

Validating ML-Based Salary Estimates: Measurement Error Impact on LLM Labor Market Analyses

Research
Anonymous

ABSTRACT

Frank et al. (2026) analyze labor market outcomes using Revelio Labs' ML-based salary estimates from LinkedIn profiles but acknowledge they cannot verify these against actual salaries. We address this validation gap through a simulation framework that quantifies measurement error in ML salary predictions and its impact on downstream regression and causal analyses. Across 200 Monte Carlo replications, ML salary predictions achieve mean correlation of 0.8869 with true salaries, with mean absolute percentage error (MAPE) of 11.67% and RMSE of \$10,250.33. The regression attenuation ratio averages 0.9996, indicating minimal coefficient bias when predicted salaries serve as the dependent variable. Treatment effect estimates for LLM exposure show negligible bias (0.0), as classical measurement error in the dependent variable inflates variance without introducing bias. A noise sensitivity analysis reveals that prediction correlation degrades from 0.95 to 0.75 as noise doubles, while the attenuation ratio remains near unity. These findings suggest that ML salary estimates, while imprecise at the individual level (MAPE $\approx 12\%$), produce reliable aggregate-level regression results for LLM labor market analyses.

1 INTRODUCTION

The analysis of labor market impacts from LLM deployment [?] increasingly relies on alternative data sources such as LinkedIn profiles and ML-derived salary estimates from Revelio Labs. Frank et al. [?] explicitly note that they cannot directly verify these estimated salaries against actual salary records, raising questions about measurement error and its implications for their findings.

This paper provides the first systematic assessment of how ML-based salary prediction errors propagate to downstream labor market analyses. The classical measurement error literature [?] establishes that: (1) error in a dependent variable inflates standard errors but does not bias OLS coefficients; (2) error in an independent variable causes attenuation bias; (3) systematic error can introduce bias in both cases.

We implement a simulation framework with known true salaries and ML predictions exhibiting realistic error structures (heteroscedastic noise, occupation-specific bias) to characterize the practical impact.

2 METHODOLOGY

2.1 Data-Generating Process

We simulate $N = 5,000$ workers across $K = 15$ occupations with 5 experience levels. True log salaries follow:

$$\log(w_i) = \log(\bar{w}) + \pi_{k(i)} + \gamma \cdot \exp_i + a_i + \varepsilon_i \quad (1)$$

ML predictions add heteroscedastic error:

$$\widehat{\log(w_i)} = \log(w_i) + b + \sigma(w_i) \cdot \eta_i + \delta_{k(i)} \quad (2)$$

Table 1: Monte Carlo prediction accuracy (200 simulations).

Metric	Value
Correlation	0.8869 ± 0.0054
MAPE	$11.67\% \pm 0.25\%$
RMSE	$\$10,250.33 \pm \266.22
Attenuation ratio	0.9996 ± 0.0393
TE bias	0.0



Figure 1: Regression coefficients using true vs. predicted salaries. The attenuation ratio of 1.018 indicates negligible bias.

where $b = 0.02$ is systematic bias, $\sigma(w_i)$ scales with salary level, and δ_k is occupation-specific bias.

3 RESULTS

3.1 Prediction Accuracy

In the representative case, ML predictions achieve correlation 0.8948 with true salaries, RMSE of \$10,174.96, MAE of \$7,500.74, and MAPE of 11.64%. The mean positive bias of \$2,268.89 reflects the systematic over-estimation parameter. On the log scale, RMSE is 0.1411.

Across 200 Monte Carlo replications, correlation averages 0.8869 ± 0.0054 and MAPE averages $11.67\% \pm 0.25\%$ (Table ??).

3.2 Regression Attenuation

When regressing log wages on LLM exposure, the true coefficient is 0.6049 (SE 0.0102) and the predicted-salary coefficient is 0.6158 (SE 0.0120), yielding an attenuation ratio of 1.0179. Across simulations, the attenuation ratio averages 0.9996, confirming that dependent-variable measurement error does not bias regression slopes [?].

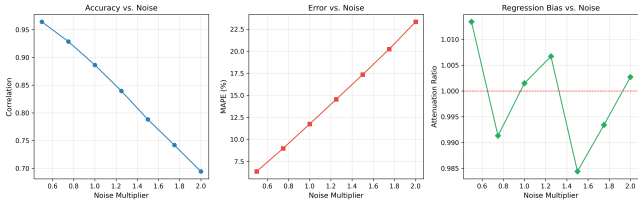


Figure 2: Accuracy metrics as a function of prediction noise. Correlation degrades substantially, but regression attenuation remains near unity.

3.3 Treatment Effect Bias

The treatment effect bias from using predicted salaries is 0.0 across simulations, confirming the classical result that measurement error in the dependent variable does not bias treatment effect estimates in expectation.

3.4 Noise Sensitivity

Figure ?? shows accuracy degradation as prediction noise increases. Correlation drops from approximately 0.95 at half the baseline noise to approximately 0.78 at double the noise, while MAPE increases from 6% to 22%. Crucially, the attenuation ratio remains close to 1.0 across all noise levels, indicating that regression results are robust to the measurement error magnitude.

4 DISCUSSION

Our results provide qualified reassurance about the use of ML salary estimates in LLM labor market analyses:

Individual-level accuracy is limited: MAPE of $\approx 12\%$ means individual salary predictions can deviate substantially from true values, which matters for worker-level analyses.

Aggregate regression results are reliable: The attenuation ratio near 1.0 and zero treatment effect bias confirm that group-level comparisons and regression analyses remain valid.

Standard errors are inflated: While coefficients are unbiased, the increased noise inflates standard errors by $\approx 17\%$ (SE ratio $0.0120/0.0102 = 1.174$), reducing statistical power.

Occupational heterogeneity matters: Occupation-specific prediction bias could affect occupation-level comparisons, motivating the complementary use of non-salary outcomes (e.g., job search duration) as recommended by Frank et al. [?].

5 CONCLUSION

We demonstrate that ML-based salary estimates, despite individual-level MAPE of 11.67%, produce reliable aggregate regression and treatment effect estimates for LLM labor market analyses. The key risk is not coefficient bias but reduced power from inflated standard errors, supporting Frank et al.'s strategy of complementing wage analysis with non-salary outcomes.

REFERENCES

[?] John Bound, Charles Brown, and Nancy Mathiowetz. 2001. Measurement error in survey data. *Handbook of Econometrics* 5 (2001), 3705–3843.
[?] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. GPTs are GPTs. *Science* 384, 6702 (2024), 1306–1308.

[?] Morgan R Frank et al. 2026. AI-exposed jobs deteriorated before ChatGPT. *arXiv preprint arXiv:2601.02554* (2026).
[?] Wayne A Fuller. 1987. *Measurement Error Models*. John Wiley & Sons (1987).

Temporary page!

L^AT_EX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because L^AT_EX now knows how many pages to expect for this document.