Unscrambling the Omelette: Recovering Hours and Wages from Total Earnings using Machine Learning Methods

John Green, Wonsik Ko, Su Wang

Applied Micro Student Workshop
December 4, 2025

Motivation

- **Administrative Data** is widely used in Economics (Cole et al., 2021).
  → Clear Advantage: Accurate records for full populations.
- **Main Caveat:** Limited set of variables.
  → US Income data from UI records **only** provides **quarterly earnings**.
- However, earnings are a composite: wage rates and the amount of labor supplied.
  → $Y = W \times H$.

Suppose we want to understand some policy's impact on individual labor market outcomes.

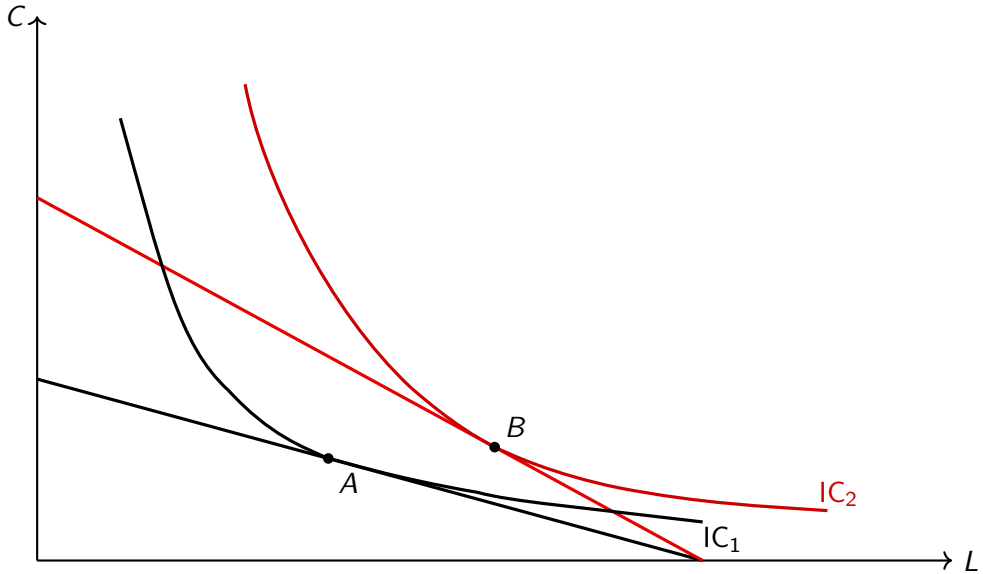
1. When there is **zero** impact on earnings (Y)
   - $\rightarrow$ It does not necessarily mean zero policy impact.
   - $\rightarrow$ Policy may have **offsetting effects** on the wage rate (W) and working hours (H).
2. When there is **some** impact on earnings (Y)
   - $\rightarrow$ A policy impact on Y may be driven **solely** by an impact on W or H.

For policy evaluations, we need to know:
Is the policy effect on W (productivity) or H (labor supply)?

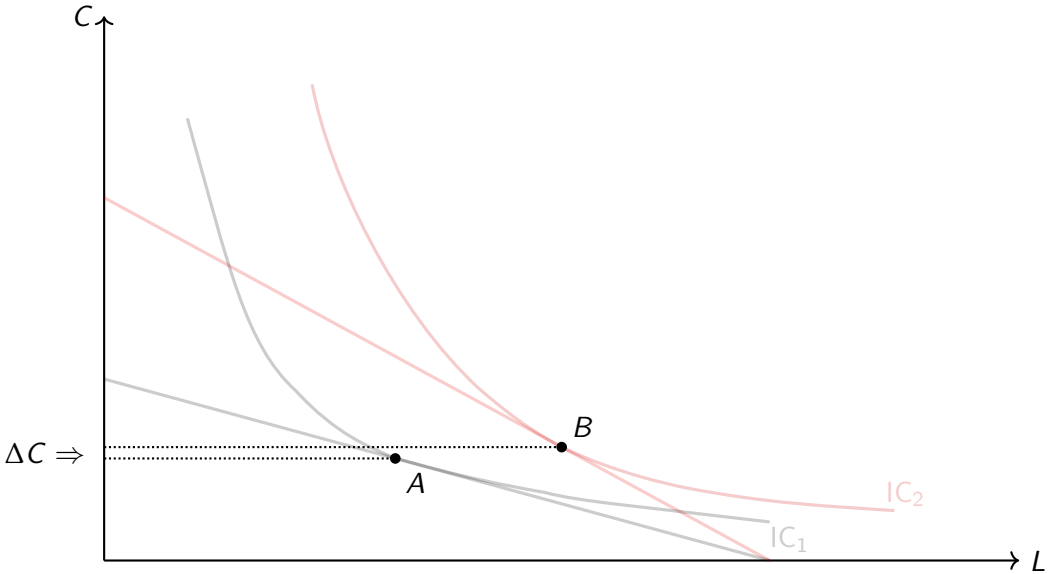
This paper...

- Can we use an auxiliary data set which contains H and W separately to *disentangle* H and W from Y in the administrative data?
- What are the effects of MD workforce training programs on earnings of displaced workers during the COVID-19 recession?
  - → How does the effect on wages differ from the effect on hours?
  - → Maryland Longitudinal Data System: (relatively) new administrative data on Maryland residents

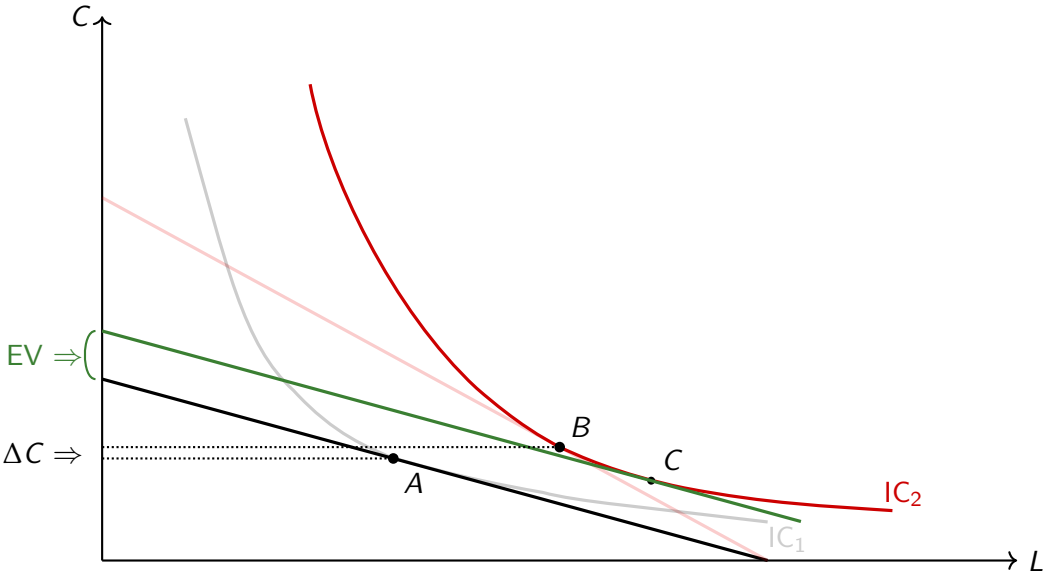
The Intuition



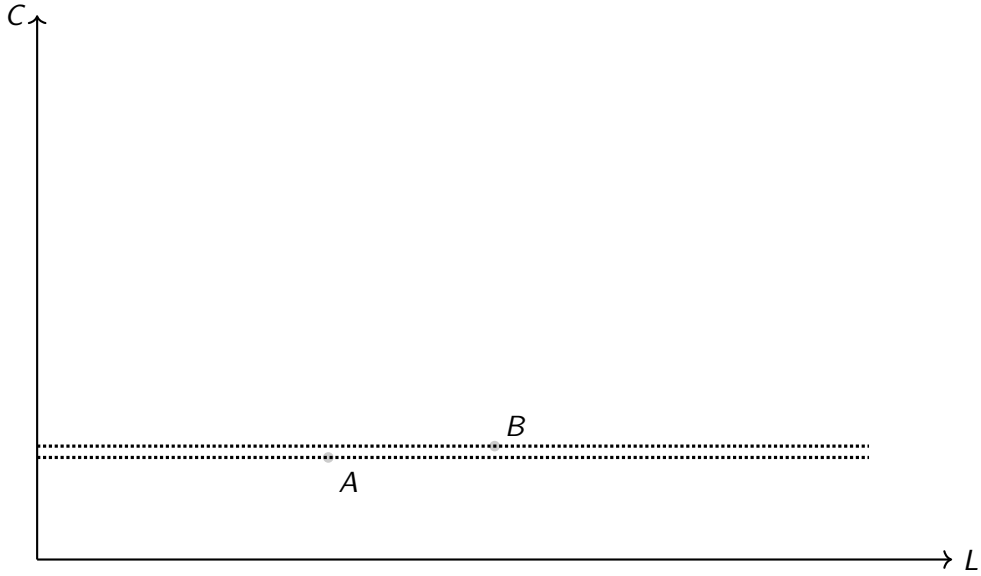
The Intuition



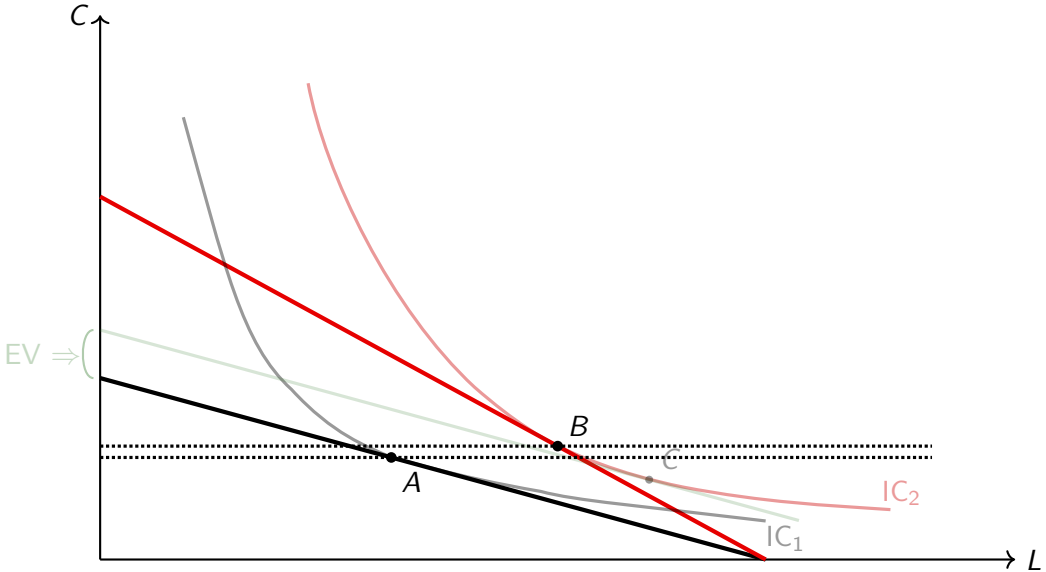
The Intuition



Earnings Data



With Imputed Wage



Method

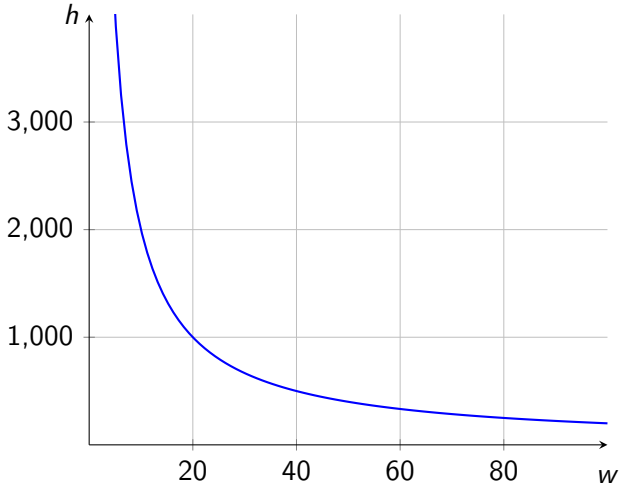
- In the administrative data, we have earnings (Y) and some individual characteristics (X).
- What if we have information on the joint distribution (X, W, H)?
  - → Survey data (e.g. CPS)
- Learn the distribution (X, W, H) using Machine Learning methods and decompose the earnings administrative data
  - → We attempt to guide the algorithm with some knowledge from economic theory: a labor supply function

Method

- In the administrative data, we have earnings (Y) and some individual characteristics (X).
- What if we have information on the joint distribution (X, W, H)?
  - $\rightarrow$ Survey data (e.g. CPS)
- Learn the distribution (X, W, H) using Machine Learning methods and decompose the earnings administrative data
  - $\rightarrow$ We attempt to guide the algorithm with some knowledge from economic theory: a labor supply function

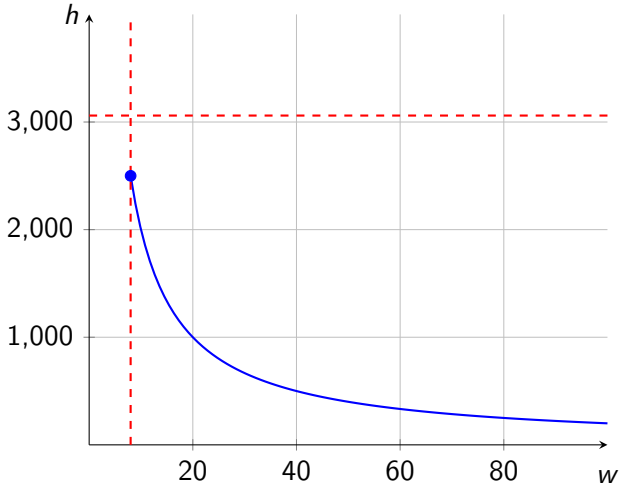
Illustration

Given a particular income level, we know that the individual's wage and hours must be:



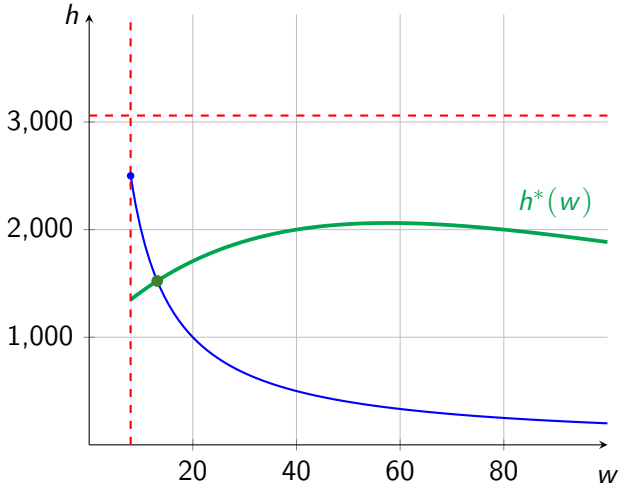
Illustration

Cut off parts of this curve with constraints: maximum working hours, minimum wage



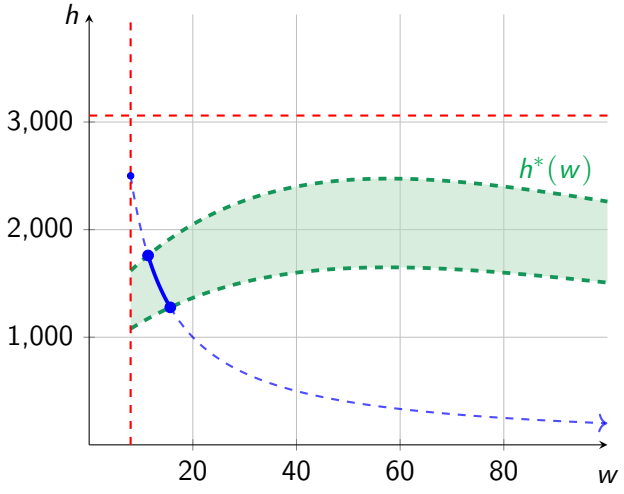
Illustration

If we have a labor supply function:



Illustration

If there are some uncertainty around the labor supply:



Specifications

We assume that consumers are hand-to-mouth with CRRA utility:

$$u(c_i, w_i; X_i) = \frac{c_i^{1-\gamma(X_i)}}{1-\gamma(X_i)} - \beta(X_i) \cdot \frac{h_i^{1+\sigma(X_i)}}{1+\sigma(X_i)} \tag{1}$$

which means estimating the following model:

$$\ln h_i = \varpi(X_i) + \alpha(X_i) \ln w_i + \epsilon_i \tag{2}$$

With an estimated labor supply, we can invert and get the imputed wage

$$\ln Y_i = \ln h_i + \ln w_i = \varpi(X_i) + [\alpha(X_i) + 1] \ln w_i + \epsilon_i$$

$$\ln w_i = \frac{\ln Y_i - \varpi(X_i) - \epsilon_i}{\alpha(X_i) + 1}$$

Specifications

We assume that consumers are hand-to-mouth with CRRA utility:

$$u(c_i, w_i; X_i) = \frac{c_i^{1-\gamma(X_i)}}{1-\gamma(X_i)} - \beta(X_i) \cdot \frac{h_i^{1+\sigma(X_i)}}{1+\sigma(X_i)} \tag{1}$$

which means estimating the following model:

$$\ln h_i = \varpi(X_i) + \alpha(X_i) \ln w_i + \epsilon_i \tag{2}$$

With an estimated labor supply, we can invert and get the imputed wage

$$\ln Y_i = \ln h_i + \ln w_i = \varpi(X_i) + [\alpha(X_i) + 1] \ln w_i + \epsilon_i$$
$$\ln w_i = \frac{\ln Y_i - \varpi(X_i) - \epsilon_i}{\alpha(X_i) + 1}$$

Estimation

- Observables X_i include state, year, gender, race, age, and education
- $\varpi(X_i)$ is a flexible function of X_i, specified as a neural net:
  - 3 layers of 256, 128, and 64 neurons
  - Significant regularization: state and industry are given *embeddings*, dropout of 0.25, gradient clipping, etc.
  - Experiment with activation functions and other hyper-parameters
- $\alpha(X_i) = \pi_1 \text{age}_i + \pi_2 \text{age}_i^2 + \pi_3 \text{female}_i$
- Estimate the model on CPS ASEC using non-linear least squares
- **But** customize the loss function to place additional penalties on predictions that violate the minimum wage

Estimation

- Observables X_i include state, year, gender, race, age, and education
- $\varpi(X_i)$ is a flexible function of X_i, specified as a neural net:
  - 3 layers of 256, 128, and 64 neurons
  - Significant regularization: state and industry are given *embeddings*, dropout of 0.25, gradient clipping, etc.
  - Experiment with activation functions and other hyper-parameters
- $\alpha(X_i) = \pi_1 \text{age}_i + \pi_2 \text{age}_i^2 + \pi_3 \text{female}_i$
- Estimate the model on CPS ASEC using non-linear least squares
- **But** customize the loss function to place additional penalties on predictions that violate the minimum wage

Two replication exercises

To examine the performance of our methods, we conduct the following replication exercises:

1. Beland (2015) - CPS
   - → Estimates the impact of the party allegiance (Republican or Democratic) of US governors on labor-market outcomes
   - → Regression discontinuity design
2. Huckfeldt (2022) - PSID
   - → Compare the cost of job loss across occupation stayers and switchers
   - → Event study

Summary of Results - CPS

- Blacks' **hours worked** increase relative to whites under a Democratic governor.
- Suppose we only observe earnings in administrative data.
- What we observe: Blacks' **earnings** increase relative to whites under a Democratic governor.
  $\rightarrow$ This masks that the policy effect operates through hours, not wages.

Results - CPS

Table 1: Replication: Beland (2015) Table 2

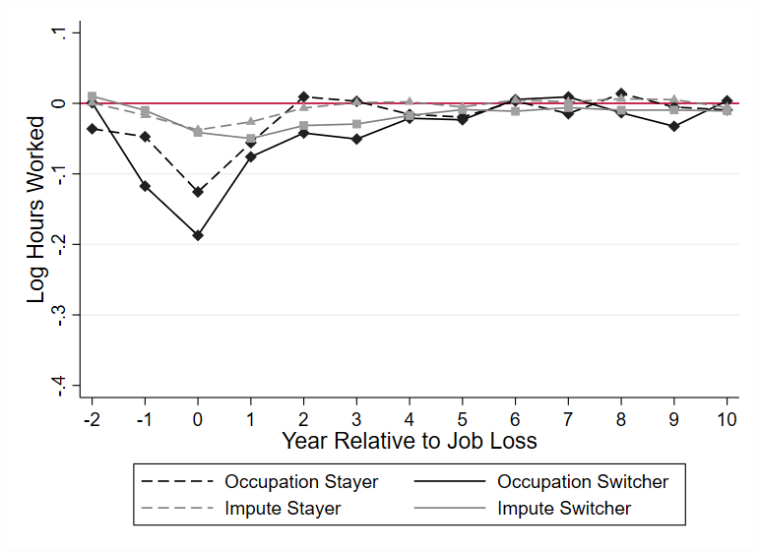
	(1)	(2)	(3)	(4)	(5)
	Hours		Hourly Wage		Earnings
	True	Imputed	True	Imputed	True
Democrat (=1)	-0.011***	-0.009***	-0.017***	-0.018***	-0.028***
	(0.004)	(0.002)	(0.006)	(0.006)	(0.010)
Democrat × Black	0.033***	0.028***	0.017	0.023**	0.050***
	(0.012)	(0.005)	(0.014)	(0.010)	(0.015)
Black (=1)	-0.046***	-0.047***	-0.092***	-0.091***	-0.138***
	(0.008)	(0.004)	(0.012)	(0.012)	(0.017)
Observations	1,145,032	1,145,032	1,145,032	1,145,032	1,145,032
R^2	0.109	0.460	0.306	0.193	0.283
SSPE		0.041		0.094	

Summary of Results - PSID

- Earnings losses from job loss are driven primarily by wage declines.
- These losses are concentrated among occupation switchers, not stayers.
- Our replication is somewhat off and requires additional work.
- Again, observing only earnings masks the main driver of the cost of job loss.

Results - PSID

Figure 1: Replication: Huckfeldt (2022) Figure 1, 2



Results - PSID

Figure 2: Replication: Huckfeldt (2022) Figure 1, 2



Future Plan

- One additional replication (NLSY?)
- Uncertainty quantification (e.g. mixed density networks, kernel variance regression)
- Then, move to our empirical application on MD workforce training programs
  - Standard matching approach, possible 2SLS model using COVID closures
  - This literature almost always focuses on earnings; we will examine whether training programs operate through W or H