Computation of MSA-Level Wage Premia

Project Workflow Automation

June 3, 2025

1 Introduction

This document outlines the methodology used to compute MSA-level wage premia based on ACS data. The goal is to estimate the extent to which wages in different Metropolitan Statistical Areas (MSAs) differ after controlling for individual worker characteristics.

2 Data Source

The primary data source for this analysis is a Stata file ('extract19.dta') containing pooled cross-sections from the American Community Survey (ACS) for the years 2005-2023. Key variables utilized from this dataset include:

- `incwage`: Individual's annual wage and salary income.

- `wkswork1`: Weeks worked last year.

- `uhrswork`: Usual hours worked per week.

- `perwt`: Person weight.

- `age`: Age of the individual.

- `sex`: Sex of the individual.

- `race`: Race of the individual.

- `educ`: Educational attainment of the individual.

- `met2013`: Metropolitan Statistical Area identifier (CBSA 2013 definition).

- `year`: Survey year.

3 Methodology

3.1 Hourly Wage Calculation

First, an hourly wage is computed for each individual i as:

$$\text{wage\_hr}_i = \frac{\text{incwage}_i}{\text{wkswork1}_i \times \text{uhrswork}_i} \tag{1}$$

Individuals with non-positive or non-finite hourly wages are excluded from the analysis. The natural logarithm of the hourly wage is then taken:

$$\ln(w_i) = \log(\text{wage\_hr}_i) \tag{2}$$

3.2 Regression Model

For each year t from 2005 to 2023, the following regression model is estimated using Ordinary Least Squares (OLS), weighted by `perwt`:

$$\ln(w_{ijmt}) = \beta_0 + \beta_1 \text{age}_{it} + \beta_2 \text{age}_{it}^2 + \mathbf{X}_{it}'\gamma + \alpha_{mt} + \epsilon_{ijt} \tag{3}$$

where:

- w_{ijmt} is the hourly wage of individual i in MSA m in year t.

- age_{it} and age_{it}^2 control for age in a quadratic form.

- $\mathbf{X}_{it}$ is a vector of dummy variables for individual characteristics: sex, race, and educational attainment. γ is the corresponding vector of coefficients.

- α_{mt} represents the fixed effect for MSA m in year t. These are the key parameters of interest.

- ϵ_{ijt} is the error term.

The model is specified in 'fixest::feols' syntax as:

`lnw ~ age + I(age extasciicircum2) + factor(sex) + factor(race) + factor(educ) | met2013`

This effectively means that `met2013` (representing α_{mt}) captures the average log-wage difference for MSA m in year t relative to a baseline MSA (omitted by the estimation procedure), after controlling for the specified individual characteristics.

3.3 Wage Premia

The estimated fixed effects, $\hat{\alpha}_{mt}$, are interpreted as the skill-adjusted log-wage premia for each MSA m in year t. These are denoted as `fe_adj_lnw` in the output data.

These log-wage premia (`fe_adj_lnw`$_{mt}$) represent the component of wages attributable to geographic location (MSA) after accounting for differences in the observable skill composition of the workforce across MSAs.

4 Output

The primary output containing these premia is `data/output/msa_wage_premia_2005_2023_simplified.csv`. This file includes `met2013`, `year`, and `fe_adj_lnw`, among other variables.