Analytical Report - DA 03 HMW 1

Submitted by: Hanna Asipovich, MSc'23 on January 26, 2023

1.Introduction

The purpose of this assignment is to create four predictive models for earnings per hour using linear regressions. I chose occupation 2100 Lawyers, Judges, Magistrates and other judicial workers(1,027 obs.).

2. Data Analysis and Transformation

In the process of data exploration, I carried out data summaries for the data overview. E.g. the sample is varied with a high enough number of women (405) and men (615). Hence, we will use the gender variable for the analysis and carried out respective transformation into a binary. I have also looked into a race variable, however, the number of non-white representatives was rather low, so it was not taken into consideration in the end. The data contains observations for people from 21 years to 64 years old. This may be important in our consideration of the drop in the pay level for older women based on earlier retirement or the education level gap in the older generations. One more variable, quadratic age predictor, is created for model building. In the data filtering process, education level is included from college to PhD due to the high level of competence required for the job. Only at least 20 hours a week jobs are included to ensure that our comparison is valid. In further munging and transformation, for the target variable I created hourly wage_**(earnhrs)**_ through dividing the weekly earnings (earnwke) by the number of hours (uhours). For additional consideration, I also created the log of this variable _**(ln_earnhrs)**_, which however I dropped after seeing the skewed distribution of the log(see Annex).

3. Variables, their interactions, and regression models

Gender showed as a strong predictor for earnings per hour. Based on a simple linear regression, on average women in this professional category earn 4 USD per hour less than men. Further on, we can see in the loess graph in the Annex that older age also leads to an increase in earnings. However, it is also interesting to see interaction on gender and age (see Annex), as clearly there is a different dynamic for women in their middle careers (quite likely connected to the maternity leaves and childcare) and further drop closer to 64 years old (possibly connected to an earlier retirement). To further capture the interplay of independent variables, interactions are used on gender and education. Four linear regression models are built to prediction analysis. Model 1 is the simplest containing gender, Model 2 has the Model 1 explanatory variable along with age and age squared. In the Model 3, more explanatory variables on education levels are added. Model 4, is the most complex as it contains all the mentioned independent variables and the respective interaction on gender and education.

4. Comparison of models performance After building models, Model 4 looks the best positioned based on the lower average RMSE and highest R2. However, there is only insignificant difference with model 3. Most likely, due to the fact that pay level difference within the same higher category of education for both genders is not statistically significant, despite the fact that the combined regression table shows that income on average is still lower to highly educated females as opposed to males who are highly educated. BIC as the measure of fit penalizes the model complexity and helps to avoid over-fitting, so lower BIC models are preferred. During cross-validation, among the models, Model 3 has the lowest BIC and cross validation RMSE average as it is less complex.