Predicting U.S. Wage Growth from Occupation and Regional Characteristics

Jiarong Wang
Stats 507
University of Michigan, Ann Arbor
Ann Arbor, USA
jiarong@umich.edu

***Abstract*—This study investigates the drivers of wage growth in the United States from 2020 to 2024 using data from the Bureau of Labor Statistics (BLS). By employing both Ordinary Least Squares (OLS) and Two-Way Fixed Effects models, we analyze the influence of occupational attributes and regional characteristics on wage dynamics. Our findings suggest that while baseline employment levels and industry specialization (Location Quotient) appear significant in naive models, they are statistically insignificant when controlling for unobserved heterogeneity. Instead, historical wage levels and short-term employment growth emerge as the true, significant drivers. The fixed effects model reveals that intrinsic regional economic conditions absorb much of the variance often attributed to simple cross-sectional predictors.**

***Keywords*—wage growth, labor economics, fixed effects, predictive modeling, BLS data**

I. Introduction

Wage growth is a fundamental indicator of economic health, reflecting productivity, labor demand, and regional economic conditions. Understanding the drivers of wage dynamics across different occupations and locations provides critical insights for policymakers, job seekers, and economists alike. While traditional economic models, such as Mincer's human capital earnings function [1], emphasize individual characteristics like schooling and experience, recent scholarship has shifted focus toward structural and spatial determinants. Autor and Dorn [2] and Goos and Manning [3] documented the polarization of the labor market, where wages have risen fastest in high-skill cognitive and low-skill service occupations. Simultaneously, regional disparities have widened, with dense, skill-concentrated areas generating significant wage premiums due to agglomeration economies [4].

Despite these theoretical advancements, identifying the unbiased, causal drivers of wage growth remains challenging due to the presence of unobserved heterogeneity. Simple cross-sectional analyses often fail to distinguish whether wage premiums are driven by intrinsic occupational attributes (e.g., specialization) or by time-invariant regional characteristics. For instance, variables such as total employment (TOT_EMP) and location quotient (LOC_QUOTIENT) are widely used to predict spatial variations in growth [5], yet their estimated effects can be confounded without proper controls for entity-specific and macro-level shocks.

Our project seeks to build a predictive model of wage growth using open data from the U.S. Bureau of Labor Statistics (BLS) Occupational Employment and Wage Statistics (OEWS) program from 2020 to 2024. We aim to quantify how occupation-specific attributes, regional scale, and industry mix influence wage trends over this period. Specifically, we investigate the question: Which types of jobs and which regions experience faster wage growth, and why? To answer this, we employ a two-stage econometric approach. We first establish a baseline using a standard Ordinary Least Squares (OLS) model, and then we contrast these findings with a more robust Two-Way Fixed Effects (PanelOLS) model. This comparative framework allows us to correct for omitted variable bias by controlling for unobserved time-invariant heterogeneity across state occupation entities.

II. Methodology

A. Data acquisition and Preprocessing

The primary dataset was sourced from the BLS OEWS program, covering annual state-level estimates from 2020 to 2024. The raw dataset contains granular estimates for employment and wages across varying occupation codes ('OCC_CODE') and primary states ('PRIM_STATE'). To ensure data quality and model stability, the following preprocessing steps were applied:

1) Filtering: Observations with missing mean annual wage ('A_MEAN') or missing total employment ('TOT_EMP') figures were removed. Low-wage entries (< 1000) were filtered out and data were sorted by occupation/state and year. And duplicate rows were removed based on unique identifier (year, PRIM_STATE, OCC_CODE).

2) Type Correction: Core metrics, including employment (TOT_EMP), mean wage (A_MEAN), Location Quotient ('LOC_QUOTIENT'), were forced to numeric types to handle non-numeric placeholders (e.g., "*" used by BLS for estimates with low-reliability).

3) Panel Structure: A unique entity identifier, 'state_occ', was created by concatenating the state abbreviation and occupation code. The data was then indexed by 'state_occ' (Entity) and 'year' (Time).

B. Feature Engineering

We constructed a set of log-transformed and growth based the skewed data features to capture interpretability, elasticities

and dynamic trends.

Target Variable: The response, Y_{it}, is the 2-Year Wage Growth (wage_growth_2yr). Unlike annual volatility, a two-year window effectively lessons short term noise and captures structural wage trends. It is calculated as:

$$Y_{it} = \frac{Wage_{it} - Wage_{i,t-2}}{Wage_{i,t-2}} \quad (1)$$

Predictor Variables: The feature space X_{it} consists of:

• Log Mean Wage $(\ln(Wage_{it}))$: The natural logarithm of the mean annual wage, used to control for the base wage level and test for convergence or divergence.

• Log Employment $(\ln(Emp_{it}))$: The natural logarithm of total employment, proxying for the absolute size of the labor market.

• Log Location Quotient $(\ln(LQ_{it}))$: The natural logarithm of the Location Quotient, capturing the relative concentration of an occupation compared to the national average.

• Employment Growth $(\Delta \mathrm{Emp}_{it})$: The year-over-year percentage change in total employment, representing short-term labor demand shocks.

• Interaction Term: An interaction between Log Employment and Log LQ $(\ln(Emp) \times \ln(LQ))$ was included to test if the effect of market size is conditional on specialization.

C. Econometric Models

We specified two distinct linear models to isolate the drivers of growth.

Model 1: Pooled OLS (Baseline): This model treats all observations as independent, ignoring the panel structure. It assumes that error terms are uncorrelated with the regressors:

$$Y_{it} = \beta_0 + \beta X_{it} + \mathcal{E}_{it} \quad (2)$$

Model 2: Two-Way Fixed Effects (PanelOLS): To address Omitted Variable Bias (OVB), we implemented a Two-Way Fixed Effects model using the linearmodels library. This model controls for unobserved time-invariant characteristics (α_i) and common time shocks (γ_t) while also controls for job-state pairs, which represents that we focus on time-invariant characteristics (α_i) within that entity, rather than across different entities:

$$Y_{it} = \beta X_{it} + \alpha_i + \gamma_t + u_{it} \quad (3)$$

This method uses a "within-transformation" to absorb α_i and γ_t, allowing us to estimate the causal impact of time-varying features on wage growth within specific state occupation pairs.

III. Results

To identify the causal drivers of wage growth, we compared a Baseline OLS model against a Fixed Effects (PanelOLS) model. Both models utilized the same dataset though the number of observations deviate gently for data-cleaning part, and the same target variable (2-Year Wage Growth).

A. Model Performance Comparison

TABLE 1 represents a side-by-side comparison of the coefficients and the model fit statistics.

TABLE 1 Comparative Regression Result: OLS vs Fixed Effects

	Model 1: OLS		Model 2: Fixed Effects	
Variable	**Coef.**	**P-Value**	**Coef.**	**P-Value**
Constant	-0.0815	0.000	-15.276	0.0000
ln(Emp)	-0.0001	0.391	0.0067	0.1479
ln(LQ)	-0.0116	0.000	-0.0052	0.4308
ln(emp_LQ)	0.0015	0.000	-3.423e-05	0.9674
Emp.Growth	-0.0056	0.000	0.0023	0.0088
ln(mean wage)	0.0117	0.000	1.3965	0.0000
R^2	0.004		0.3376 (Within)	
F-Test(Poolability)	—		6.227 (p=0.00)	

Note: Bold coefficients indicate statistical significance at $p < 0.05$. The Fixed Effects model controls for Entity(State-Occupation) and Time(Year)

B. Analysis of Results

The Failure of OLS and Omitted Variable Bias:

The Baseline OLS model performs poorly, explaining less than 1% of the variance ($R_2 = 0.004$). More critically, it yields misleading results regarding industry specialization. In the OLS model, the variable log_LQ (Location Quotient) has a statistically significant negative coefficient (-0.0116). A naive interpretation would suggest that specialized industries inherently experience slower wage growth. However, this is likely a case of Omitted Variable Bias (OVB). High-LQ jobs (e.g., technology roles in California or manufacturing in Michigan) are intrinsically tied to specific regions with unique cost-of-living pressures, tax policies, and amenities. The OLS model cannot distinguish between the effect of the job concentration and the unobserved effects of the region itself.

The Superiority of Fixed Effects:

The PanelOLS model dramatically improves the model fit, raising the within-group R_2 to 0.3376. The F-test for poolability returns a value of 6.2277 ($p = 0.00$). This strongly rejects the null hypothesis that the fixed effects are zero. So the OLS model was confirmed to be biased. By controlling for entity-specific heterogeneity (α_i) in Model 2, log_LQ becomes statistically insignificant ($p = 0.431$). This implies that once we control for the specific state and occupation context, just the concentration of an industry does not coincidentally drive wage growth.

Drivers of Growth: Stock vs. Flow:

The results highlight a crucial distinction between economic "stock" (static levels) and "flow" (changes):

Stock (Insignificant): Static measures like the size of the workforce (log_emp) and specialization (log_LQ) do not predict wage growth rates.

Flow (Significant): However, the variable employment_

growth is positive and significant in the robust model (Coef: 0.0023, p = 0.009). This confirms basic supply and demand theory: an active increase in demand for labor (hiring) puts upward pressure on the price of labor (wages).

IV. Conclusion

By comparing a naive OLS baseline with a rigorous PanelOLS specification, we demonstrated that cross-sectional analyses of wage growth are prone to significant bias. Static measures of market structure (Market Size, Specialization) appear significant only when failing to control for regional heterogeneity. The robust model reveals that wage growth is dynamically driven by immediate labor demand shocks (employment growth) and a structural momentum where high wages indicates further high growth.

REFERENCES

[1] J. Mincer, Schooling, Experience, and Earnings. New York: National Bureau of Economic Research, 1974.

[2] D. Autor and D. Dorn, "The Growth of Low-Skill Service Jobs and the Polarization of the U.S. Labor Market," American Economic Review, vol. 103, no. 5, pp. 1553-1597, 2013.

[3] M. Goos and A. Manning, "Lousy and Lovely Jobs: The Rising Polarization of Work in Britain," Review of Economics and Statistics, vol. 89, no. 1, pp. 118-133, 2007.

[4] E. Moretti, "Local Labor Markets," in Handbook of Labor Economics, vol. 4B, D. Card and O. Ashenfelter, Eds. Amsterdam: Elsevier, 2011, pp. 1237-1313.

[5] T. J. Bartik, "Who Benefits from State and Local Economic Development Policies?" Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 1991.

[6] J. Azar, I. Marinescu, and M. Steinbaum, "Labor Market Concentration," The Journal of Human Resources, vol. 57, no. S, pp. S167-S199, 2022.