Gender Wage Ratios Are Increasing Rapidly

DS 4002
10/29/24
Group 8: Grace Brasselle, Kristy Luk (Leader), & Isabel O’Connor

Intro & Hypothesis → Data Acquisition → Analysis Plan → Testing & Analysis → Results → Next Steps

Motivation

- Since women entered the workforce, there has been a consistent gap between the average salaries for men and women
- Gender-based wage disparities have persisted, despite women participating more in the labor force and holding positions once dominated by men
- Various social factors continue to contribute to this inequality

Goal: Analyze how different factors can impact the gender wage gap

Research Question

How do predicted future gender wage gaps compare to historical data from 1960-2000?

Modeling Approach

Use ARIMA (Autoregressive Integrated Moving Average) modeling in Python to predict future values of the gender wage gap based upon historical socioeconomic factors

Data Acquisition/Explanation

Acquisition

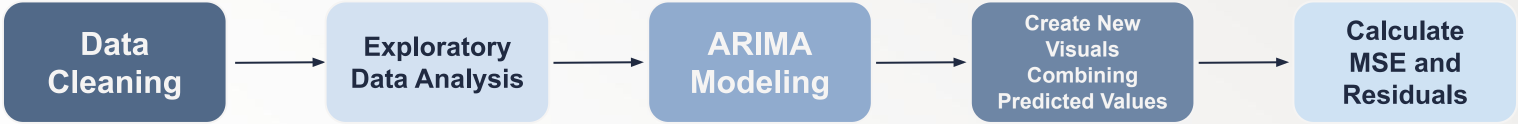
- Acquired from multiple reputable sources like the U.S. BLS
- No licensing or ethical concerns
- Text data
- Original dataset has 60 rows and 5 columns

Data Cleaning

- Cleaned by deleting the rows 1961 and 1963 because of N/A values

Column	**Description**	**Potential Response**
Year	Year in which data was recorded	1980
Female_LFPR	The percentage of US women 16 years or older who participate in the labor force	51.6
Bachelor_percentage	The percentage of US women 25 years or older who have attained at least a bachelor's degree	13.6
Wage_ratio	The ratio, out of 100, of female to male earnings for full-time, year-round workers	60.2
First_Birth_Median Age	Median age of US women to give birth to their first child	22.32

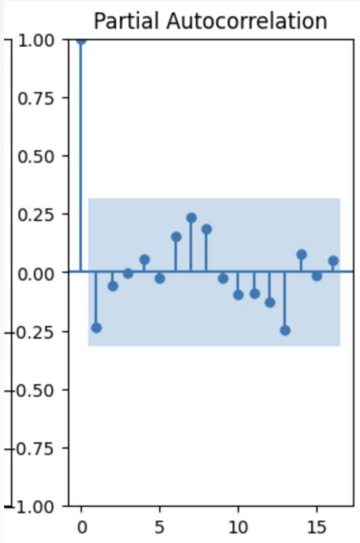
Analysis Plan

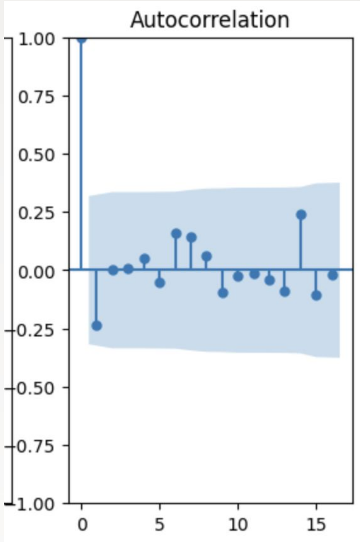


- After cleaning the data and conducting initial EDA
  - Use data from 1960–2000 to predict wage ratios for 2001–2019 with ARIMA
  - Plot both the predicted and actual values on a graph to display any differences
  - Calculate MSE and residuals to evaluate model accuracy

Tricky Analysis Decision

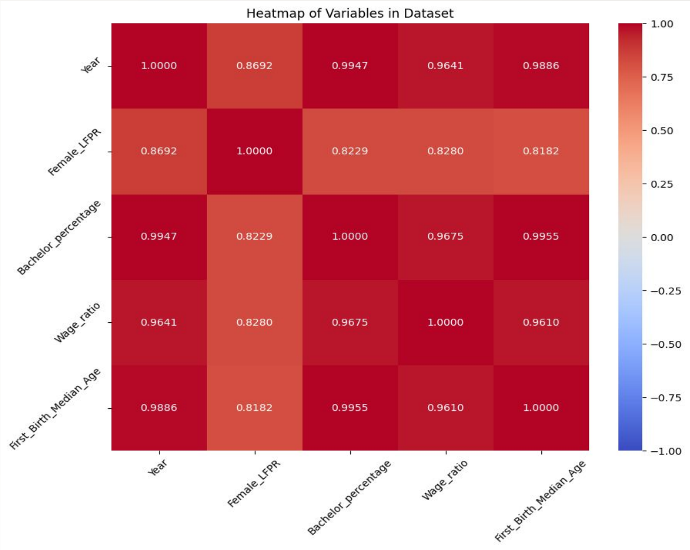
- Difficult to decide the best model parameters (p, q, d) that would be appropriate for ARIMA
- Initially started with trial and error to get an idea of what would work best
- Used the KPSS test to determine if our data was stationarity or not (d)
- Generated autocorrelation graphs and partial autocorrelation graphs to determine p and q





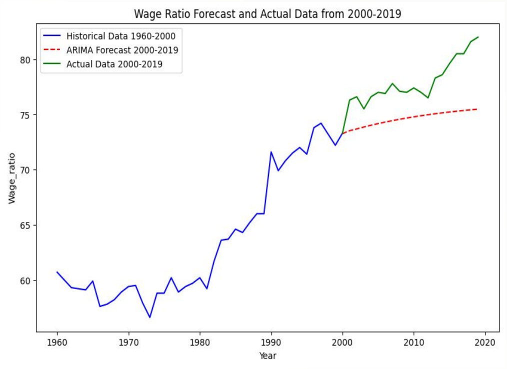
Bias and Uncertainty Validation

- Multicollinearity was a big concern after doing the EDA
    - But we agreed that omitted variable bias was arguably worse, so we decided to still go through with the ARIMA modeling
    - ARIMA is also designed to handle this
    - However, multicollinearity could reduce precision in the model
- Small sample size
    - Only 39 years accounted for when using ARIMA
    - This could affect accuracy of the predictions

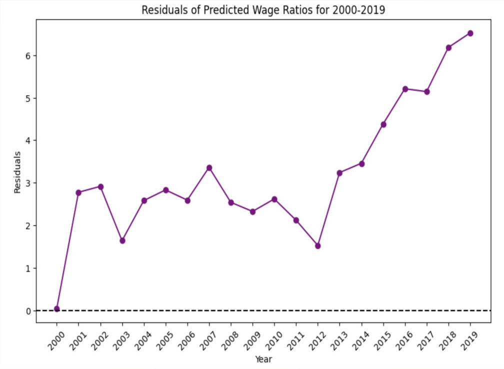


Results and Conclusions

ARIMA Graph



ARIMA Residuals



Moderate level of prediction error in our model

MSE = 12.6607
RMSE = 3.5582
Predicted, Actual Wage Ratio, and Residuals for 2000-2019:

	Year	Predicted_Wage_Ratio	Actual_Wage_Ratio	Residuals
0	2000	73.256089	73.3	0.043911
1	2001	73.523076	76.3	2.776924
2	2002	73.684339	76.6	2.915661
3	2003	73.860579	75.5	1.639421
4	2004	74.018129	76.6	2.581871
5	2005	74.167588	77.0	2.832412
6	2006	74.306782	76.9	2.593218
7	2007	74.437149	77.8	3.362851
8	2008	74.559038	77.1	2.540962
9	2009	74.673060	77.0	2.326940
10	2010	74.779707	77.4	2.620293
11	2011	74.879459	77.0	2.120541
12	2012	74.972761	76.5	1.527239
13	2013	75.060031	78.3	3.239969
14	2014	75.141659	78.6	3.458341
15	2015	75.218009	79.6	4.381991
16	2016	75.289422	80.5	5.210578
17	2017	75.356219	80.5	5.143781
18	2018	75.418696	81.6	6.181304
19	2019	75.477135	82.0	6.522865

Our model predictions for the wage ratio get increasingly more inaccurate compared to the actual data obtained for 2001-2019

Next Steps

New Lines of Exploration

- Expand the model to include other countries for analysis of global trend
- Examine how race/ethnicity affects gender gaps

Improvements

- Use more historical data to improve model predictions
- Incorporate more predictive variables related to different social factors
- Use a different modeling approach, such as exponential smoothing

New Questions

- When do we predict the wage gap to close, if it does?
- How does ethnicity play a role in the wage ratio?

References and Acknowledgements

[1] B. Etienne, “Time Series in Python — Exponential Smoothing and ARIMA processes,” TowardsDataScience.com, https://towardsdatascience.com/time-series-in-python-exponential-smoothing-and-arima-processes-2c67f2a52788 (accessed Oct. 23, 2024).

[2] D. Abugaber, “Chapter 23: Using ARIMA for Time Series Analysis,” University of Illinois Chicago, https://ademos.people.uic.edu/Chapter23.html/ (accessed Oct. 10, 2024).

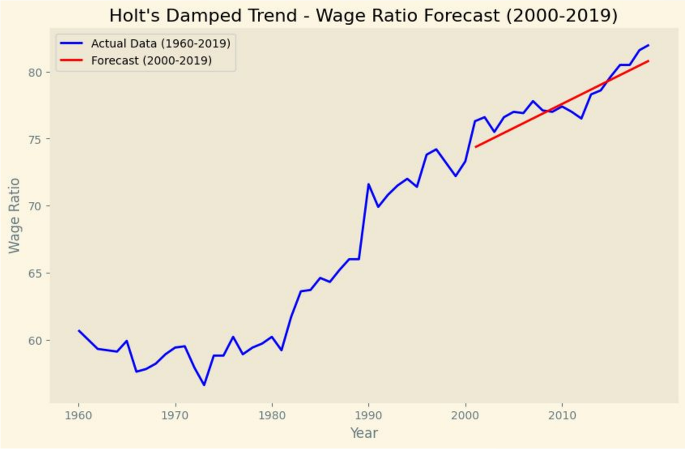
[3] J. Brownlee, “How to create an Arima model for time series forecasting in Python,” MachineLearningMastery.com, https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/ (accessed Oct. 18, 2024).

[4] Fuqua School of Business, Introduction to ARIMA models, https://people.duke.edu/~rnau/411arim.htm (accessed Oct. 18, 2024).

https://github.com/kristyluk/DS4002Project2

Thank you to Professor Alonzi and our TA Layla for their assistance on this project!

Holt Graph



Holt Residuals

Thank you!