Gender Wage Gap Analysis

Notebook: da1_assignment1_artyom-ashigov.ipynb

Objective:

This data analysis notebook investigates the gender wage gap using the U.S. Census Bureau's Current Population Survey (CPS) dataset. The analysis is conducted in Python using various libraries such as Pandas, Seaborn, Plotnine, and Statsmodels.

The primary aim is to comprehensively examine both conditional and unconditional gender wage gaps, considering factors such as education, age, and other demographic variables. The dataset used is the cps-earnings dataset, which contains information on 149,316 individuals.

The analysis focuses on **two occupation types**: web developers and office and administrative support occupations.

The data is filtered accordingly, and additional variables like female, hourly wages, log of hourly wages, and age squared are created for analysis.

The notebook employs visualizations to provide clear insights into the distribution of hourly earnings and the impact of natural logarithm transformation. Regression models are then used to analyze the gender wage gap, considering different factors.

Key Findings:

- **Unconditional Gender Wage Gap (Web Developers):** A basic regression model reveals that female web developers earn, on average, 18% less than their male counterparts.
- **Minimal Difference between Robust and Standard Models (Web Developers):** The disparity between the robust model, which considers heteroskedasticity-robust standard errors, and the standard model is minimal for web developers. This implies that, for web developers, the choice between robust and standard models has a limited impact on the estimated coefficients and overall model interpretation.
- **Conditional Gender Wage Gap with Education (Office Workers):** Considering education levels, the gender wage gap persists but is influenced by factors such as having a Ph.D. or professional degree.
- **Conditional Gender Wage Gap with Age (Office Workers):** Analyzing age-related factors indicates that, on average, female office workers can expect a 1% higher salary for each year of experience. However, the positive effect of age on earnings is weaker for females compared to males(1.6%).
- **Quadratic Form Better Explains Age-Earnings Relationship (Office Workers):** Incorporating quadratic(especially), cubic, and quartic forms of age into the regression model demonstrates a notable improvement in explanatory power. The R-squared value for the quadratic regression model surpasses that of the linear model, indicating

a closer fit to the underlying data. This suggests that the functional relationship between age and hourly wages is better captured by a quadratic form.

- **Polynomial Regression Model Reveals Dynamic Gender Wage Gap (Office Workers):** A more complex model involving polynomial terms and interactions between gender and age reveals that the gender wage gap varies over the life course. Initially, there is no visible difference, but the gap starts expanding after the age of 30, reaching about 20% around age 50, and then decreasing to around 15%.
- **Extended Regressions:** The analysis is extended to include additional variables such as race, citizenship status, marital status, and job characteristics. Even after accounting for these factors, being female is associated with a statistically significant decrease in hourly wages, with an estimated effect ranging from -3.2% to -7.7%.
- **Causal Analysis:** The notebook considers potential discriminatory factors by incorporating variables related to race, citizenship, marital status, and job characteristics. The analysis suggests that even after controlling for these factors, a significant gender wage gap persists, but the most significant factors remain age and education.

Conclusion:

This notebook systematically explores the gender wage gap, considering various factors that may contribute to disparities in earnings. The results indicate that being female is associated with lower hourly wages, and while the gap varies across different conditions, it remains statistically significant. The inclusion of additional covariates improves the models' explanatory power, providing a comprehensive understanding of the factors influencing the gender wage gap in the given dataset.