Wage Inequality in the U.S. Labor Market

Emily Eames and Layne DiBuono
American University

Introduction

“The Gender Pay Gap” continues to be a hot topic in today’s political climate. It is a term that is often thrown around in conversation and misused or misrepresented. The Gender Pay Gap itself is looking at the difference in wages **based solely on gender**. On the surface this is a relatively simple question, but the implications of it’s results are really what spark the controversy.

We aim to answer this original question. To what extent just “gender” determine an individual's wage? In addition, we attempt to make an educated and accurate guess as to weather an individual will make more or less than their opposite sex coworker based solely on gender?

Methodology

This project investigates gender as a major determinant for wage inequality in the U.S. labor market using advanced regression (OLS, Ridge, and Lasso) and classification methods (Logistic Regression, Random Forest, and SVMs) to compare predictive accuracy along with interpretability. We then used cross validation to select the optimal regularization parameter by minimizing cross-validated prediction error. Lastly, we used the F1 score as a key performance metric for evaluating our binary classifiers.

Data

Our data is from the Bureau of Labor Statistics (BLS). Specifically, we are using the 2024 Labor Force Statistics from the Current Population Survey. This data set looks at household median weekly earning of full-time wage and salary workers, broken down my occupation and gender. The dataset contains just under 600 different occupations ranging from metal and plastic workers to engineers and podiatrist. This dataset alone does not fully account for education and experience
.

Stakeholders, Limitations, & Ethical Consideration

Stakeholders: Policymakers, students, academics, employees and employers, thinktanks, and unions.

Limitation: Data only accounts for gender and occupation. It does not account for other factors such as education and experience.

Ethics: Avoid implying causation from correlation. Data uses Median Wages - we cannot report on the individual level.

Regression

What are the main factors that explain variation in individual wages? Typically , education, experience, occupation, gender, region, and race. Focusing on just one determinant, gender, how do regularization methods like Ridge and LASSO improve model performance and interpretability?

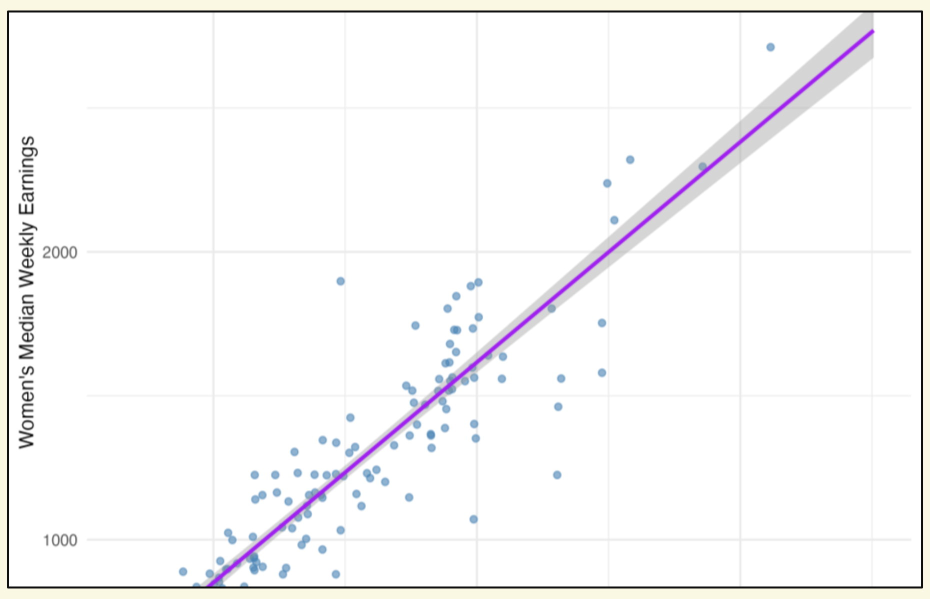


Fig. 1. Correlation plot

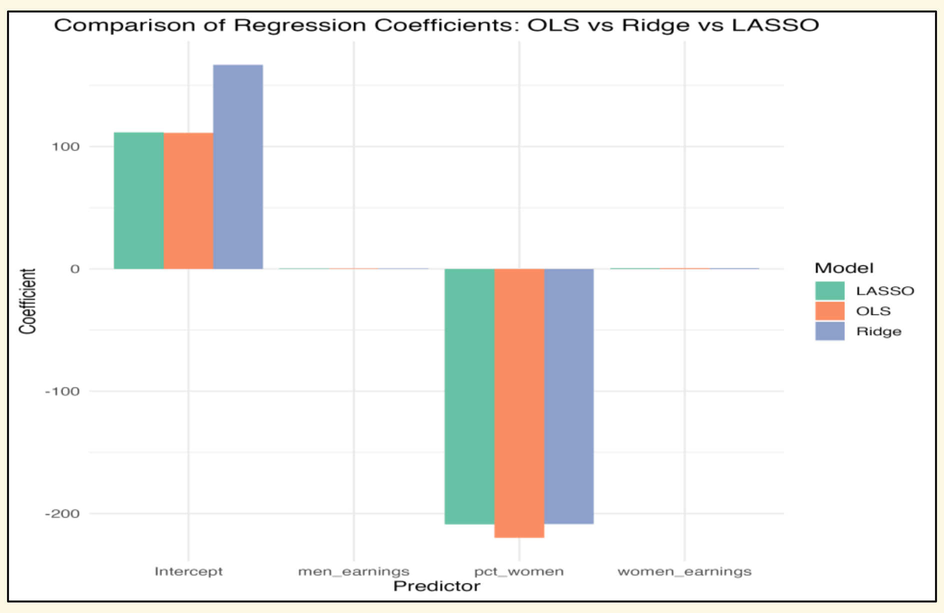


Fig. 2. Comparison plot

total_earnings=β0+β1(men_earnings)+β2(women_earnings)+β3(pct_women)+ε

Model/Predictor	OLS	Ridge	Lasso
Men_earnings	0.4610513	0.4347357	0.4620192
Women_earning	0.5428092	0.5224094	0.5365962
Pct_women	-219.7951658	-208.5343685	-208.8967670
RMSE	49.7051889	54.8905951	49.8138623

Classification

Can we predict whether an individual earns above or below the median national wage based on their gender?

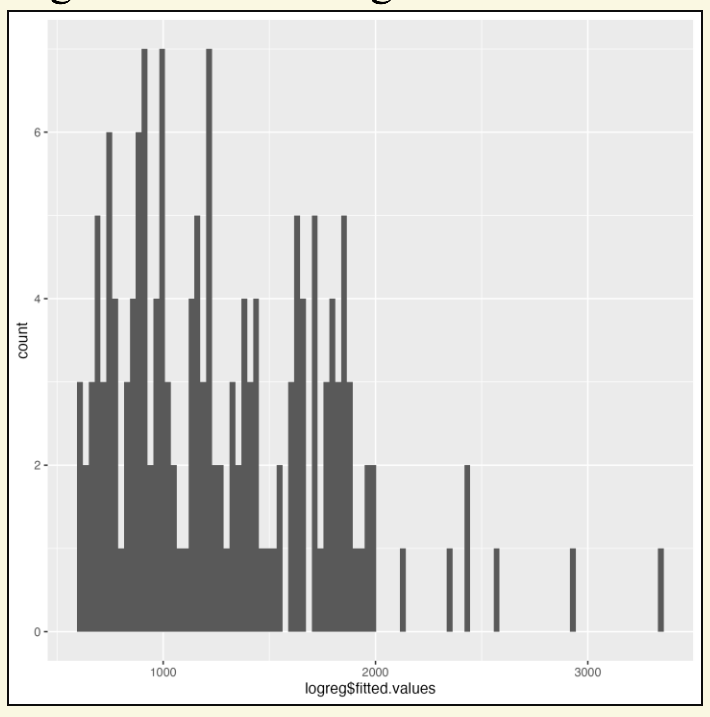


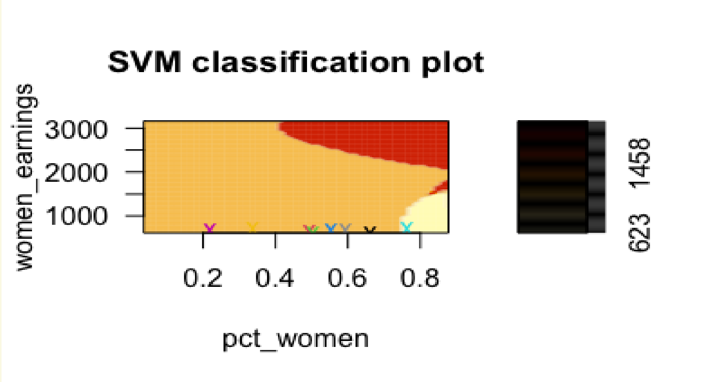
Fig. 3. Fitted value plot

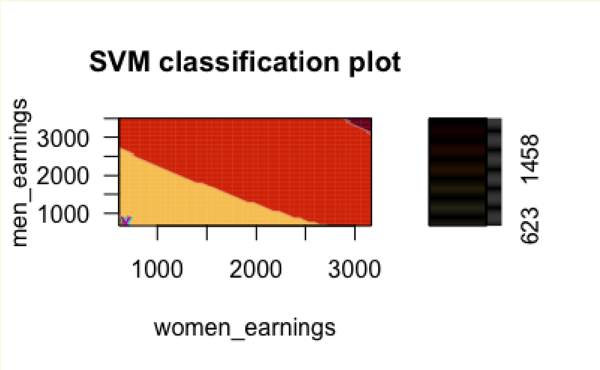
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    111.08107   15.12506   7.344 1.06e-11
men_earnings     0.46105    0.02149  21.449 < 2e-16
women_earnings   0.54281    0.02632  20.627 < 2e-16
pct_women     -219.79517   19.15350 -11.475 < 2e-16

(Intercept)    ***
men_earnings   ***
women_earnings ***
pct_women      ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

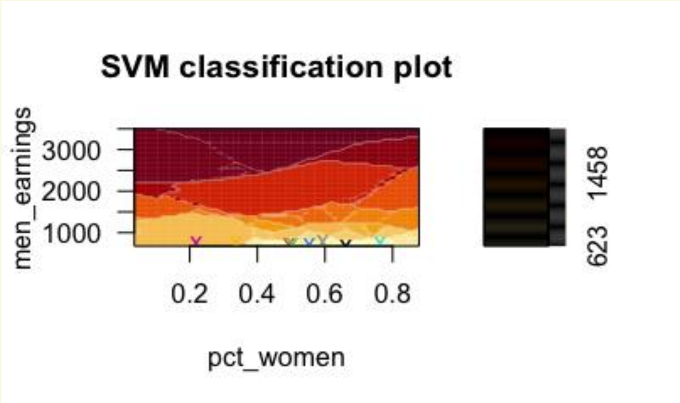
P-values of all variables used are <0.001, making them significant predictors of total earnings. Fig 3. displays large peaks around 0-2,000 and 4,000 – 6,000.

VIF Output	
Men_earnings	8.6566
Women_earnings	8.608
Pct_women	1.018

VIF shows high multicollinearity between men_earnings and women_earnings. Pct_women has no meaningful correlation with the other variables







These three SVM decision boundary plots illustrate how earnings predict gender in occupational data. The top plot, using women’s earnings and the proportion of women in the occupation, shows a nearly horizontal boundary—women’s earnings alone strongly predict gender, while the occupational composition adds little. The middle plot, using men’s and women’s earnings, shows that men almost always earn more than women, allowing the SVM to separate genders almost perfectly along the line where men’s earnings equal women’s. The bottom plot, using men’s earnings and occupational gender composition, again reveals a near-horizontal boundary, indicating men’s earnings alone are sufficient to predict gender. Overall, the visualizations demonstrate that earnings are a highly effective proxy for gender, far more so than occupational composition, highlighting systemic pay gaps and occupational segregation.

Random Forest:
92% of the variance is explained by the model which means it fit the data very well. The high correlation between men's and women's earnings does not harm the model

LOOCV:
Your LOOCV RMSE: 14.50
Your RMSE as % of the mean: 1.1 %
Approximate R^2 from LOOCV: 0.999

F1 and Accuracy Scores	
Logistic Regression Accuracy score	93.9%
Recall	95%
F1 score	94%

Results and Analyses

Our regression results show that both men’s and women’s earnings positively predict total earnings, but the percentage of women in an occupation has a strong negative association.

The classification methods consistently showed that gender-related features reliably distinguish between high- and low-earning occupations.

Overall, cross-validation confirmed that regularization does not substantially improve prediction accuracy, reinforcing the strength and consistency of the baseline regression model.

Conclusion

So, do women make less than men? In short, yes!

Our research shows that women consistently earn less than men across the sampled occupations in 2024, demonstrating that gender continues to be a strong predictor of wage differences. Both our regression and classification methods reveal a clear pattern: occupations with higher shares of women are systematically associated with lower total earnings, even when controlling for men’s and women’s pay levels.

References:

Gould, E., Mishel, L., & Bivens, J. (2015, January 6). *Wage stagnation in nine charts*. Economic Policy Institute . https://www.epi.org/publication/charting-wage-stagnation/
Sherman, A., Trisi, D., & Cureton, J. (2024, December 11). *A guide to statistics on historical trends in income inequality*. Center on Budget & Policy Priorities. https://www.cbpp.org/research/poverty-and-inequality/a-guide-to-statistics-on-historical-trends-in-income-inequality
Yellen, J. L. (2021, January 6). *The history of women’s work and wages and how it has created success for us all*. Brookings. https://www.brookings.edu/articles/the-history-of-womens-work-and-wages-and-how-it-has-created-success-for-us-all/