DA2

Assignment 1

DATA SELECTION

Selection of the “Food preparation and related services” sector: beyond professional interest in that sector, it has specific characteristics on terms of working hours, low educational level required, high pressure on cost (hence on wages), extremely fragmented, …

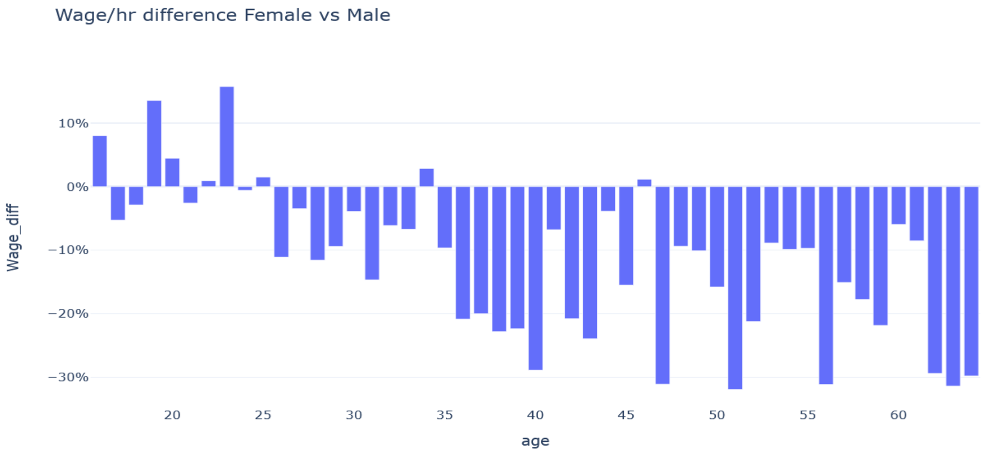
It represents 5% of the total population in the dataset with 8694 observations. The average profile is much younger with 32 years of age (vs 40 for the total population). Working hours are less with 31 hours per week (vs 39).

The data is clean. Outliers are observed in terms of wage per hour: some people declared over 2000 USD per week with less than 10 hours of work which puts the wage per hour at some extreme!

INITIAL DATA ANALYSIS

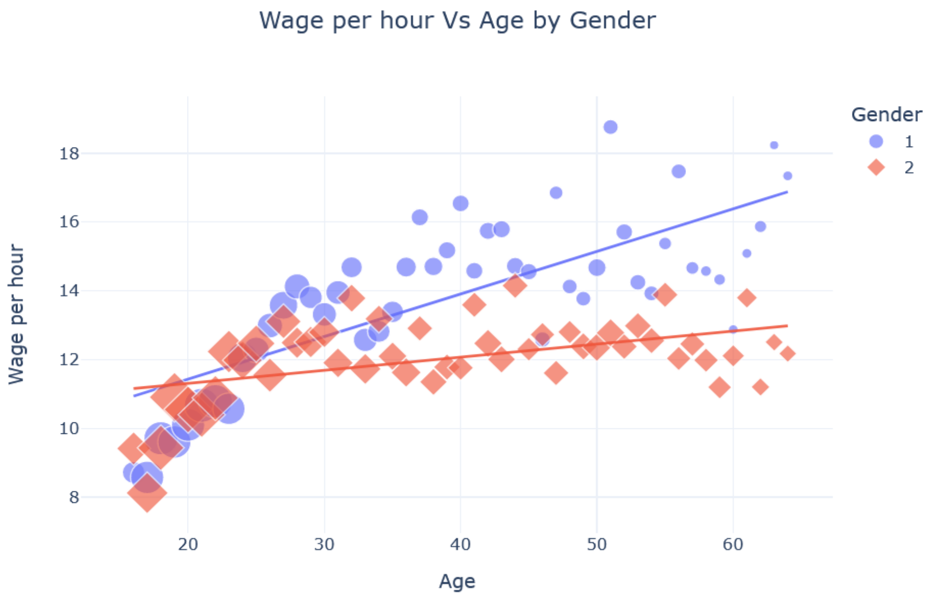
Wage gap is much lower than the population average with an unconditional bias at -7% vs -22%.

Interestingly there is a positive bias on female workers under the age of 25, whilst the bias is increasingly negative beyond that age.



An initial view of hourly wages by gender illustrates é to 3 different stages:

- The initial stage previously mentioned which is favorable to women (influence of tips?)
- A second stage where female salary tends to plateau whilst male salaries increase
- A possible final stage beyond 45-50 years old, more fragmented, where men seems to be fairly steady as well.



REGRESSION ANALYSIS AND FURTHER CONSIDERATIONS

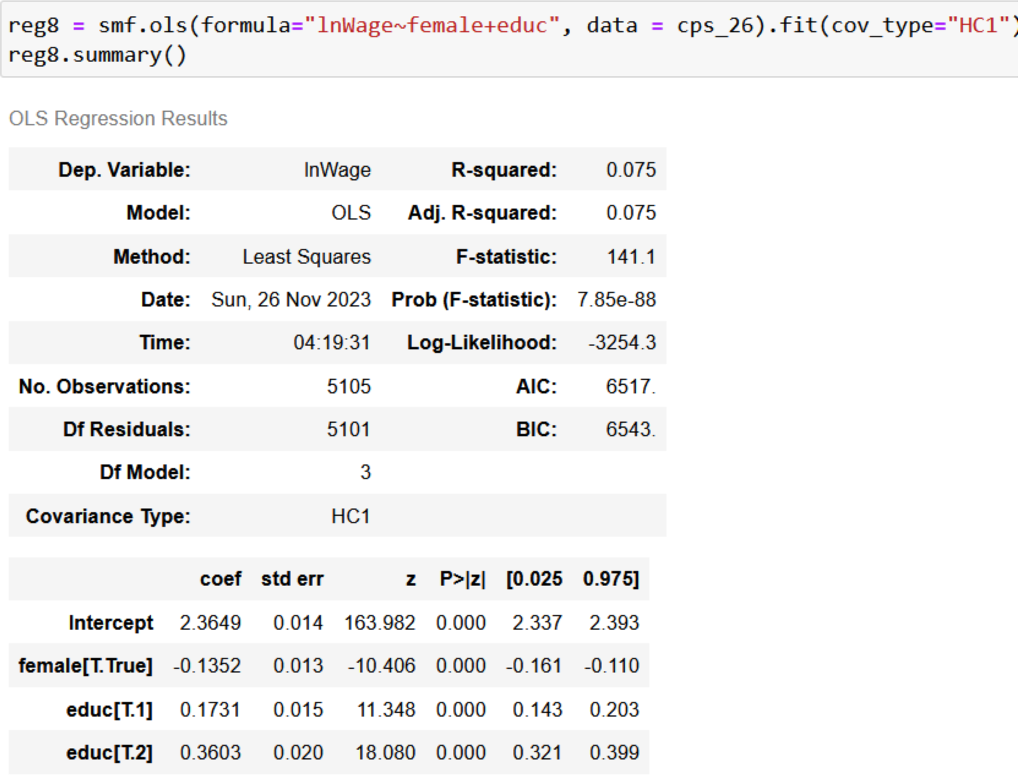
Not surprisingly, regression against log of wages is stronger than against wages. The variable “age” is giving limited results though: whilst dependency of wages over age is confirmed, it is however limited in its scope with only 0.8 log point per year of age. Gender bias increases slightly to 9 pts.

Education was analyzed both with raw data (indices 31 to 46). Whilst this index gives more data to play with, it is not proportional, with too high a weight of pre-high school education. By combining it into 3 categories (“no diploma”, “high school” and “graduate”), the results obtained are more sensible and give better results: impact of high school is estimated at +18% for attending high school and +39% for graduation/diplomas. In this scenario, the gender bias remains “low” at 7.7.

To get better regression results, we split the data, with regression analysis for people aged over 25. In that scenario, correlation increases. The impact of age becomes 0.2 log point per year and the gender bias raises to 14.4.

Multiple regression integrating age, education and child presence gave the highest correlation but the simple regression of log(wages) depending on education gave a strong result with -13.5 log points of female bias, 17 log points for high school education and 36 log point for graduate education.

In all regression work done, 95% confidence intervals range from 5 to 6 points on the gender bias. P tests all reject the null and all variables confirm the dependance of wages over gender. Not to the point of demonstrating any causality however.



```
reg8 = smf.ols(formula="lnWage~female+educ", data = cps_26).fit(cov_type="HC1")
reg8.summary()
```

OLS Regression Results

Dep. Variable:	lnWage	**R-squared:**	0.075
Model:	OLS	**Adj. R-squared:**	0.075
Method:	Least Squares	**F-statistic:**	141.1
Date:	Sun, 26 Nov 2023	**Prob (F-statistic):**	7.85e-88
Time:	04:19:31	**Log-Likelihood:**	-3254.3
No. Observations:	5105	**AIC:**	6517.
Df Residuals:	5101	**BIC:**	6543.
Df Model:	3		
Covariance Type:	HC1		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	2.3649	0.014	163.982	0.000	2.337	2.393
female[T.True]	-0.1352	0.013	-10.406	0.000	-0.161	-0.110
educ[T.1]	0.1731	0.015	11.348	0.000	0.143	0.203
educ[T.2]	0.3603	0.020	18.080	0.000	0.321	0.399