

Data Preparation

- Generate the simulated data
- (If using Google Sheets / Excel, this is done for you)

Exploratory Data Analysis

- What do you notice in each plot?
- What is unrealistic about the simulated data?
- What else would you want to know about the data?

Model Development

- Do the train and test set look “balanced”?
- What is unrealistic about the consumption distribution here?
- Are the most “important” features consistent with the way we generated the consumption variable?
- Are the most “important” features consistent with what you think indicates consumption and income in real life?

PMT Scoring System

- How might you change the scoring system, and how does that affect the PMT?
- How sensitive does the PMT seem to be to the scoring system?
- Why might we want a scoring system using “points” rather than a more complicated model?

PMT Scoring System

- How might you change the scoring system, and how does that affect the PMT?
- How sensitive does the PMT seem to be to the scoring system?
- Why might we want a scoring system using “points” rather than a more complicated model?

R Resources

- CS50R: Harvard's introduction to R course, available for free online here. Includes videos, slides, and an optional final assignment.
- swirl: An R package that teaches you R from within the R console itself, offering an interactive, hands-on learning experience.
- R for Data Science by Hadley Wickham and Garrett Grolemund:
  Written by key contributors to the R ecosystem, this book focuses on using the tidyverse for data analysis.
- An Introduction to Statistical Learning: With Applications in R by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani:
  A classic textbook for statistical learning with practical examples in R. Available for free online here.
- RStudio Education: A set of resources to begin learning R, with a clear pathway to learning more intermediate and advanced topics in R.

Python Resources

- CS50 Python: Harvard's introduction to Python course, available for free online here. Includes videos, slides, and an optional final assignment.

- An Introduction to Statistical Learning: With Applications in Python by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani:
  A classic resource for statistical learning with practical examples in Python. Available for free online here.

- Python for Data Analysis by Wes McKinney: Written by the creator of the pandas library, this book is great for understanding how to work with large datasets.

- Python Data Science Handbook by Jake VanderPlas: This comprehensive book provides a guide to essential Python data science tools like NumPy, pandas, and Matplotlib.

- Data Science from Scratch by Joel Grus: This book covers the fundamental concepts of data science by having you implement the algorithms from scratch using Python.

Stata Resources

- Statalist: The official forum for Stata users to engage in discussions about statistics and Stata.
- Stata Blog and Video Tutorials: The official Stata website offers short, focused video tutorials and blog posts on specific features and techniques.
- Stata Documentation and Web Resources: The official documentation is comprehensive, and there are web resources that provide step-by-step instructions.

General Coding Resources

- Leetcode: A repository of coding problems that varies in difficulty from introductory problems to problems people practice on for interviews at top tech companies

- Kaggle Learn: Kaggle is a site where people compete to see how well they can solve problems. Kaggle Learn is a set of courses they provide to get you from just starting out to competing on their platform

- Khan Academy: An enormous repository of lectures and courses, including foundational math and computer science.

The following are a set of
- freeCodeCamp
- Codecademy
- DataCamp

Massive Online Open Courses (MOOCs)

These are all platforms that offer open courses online for free:

- EdX
- Coursera
- Udacity
- MIT OpenCourseWare