# DA3 HOMEWORK1

Shahana Ayobi

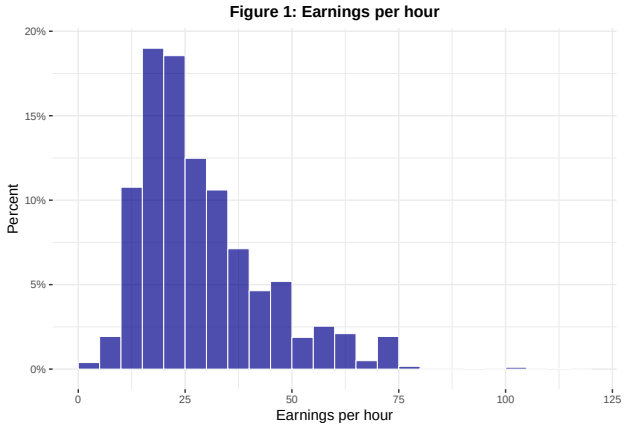
2023-01-21

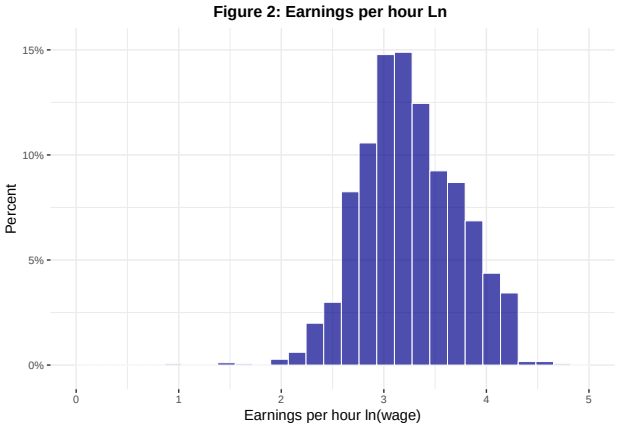
## Introduction

The purpose of this report is to develop a prediction model for *Accountants and Auditors* occupation in the United States using the data from OSF. The prediction models in this study were created using OLS regressions, and the best model was determined by considering lowest BIC and RMSE score as well as performing 4-fold cross-validation using the average RMSE of the individual models. A total of 4 models were created, each of which was initially tested on the entire sample before being divided into 4 folds for training and cross-validation.

## Data Cleaning

The OSF data was filtered for the occupation of **Accountants and Auditors** with census code of 0800 and 1812 observations. Starting withe the independent variable, the hourly wage is calculated by dividing the weekly earnings **earnwke** by the number of hours **uhours**, and the log of the mentioned variable **lnwage** is also created. Additionally, the chosen data set is modified to include observations with hourly wages of at least one US dollar. The histogram shows that some accountants and auditors have higher wages which pulled mean to the right and made the distribution of wage rightly skewed.




The **Female** variable was transformed into one if a person is female is zero otherwise. Other variables are created and transformed to model, including education levels, ethnicity, marital status, and whether an individual has a child. The data is filtered for observations with more than 20 years of age and a quadratic age predictor is created to model non-linearity in age. The education levels below college are added to the **No Diploma** category, and both vocational and academic associate certificates are added to the associate category.The rest of categories include having a BA, MA, PhD, and Professional degree.

Figure 3: Loess: Wage ~ Age

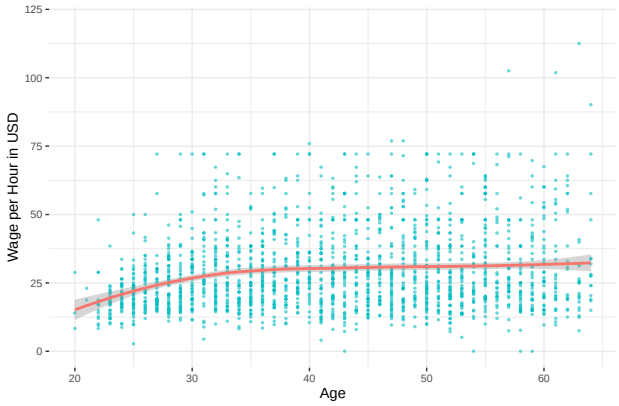
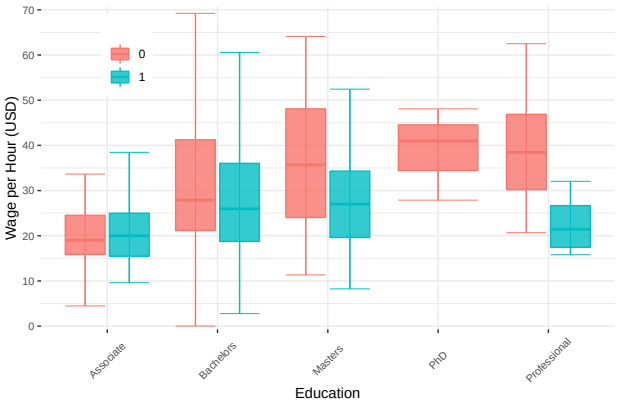



Figure 4: Wage Distribution for Male and Female for Educational Categories




# Models and Predictors

Education is likely one of the main predictors of earnings, thus the the simplest model, Model 1, includes all above mentioned categories taking the **No Diploma** as the base category. As Figure 4 shows that except for the Associate degree, female accountants and auditors tend to earn lower than their male counterparts. Also, there are only 5 female observations that have PhD in this occupation, this is way it is not shown in the graph(Check Table 5 on Appendix). The second model add more variables such as age and age squared since as a person ages, they gain work experience, thus, increasing their wages. Female variable is also added since there is a clear pay gap of 6.1 USD as shown in Table 4 in the Appendix. Third model includes more variables like marital status, and having a child since it is apparent from table 7 in the Appendix that married individuals earn 4.12 USD more. In the more complex model, Model 4, interactions are used to further capture the interaction of independent variables. Gender and education interactions, marital status and gender interactions, gender and having a child, and age and having a child interactions are added.

# Model Performance

A model's fit is measured using all of the original data, and the BIC penalizes model complexity and aids in preventing over-fitting. In general, models with a lower BIC are recommended. Model 4 of the models has the lowest BIC. The root mean squared loss over a number of target observations, or RMSE, is the second metric used to assess model performance. Once again the most complex model with 19 coefficients, Model 4, has the lowest RMSE. Considering these measures, the Model 4th is considered to be the best wage prediction model which explains approximately 36% of variation in wage.

Table 1: Evaluation of Models

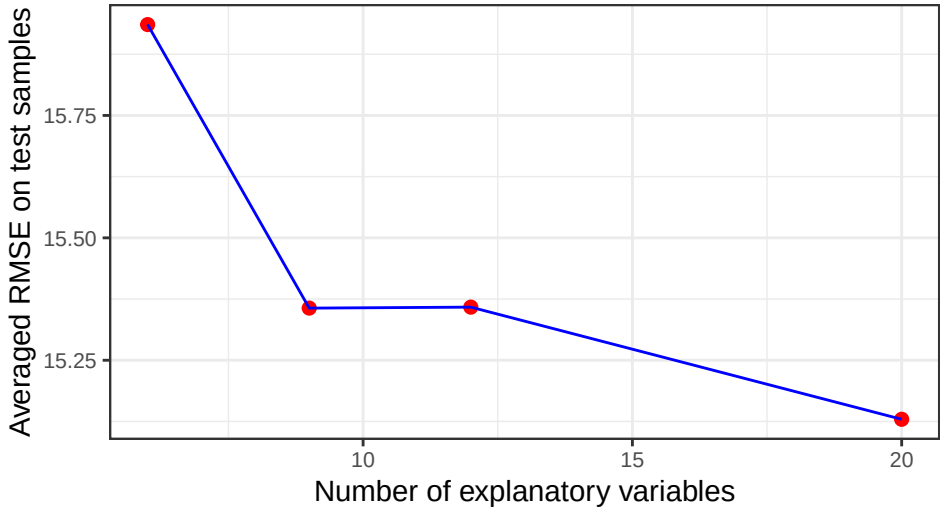
| Model | N predictors | R-squared | Training RMSE | BIC |
|---|---|---|---|---|
| (1) | 5 | 0.1378 | 15.3287 | 15087.29 |
| (2) | 8 | 0.2031 | 14.7368 | 14967.08 |
| (3) | 11 | 0.2069 | 14.7009 | 14980.75 |
| (4) | 19 | 0.3585 | 13.2214 | 14656.35 |

The result of the 4-fold cross validation after training the models on the training sets and validating it on the test sets, shows the 4th Model as the best one with lowest average RMSE of approximately 15.13. While the Average RMSE decreases at first as models become more complex, it slightly increases for the third model and decreases back on the 4th, suggesting that the 4th Model is preferred.

Table 2: Four Fold Cross Validation Average RMSE

| Resample | Model1 | Model2 | Model3 | Model4 |
|---|---|---|---|---|
| Fold1 | 19.8219 | 19.3579 | 19.3972 | 19.5081 |
| Fold2 | 14.7037 | 13.8687 | 13.9380 | 13.3082 |
| Fold3 | 14.2690 | 13.7137 | 13.6377 | 13.0961 |
| Fold4 | 14.2501 | 13.7171 | 13.6758 | 13.6536 |
| Average | 15.9356 | 15.3565 | 15.3585 | 15.1295 |

**Figure 5: Prediction Performance and Model Complexity**



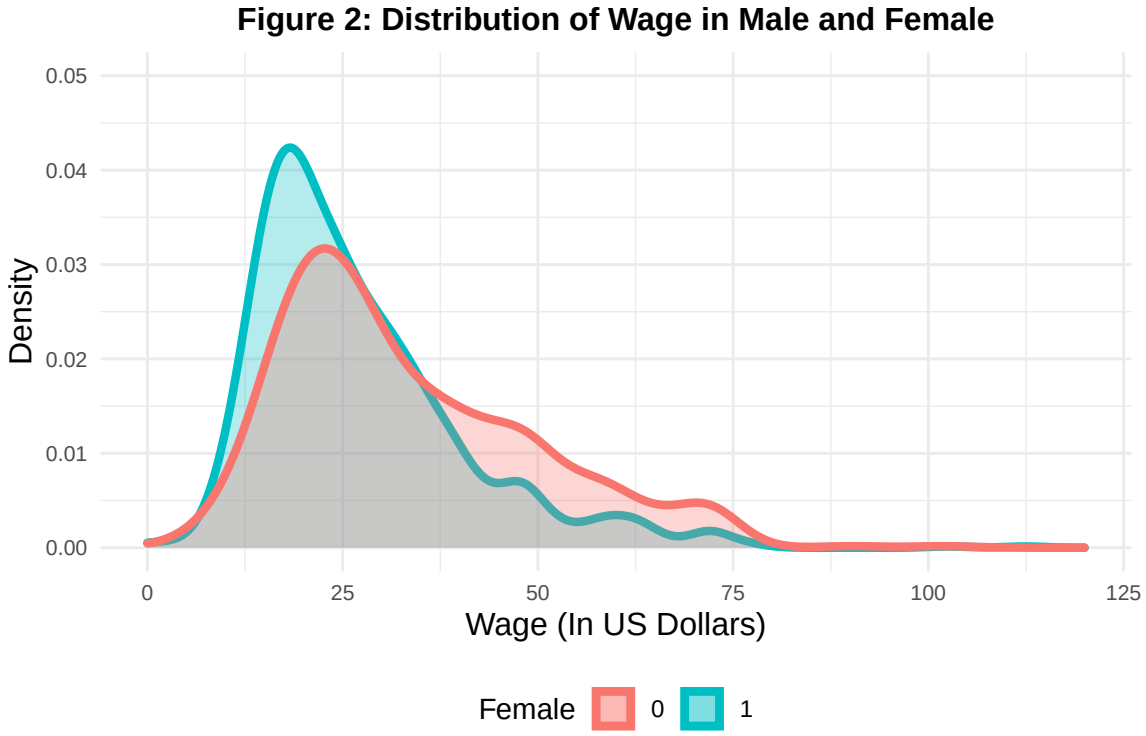
# Appendix

### Summary Statistics

Table 3: Summary Statistics

| | Mean | Median | SD | Min | Max | P05 | P95 | N |
|---|---|---|---|---|---|---|---|---|
| Earnings per hour | 29.30 | 25.00 | 16.51 | 0.01 | 346.15 | 12.50 | 59.91 | 1812 |
| Female | 0.66 | 1.00 | 0.47 | 0.00 | 1.00 | 0.00 | 1.00 | 1812 |
| Associate | 0.10 | 0.00 | 0.30 | 0.00 | 1.00 | 0.00 | 1.00 | 1812 |
| BA Degree | 0.57 | 1.00 | 0.49 | 0.00 | 1.00 | 0.00 | 1.00 | 1812 |
| MA Degree | 0.18 | 0.00 | 0.38 | 0.00 | 1.00 | 0.00 | 1.00 | 1812 |
| Professional Degree | 0.01 | 0.00 | 0.07 | 0.00 | 1.00 | 0.00 | 0.00 | 1812 |
| PhD | 0.00 | 0.00 | 0.05 | 0.00 | 1.00 | 0.00 | 0.00 | 1812 |
| Age | 42.30 | 42.00 | 11.39 | 20.00 | 64.00 | 25.00 | 60.00 | 1812 |
| Age Squared | 1918.64 | 1764.00 | 975.10 | 400.00 | 4096.00 | 625.00 | 3600.00 | 1812 |
| Has child | 0.42 | 0.00 | 0.49 | 0.00 | 1.00 | 0.00 | 1.00 | 1812 |
| White | 0.81 | 1.00 | 0.39 | 0.00 | 1.00 | 0.00 | 1.00 | 1812 |
| Marital Status | 0.65 | 1.00 | 0.48 | 0.00 | 1.00 | 0.00 | 1.00 | 1812 |

### Distribution of Wage in Male and Female



**Summary Tables for Variables**

Table 4: Summary of Wages for Both Genders

| | Female | Mean | SD | Min | Max | P25 | P75 | N |
|---|---|---|---|---|---|---|---|---|
| wage | 0 | 33.32 | 16.15 | 0.02 | 101.85 | 21.15 | 43.27 | 616 |
| | 1 | 27.22 | 16.32 | 0.01 | 346.15 | 17.50 | 33.37 | 1196 |

Table 5: Education Categories for Female

| | Education | Mean | N | Percent | Min | Max |
|---|---|---|---|---|---|---|
| female | Associate | 0.83 | 187 | 10.32 | 0.00 | 1.00 |
| | Bachelors | 0.63 | 1039 | 57.34 | 0.00 | 1.00 |
| | Masters | 0.53 | 322 | 17.77 | 0.00 | 1.00 |
| | PhD | 0.20 | 5 | 0.28 | 0.00 | 1.00 |
| | Professional | 0.40 | 10 | 0.55 | 0.00 | 1.00 |

Table 6: Summary of Wages for Education Categories

| | Education | Mean | N | Percent | Min | Max |
|---|---|---|---|---|---|---|
| wage | Associate | 21.42 | 187 | 10.32 | 0.01 | 72.12 |
| | Bachelors | 30.94 | 1039 | 57.34 | 0.02 | 102.53 |
| | Masters | 33.66 | 322 | 17.77 | 8.25 | 112.50 |
| | PhD | 107.05 | 5 | 0.28 | 27.88 | 346.15 |
| | Professional | 32.80 | 10 | 0.55 | 15.80 | 62.50 |

Table 7: Summary of Wages for Marital Status

| | Marital Status | Mean | N | Percent | Min | Max |
|---|---|---|---|---|---|---|
| wage | 0 | 26.61 | 631 | 34.82 | 0.01 | 101.85 |
| | 1 | 30.73 | 1181 | 65.18 | 0.01 | 346.15 |

Table 8: Summary of Wages for White and Non-white

| | White | Mean | N | Percent | Min | Max |
|---|---|---|---|---|---|---|
| wage | 0 | 29.17 | 338 | 18.65 | 4.47 | 346.15 |
| | 1 | 29.33 | 1474 | 81.35 | 0.01 | 112.50 |

Table 9: Simple Regressions Result

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Intercept | 21.0076** | −9.2758 | −4.3842 | −3.5701 |
| | (0.9730) | (5.0007) | (5.4934) | (5.3299) |
| Associate Degree | 0.4079 | 0.3494 | 0.2871 | −1.1799 |
| | (1.4858) | (1.4297) | (1.4277) | (3.1457) |
| BA Degree | 9.9343** | 9.9539** | 9.9470** | 9.6776** |
| | (1.0834) | (1.0584) | (1.0577) | (2.1443) |
| MA Degree | 12.6574** | 12.3488** | 12.3028** | 13.6645** |
| | (1.2957) | (1.2720) | (1.2715) | (2.3063) |
| Professional Degree | 11.7888* | 11.0462* | 11.2506* | 14.6184* |
| | (4.9520) | (4.7764) | (4.7695) | (5.8063) |
| PhD | 86.0377** | 83.5690** | 83.4426** | 22.7528** |
| | (6.9352) | (6.6919) | (6.7005) | (6.9844) |
| Age | | 1.3325** | 0.9893** | 0.9165** |
| | | (0.2441) | (0.2749) | (0.2540) |
| Age Squared | | −0.0120** | −0.0078* | −0.0075* |
| | | (0.0029) | (0.0032) | (0.0030) |
| Female | | −4.6194** | −4.5894** | −2.7700 |
| | | (0.7571) | (0.7567) | (2.4384) |
| White | | | 0.6658 | 1.0437 |
| | | | (0.9038) | (0.8170) |
| Has Child | | | 2.0761* | −1.5418 |
| | | | (0.8736) | (3.3519) |
| Marital Status | | | 0.4387 | 3.1366* |
| | | | (0.8367) | (1.3991) |
| female:marital_status | | | | −3.8306* |
| | | | | (1.5918) |
| Associate:female | | | | 1.8913 |
| | | | | (3.4481) |
| BA:female | | | | 0.2086 |
| | | | | (2.3997) |
| MA:female | | | | −2.9826 |
| | | | | (2.6977) |
| Prof:female | | | | −9.4247 |
| | | | | (8.8950) |
| PhD:female | | | | 302.7085** |
| | | | | (15.0603) |
| female:child | | | | 1.3916 |
| | | | | (1.5138) |
| age:child | | | | 0.0689 |
| | | | | (0.0736) |
| Num.Obs. | 1812 | 1812 | 1812 | 1812 |
| R2 | 0.138 | 0.203 | 0.207 | 0.359 |

* $p < 0.05$, ** $p < 0.01$