Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning para Economía Aplicada
MECA 4107
Créditos: 4
2026-10

Problem Set 1: Predicting Income

1 Introduction

In the public sector, accurate reporting of individual income is critical for computing taxes and designing social policy. However, income misreporting remains widespread. According to the Internal Revenue Service (IRS), only about 83.6% of taxes are paid voluntarily and on time in the United States.[1] One of the drivers of this gap is the under-reporting of individual income. Models of labor income therefore play a central role both in understanding inequality and in designing tools for enforcement and targeting.

The objective of this problem set is to develop and evaluate models of individual labor income that could assist tax authorities in detecting potential income misreporting. This problem presents a fundamental tension in applied econometrics and machine learning: models that are interpretable and grounded in economic theory may sacrifice predictive accuracy, while flexible predictive models may lack the interpretability needed for policy implementation. This problem set navigates this trade-off systematically, illustrating how economic theory, data analysis, and predictive modeling complement each other in addressing real-world policy problems.

The problem set is organized into three sections that build progressively on each other. Section 1 establishes a baseline age–income profile grounded in human capital theory, which predicts non-linearities in the relationship between age and earnings. This theory-driven specification is interpretable and testable, but imposes strong functional form restrictions. Section 2 enriches this baseline by allowing for systematic heterogeneity across gender, introducing the challenge of distinguishing between "explained" and "unexplained" wage gaps, a classic problem in labor economics that requires careful consideration of which controls to include. Section 3 shifts the focus from interpretation to prediction, evaluating which specifications best predict out-of-sample income and analyzing which observations drive prediction error. This progression mirrors the practical workflow that a tax authority would follow: start with economic priors about income determination, incorporate observed heterogeneity, and finally develop predictive tools for targeting enforcement resources. Throughout, you will work with data from Bogotá's 2018 household survey, making economically-informed decisions about data construction and model specification.

2 General Guidelines

2.1 Deliverables

Each team must submit the following materials:

- **Slides for in-class presentation.** Each team must prepare **three separate slide decks**, one for each section. Each slide deck should be submitted as a `.pdf` file and uploaded

[1] See https://www.irs.gov/newsroom/the-tax-gap.

to the corresponding activity on Bloque Neón. Files must follow the naming convention: `section_equipo_XX.pdf`, where `section` is `age`, `gap`, or `pred`, and `XX` is the team number (use a leading zero for team numbers below 10). For example for Team 1 the files would be:

  - `age_equipo_01` (Section 1: Age–Labor Income Profile)
  - `gap_equipo_01` (Section 2: Gender Labor Income Gap)
  - `pred_equipo_01` (Section 3: Labor Income Prediction)

- **Public GitHub repository.** Each team must submit a link on Bloque Neón to a **public** GitHub repository containing all code required to reproduce the analysis and results presented in the slides. The repository must follow the provided template and show **at least five (5) substantial** from each team member on the main branch. Code must be fully reproducible, readable, and well-commented.

2.2 Presentation Format

Each team will deliver a **10-minute in-class presentation**. Teams must be prepared to present **any** of the three sections of the problem set. During class, one group and one presenter will be randomly selected to present a given section. All team members must be equally prepared to present all sections.

Presentations should clearly communicate the empirical analysis to an audience of economists. Each slide deck should tell a coherent empirical story. While teams are free to organize their slides, a **recommended** structure is: (1) the research question and main takeaway (1–2 slides), (2) data and section-specific descriptive statistics relevant for the analysis in this section (1–3 slides), and (3) specifications and results (2–4 slides).

Focus on interpretation and conclusions rather than mechanical reporting. Tables and figures should be self-contained, properly formatted, and publication-quality—not screenshots from software output.

2.3 Evaluation Criteria

A full rubric for the presentation and repository can be found in Bloque Neón.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning para Economía Aplicada
MECA 4107
Créditos: 4
2026-10

3 Data

3.1 Data Sources and Sample Construction

The data for this problem set are available at https://ignaciomsarmiento.github.io/GEIH2018_sample/. The website hosts microdata for Bogotá from the 2018 *Medición de Pobreza Monetaria y Desigualdad* report, based on the Gran Encuesta Integrada de Hogares (GEIH). The dataset is distributed in **10 separate chunks**, all of which must be scraped.

A central component of this problem set is the construction of a coherent and well-justified analysis sample. Teams are therefore responsible not only for scraping the data, but also for cleaning it and defining the final sample used throughout the problem set. Unless otherwise stated, the analysis should be restricted to individuals who **report being employed** and are **18 years of age or older**. The main outcome variable is **total monthly labor income**, which combines income from salaried work and self-employment (the corresponding variable in the dataset is `y_total_m`).

3.2 Data Challenges and Cleaning Decisions

The raw data contain missing values, zero incomes, and potentially implausible observations. Teams are expected to make explicit, economically-justified decisions regarding how these issues are handled.

There is no single "correct" way to address these challenges. However, all data construction choices should be clearly stated, defensible, and guided by economic reasoning. Since the same cleaned dataset will be used across all three sections, teams should consider how their cleaning decisions affect each empirical question: the estimation of life-cycle profiles (Section 1), the measurement of wage gaps (Section 2), and the prediction of income for tax enforcement purposes (Section 3).

3.3 Data Presentation Requirements

Each team should work with a single underlying cleaned dataset. However, the way the data are summarized and visualized **must vary across sections** of the presentation. In each section, teams should present **only those descriptive statistics and figures that are directly informative** about that section's research question and empirical strategy.

The data should not be introduced as a dry list of variables or cleaning steps, but as a set of empirical objects that motivate and support the subsequent analysis. Descriptive statistics should help the audience understand why certain modeling choices are appropriate and what patterns the inferential analysis aims to capture. For instance, Section 1 might emphasize how income, hours worked, and employment type vary over the life cycle, precisely the variables that will enter the conditional age-income profile. Section 2 could focus on gender differences in these dimensions (or others) to motivate the choice of controls in the wage gap specification. Section 3 might focus on sample characteristics that may later drive high-error observations. These are suggestions, not requirements: teams are encouraged to present descriptive evidence that they find most informative

for their empirical story. A deep and thoughtful presentation of the data helps the audience understand its structure, sources of variation, and how these empirical patterns inform your modeling decisions.

4 Section 1: Age–Labor Income Profile

A large body of evidence in labor economics suggests that labor income follows a predictable life-cycle pattern: income is typically low at young ages, rises with labor market experience, peaks in midlife, and then remains flat or declines thereafter. Human capital theory predicts this relationship should be concave, with earnings increasing at a decreasing rate as workers age. This section tests whether this theoretical prediction holds in Bogotá's labor market and establishes a baseline age–income profile that will serve as a benchmark for subsequent sections.

In this section, teams will estimate an *unconditional age–labor income profile* for individuals in the sample using the following specification:

$$\log(w) = \beta_1 + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + u,$$

where $\log(w)$ denotes the logarithm of **total monthly labor income**. The quadratic term allows the data to reveal whether the relationship between age and earnings is indeed concave, as theory suggests, and permits the profile to exhibit a peak within the observed age range. Because the peak age implied by the quadratic specification is a nonlinear function of the estimated coefficients, **confidence intervals for the implied peak age must be constructed using bootstrap**.

In addition, teams will estimate a *conditional age–labor income profile* by augmenting this specification with controls for labor supply and employment type. The conditional specification must include **only** total hours worked (`totalHoursWorked`) and employment type (`relab`). No additional controls should be included. Teams should think carefully about what the conditional specification reveals about the age–income relationship and how it differs from the unconditional profile.

4.1 Minimum Content for the Presentation

The slides for this section must include, at a minimum:

- A regression table that demonstrates the contribution of the quadratic term to model fit and the effect of including controls on the estimated age–income relationship. The table should report the implied peak age (with confidence intervals) and in-sample fit for each specification presented.
- Visualizations of the **unconditional** and **conditional** age–labor income profiles, showing the relationship between age and predicted log monthly labor income.

4.2 Interpretation and Discussion

When presenting results, teams should assess whether the empirical patterns support the theoretical predictions from human capital theory. A central prediction is that the age–income relationship is **non-linear** and concave. Does the estimated profile exhibit this pattern? Is there evidence of a peak in earnings within the observed age range?

How does conditioning on hours worked and employment type affect the estimated non-linearity? What does this reveal about the sources of the observed life-cycle pattern?

If the estimated profiles deviate from theoretical predictions, what features of Bogotá's labor market might explain these deviations? Interpret differences between specifications economically, not just statistically. The goal is to test whether the data support the theory's prediction of a non-linear age–income profile and to understand what drives the observed pattern.

5 Section 2: Gender–Labor Income Gap

Policymakers and researchers have long been concerned with gender differences in labor income. This section builds on the baseline age–income profile established in Section 1 by introducing systematic heterogeneity across gender. The analysis explores whether the age–income relationship differs between men and women and examines how much of the observed gender wage gap can be attributed to differences in observable characteristics versus unexplained factors.

As a starting point, teams will estimate the *unconditional gender labor income gap* using a specification that includes only a gender indicator:

$$\log(w) = \beta_1 + \beta_2 \text{Female} + u,$$

where β_2 captures the raw difference in log labor income between women and men.

A shortcoming of this specification is that it does not directly address the commonly invoked principle of *"equal pay for equal work"*. To explore this, teams will estimate *conditional gender labor income gap* specifications that control for worker and job characteristics. For these conditional specifications, the gender coefficient must be estimated using the Frisch–Waugh–Lovell (FWL) decomposition, partialling out the gender indicator from the chosen set of controls. Standard errors for the gender coefficient must be reported using both analytical formulas and bootstrap.

5.1 Minimum Content for the Presentation

The slides for this section must include, at a minimum:

- A regression table comparing unconditional and conditional gender labor income gap estimates. The table should demonstrate how the estimated gender coefficient changes across specifications and should report in-sample fit for each. Standard errors should be reported using both analytical and bootstrap methods.

- Using a preferred conditional specification, a visualization showing the predicted age–labor income profiles separately for men and women.
- The implied peak ages for each group (with confidence intervals) should be reported and discussed.

5.2 Interpretation and Discussion

When presenting results, the analysis should address how the estimated gender wage gap changes as controls are added. What does this pattern reveal about the sources of gender differences in earnings? Are these changes consistent with selection effects, discrimination, a combination of both, or neither? Does the *"equal pay for equal work"* principle hold once observable differences between men and women are taken into account?

Differences between analytical and bootstrap standard errors for the gender coefficient, if present, should be examined and discussed. What might explain any discrepancies, and which estimates are more appropriate for inference in this setting?

The choice of control variables is central to the analysis. The discussion should make clear which controls are included and why, with particular attention to the potential presence of *bad controls*.

Beyond the average gap, do men and women exhibit different age–income profiles? If the shapes differ—such as steeper earnings growth for one group or earlier or later peaks—what economic mechanisms might explain these patterns? Are the differences in peak ages statistically meaningful, and what do they reveal about gender differences in career trajectories?

The goal is to assess whether accounting for systematic differences between men and women can explain the observed gap in labor income, and to understand the role that gender plays in explaining variation in earnings in Bogotá's labor market.

6 Section 3: Labor Income Prediction

This section shifts the focus from interpretation to prediction. Building on the specifications estimated in Sections 1 and 2, the objective is to evaluate their out-of-sample predictive performance and to understand which observations are most difficult to predict. From the perspective of a tax authority seeking to identify potential income misreporting, this analysis highlights both the promise and the limitations of prediction-based approaches.

Predictive performance will be evaluated using a **validation set approach**. Chunks 1–7 should be used as the training sample, while chunks 8–10 should be reserved as a validation sample. Predictive accuracy is measured using the root mean squared error (RMSE) computed on the validation sample.

As a baseline, all specifications estimated in Sections 1 and 2 should be re-estimated on the training sample and evaluated on the validation sample. This establishes a reference point for out-of-sample predictive performance.

Teams will then estimate at least **five additional specifications** designed to improve out-of-sample prediction. These specifications must be economically motivated and should build on the structure of the income equations used in previous sections. Possible extensions include additional non-linearities, interactions between variables, or additional controls that economic theory suggests should predict income.

Based on validation-sample RMSE, the **best performing model** should be selected. For this model, a detailed diagnostic analysis should be conducted to understand the sources of prediction error. This includes computing the leave-one-out cross-validation (LOOCV) prediction error, comparing it to the validation-sample error, and examining which observations are most difficult to predict and how they influence the estimated coefficients.[2]

For the selected model, observation-level influence on the coefficient vector should be measured as

$$\|\hat{\beta}_{(-i)} - \hat{\beta}\|_2,$$

where $\hat{\beta}_{(-i)}$ denotes the estimated coefficients when observation i is omitted. This calculation must be implemented using the Frisch–Waugh–Lovell decomposition.[3]

6.1 Minimum Content for the Presentation

The slides for this section must include, at a minimum:

- A table reporting validation-sample RMSE for all estimated specifications, including those from Sections 1 and 2.
- A comparison of LOOCV error and validation-sample error for the best performing model.
- One or more figures and/or tables analyzing where the best performing model fails predictively, summarizing observation-level leave-one-out prediction errors and coefficient-influence measures.

6.2 Interpretation and Discussion

The discussion should address why the best performing model outperforms alternative specifications. What role do non-linearities, interactions, or additional controls play in improving predictive performance? How does the LOOCV error compare to the validation-sample error, and what does this comparison reveal about overfitting, model stability, or the representativeness of the validation sample?

[2] Note: when attempting the computation of leave-one-out cross-validation (LOOCV) prediction error, the calculations can take a long time, depending on your coding skills, so plan accordingly!

[3] Warning: this norm is not invariant to rescaling of the regressors. As a result, the magnitude of the influence measure reflects both the observation's leverage and the scaling of the regressors. You should therefore standardize the regressors or otherwise justify your choice of scale.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning para Economía Aplicada
MECA 4107
Créditos: 4
2026-10

A central component of the analysis is understanding where the model fails predictively. Which observations exhibit the largest prediction errors, and do they share common characteristics? How do high-error observations influence the estimated coefficients when they are omitted, and which coefficients are most sensitive to exclusion? These patterns should be interpreted in terms of leverage, influence, and residual variation, and related back to observable characteristics of the individuals. While aggregate summaries should form the core of the analysis, teams are encouraged to examine a small number of illustrative observations in detail to clarify the economic and statistical mechanisms at work.

From the perspective of the tax authority (DIAN), teams should discuss whether observations with large prediction errors represent potential cases for audit or whether they primarily reflect limitations of the modeling approach. Are high-error cases systematically different in ways that suggest misreporting, or do they simply reflect groups for which these models are inadequate?