DA3 - A1

Aftab

Introduction

This paper tried to build a prediction model for *Receptionists and Information Clerks* in the U.S. from the data obtained from OSF. The paper has used OLS regressions to build the prediction models and a 5-fold cross-validation to arrive at the best model using the individual models' average RMSE; in total, 4 models were built that were initially run on the full sample and then were split into 5 folds for training and cross-validation.

Data Engineering

Starting with the sample, the paper focused on the code 5400 that refers to receptionists and information clerks in the dataset. To make sure that an apples-to-apples comparison was being made, the variable wage per hour (w) was created. The sample was then filtered to contain observations with worked hours of 40 or more per week. Observations were also filtered on age greater than or equal to 18. These two filters were made to proxy for full-time workers as per minimum working hours and minimum legal age for working in the country.

The sample size was initially 1478, which reduced to 898 observations after the above mentioned filters. The observations were then filtered to check for number of NAs in the sample; the variable ethnic returned 731 NAs and was hence dropped. That said, the race variable was available, which was coded into a binary variable of 'white' and 'other' races. Similarly, the education variable 'grade92' was coded into a factor variable (educ) of 6 levels, starting from 'no diploma' to 'master' degree. The observations with a PhD and professional degree were dropped for two reasons; the number of observations in total for both were 6, it is not a common occurrence for PhD and professional degree holders to choose this occupation.

Other factors such as marriage status were clubbed together into a factor variable containing 'separated, never married, married'. Gender was put into a binary variable of male and female. Individual states were clubbed into 4 regions as per the BLS division of the U.S.; 'west, mid-west, south, east'. This decision was made to make it feasible to interact it with other variables in the sample. Similarly, citizenship or birth region was divided into born in 'Native, Born in PR or US Outlying Area' or not compared to its original variable that contained 5 different types of entries.

Model Building

In total, 4 models were built with model 1 being the simplest and model 4 being the most complex. Whereas, for the y variable, wage/hour was taken in absolute terms, even though its distribution mimicked a log-normal distribution; purely for the simplicity of the models. For the RHS variables, the core focus was on age which was considered a proxy for work experience, education being a proxy for skill, and gender to cater for gender-based wage differences.

Model 1 only contained education as the predicting variable but model 2 contained education, age, age-squared, and gender as the predicting variables. Second-degree polynomial of age was used to capture the quadratic association of age and wage per hour. For model 3, it included all variable from model 2 plus race and marriage status dummy variables. This model also included the interaction terms necessary to capture

Table 1: Model evaluation based on full sample RMSE and BIC

Model	BIC	RMSE	No. of coeff
Model 1	5,902.4	6.7637	5
Model 2	5,820.2	6.3809	8
Model 3	5,891.6	6.1288	29
Model 4	5,943.8	5.8450	49

wage differences arising from individuals from different races, education levels, and marriage status. These interaction terms were added based on statistical analysis; conditional mean wage per hour were looked at for the interacted variables and where the difference seemed significant, only those were added. The appendix contains box plots that were used to check the significance of all the interactions.

Model 4 was the most complicated with model 3 RHS variables plus the additional variables, union membership, type of work organization, place of birth, region of work, number of children below 18, along with the necessary interaction terms as per the above mentioned methodology

Analysis

The models were first run on the full sample and then each model was run using a 5-fold cross-validation methodology. The resulting RMSE and BIC of the full sample regressions and cross-validated regressions are shown in the appendix.

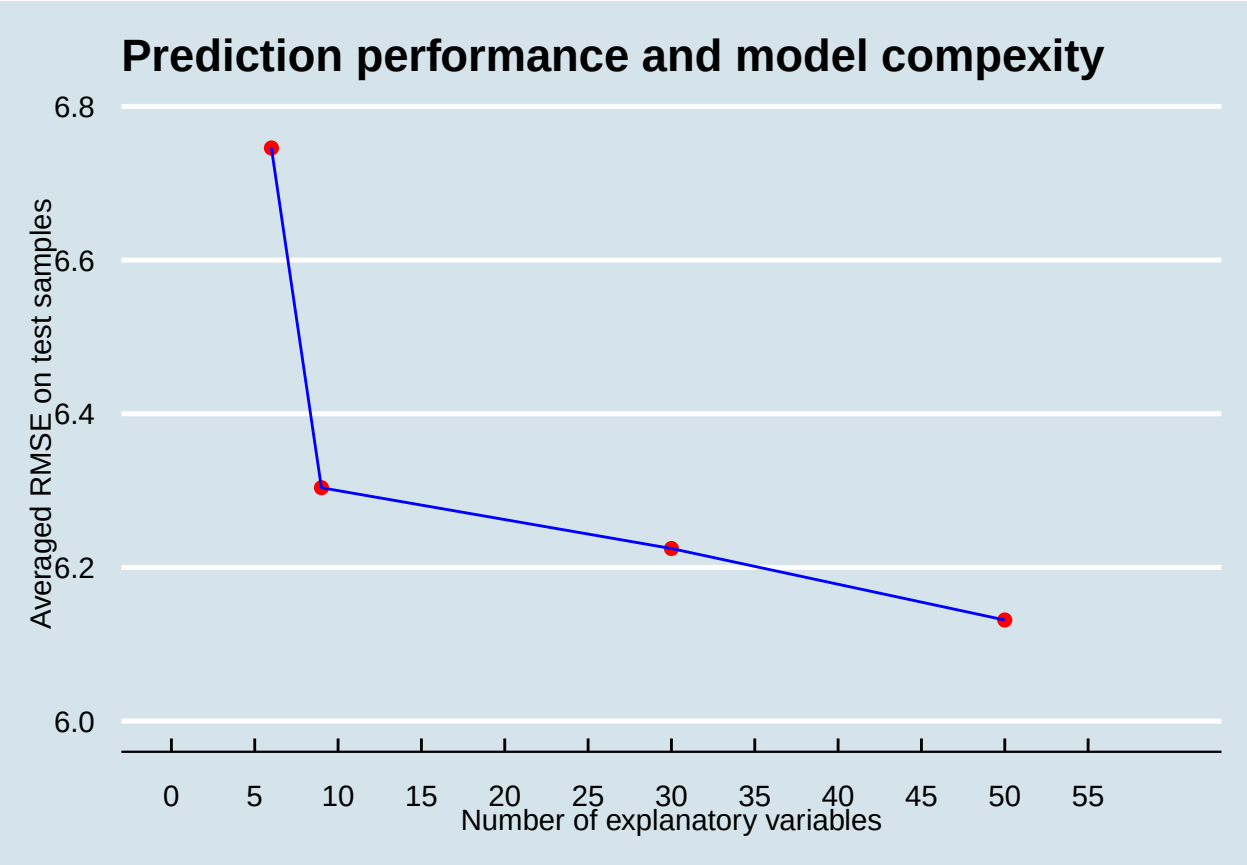
As per the full sample regressions, model 4 had the lowest RMSE, however, it can be argued that it could be because of the highest number of predictors in the model. The RMSE for the rest of the models follow a similar pattern with model 1 having the highest RMSE followed by model 2 and then model 3. Hence, looking at the BIC of these models could reveal another perspective as it penalizes for the models for additional number of RHS variables. As per BIC, the best model run on the full sample was model 2, followed by model 3, model 1 and then model 4. Based only on these results, it can be argued that model 2 is the best.

That said, averaged RMSE on test samples from the 5-fold cross-validation methodology, shown in the appendix below, suggests that model 4 is the best with the lowest averaged RMSE. However, when both complexity in terms of RHS variables and averaged RMSE, the paper suggests using model 2. The difference between model 2 and model 3 average RMSE is only around 0.2, and difference between model 2 and model 4 average RMSE is only around 0.5. When looked at these numbers from the perspective of dollar wage per hour, it can be argued that the results are not very different. Hence, this paper suggests after utilizing above-mentioned methodologies that model 2 can be best for prediction of wage per hour for the chosen occupation. Additionally, it is important that external validity is considered when using this model because of difference in time and space.

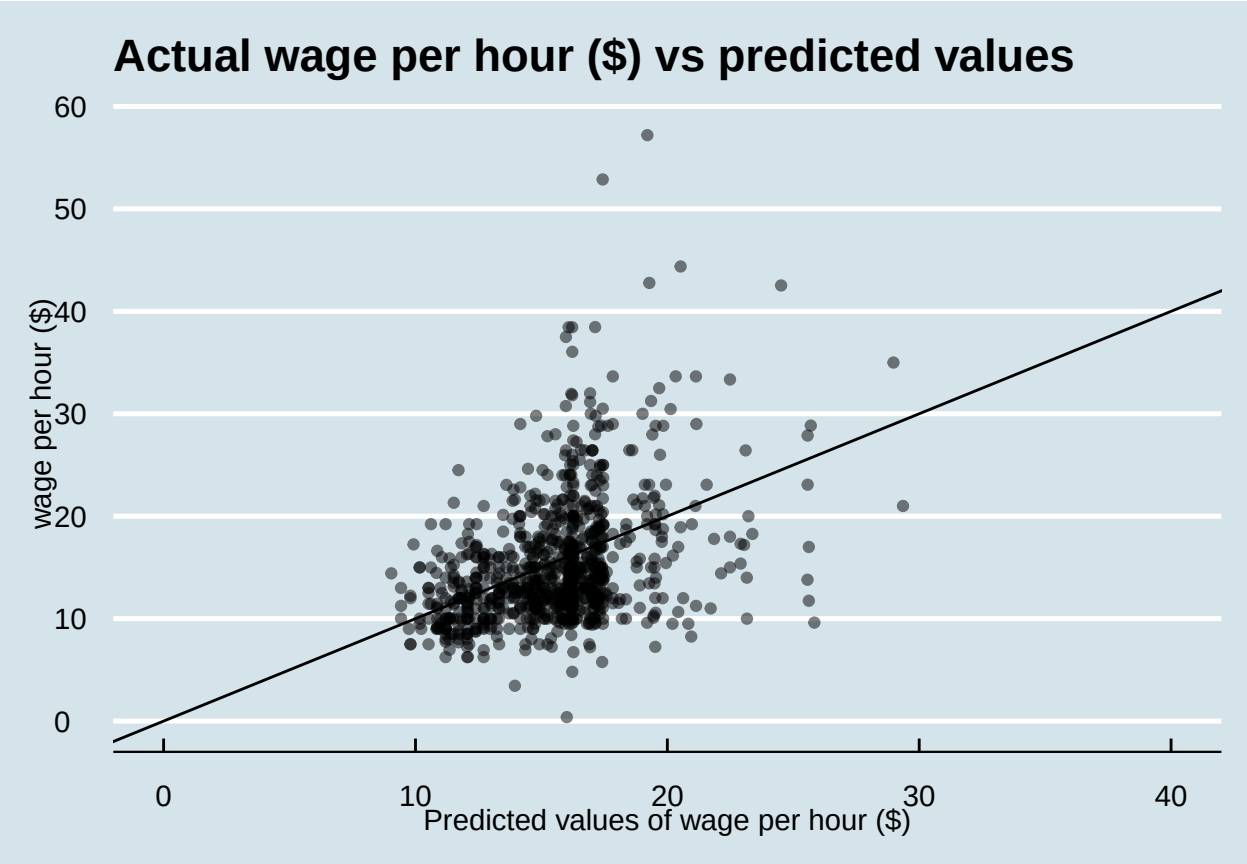
Appendix

Full sample regression stats

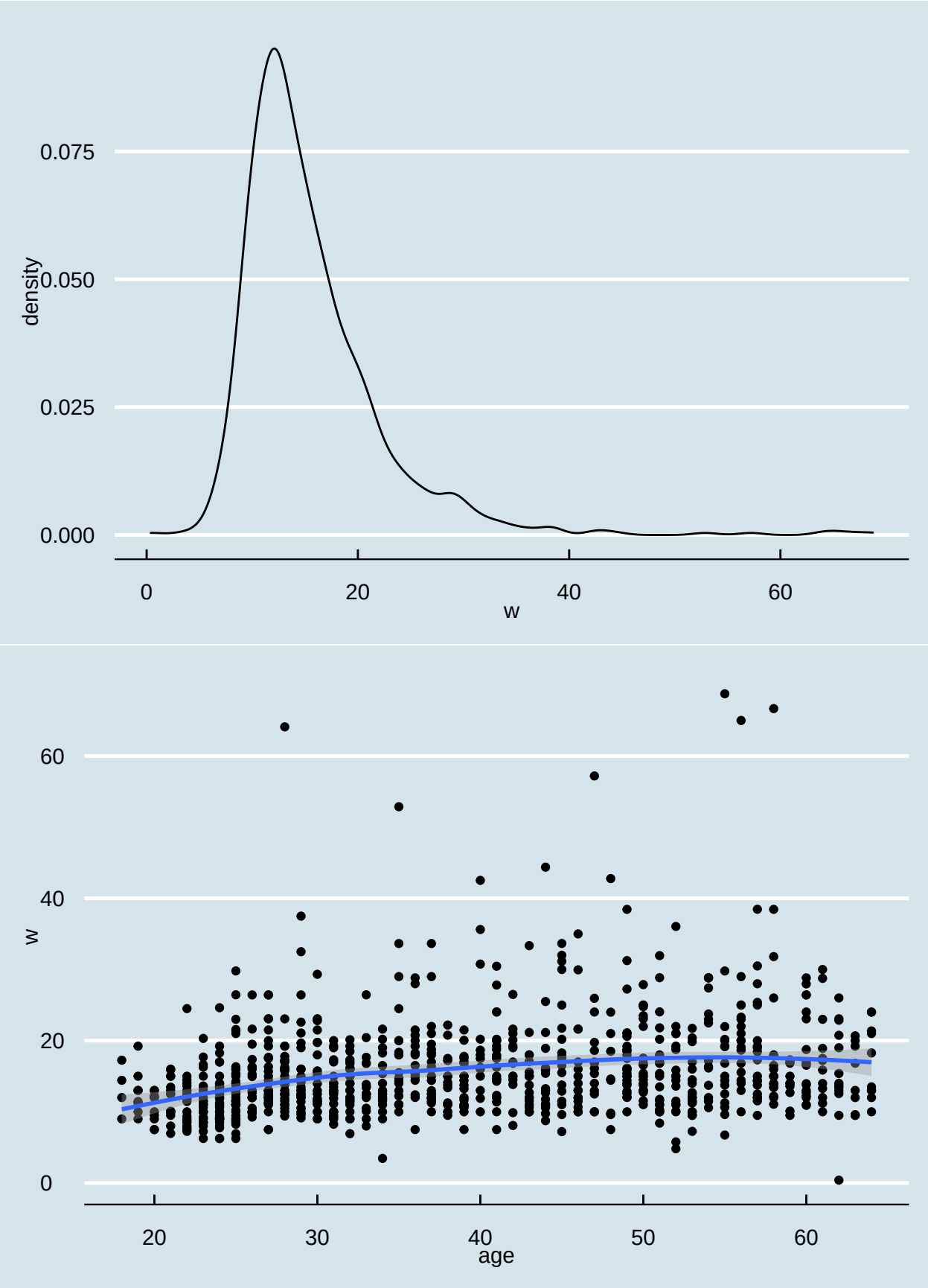
5-Fold cross-validation model complexity vs averaged RMSE



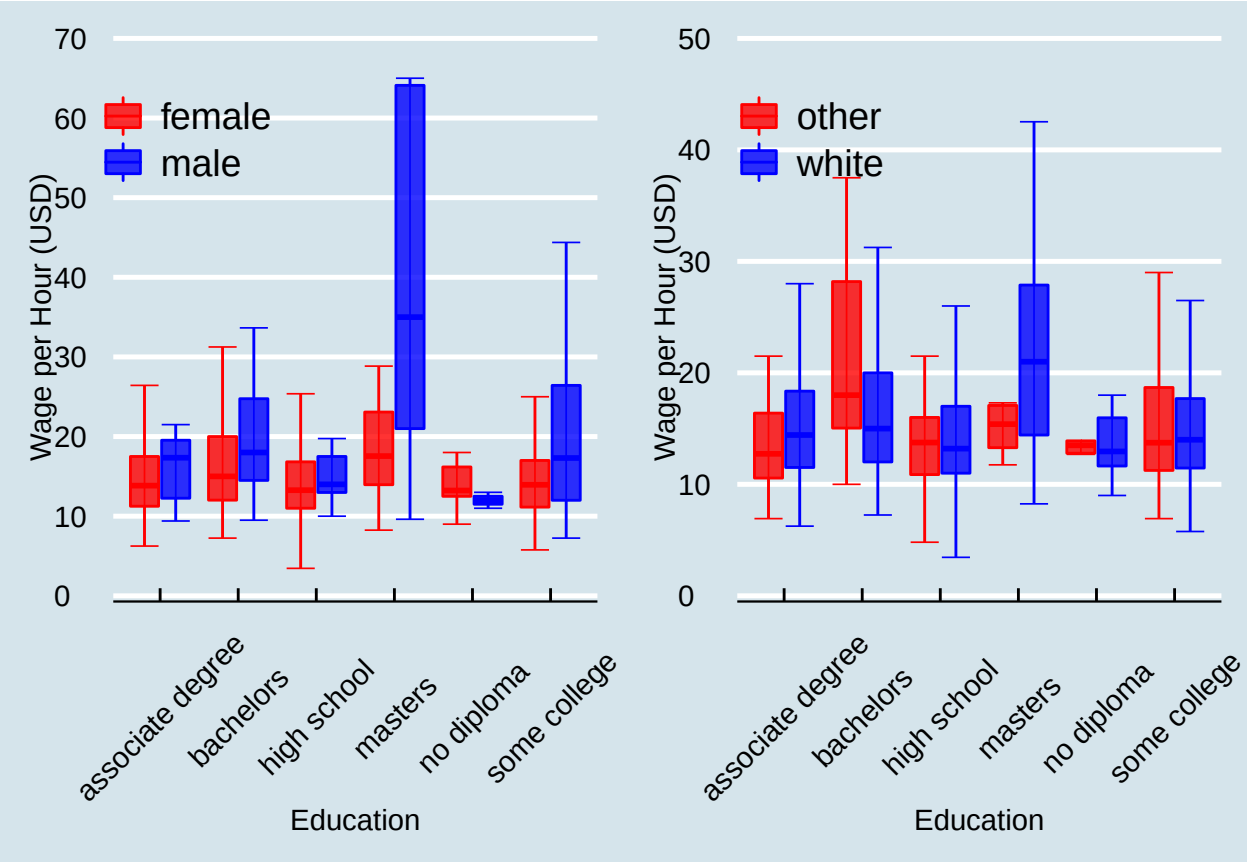
Prediction fit of model 2

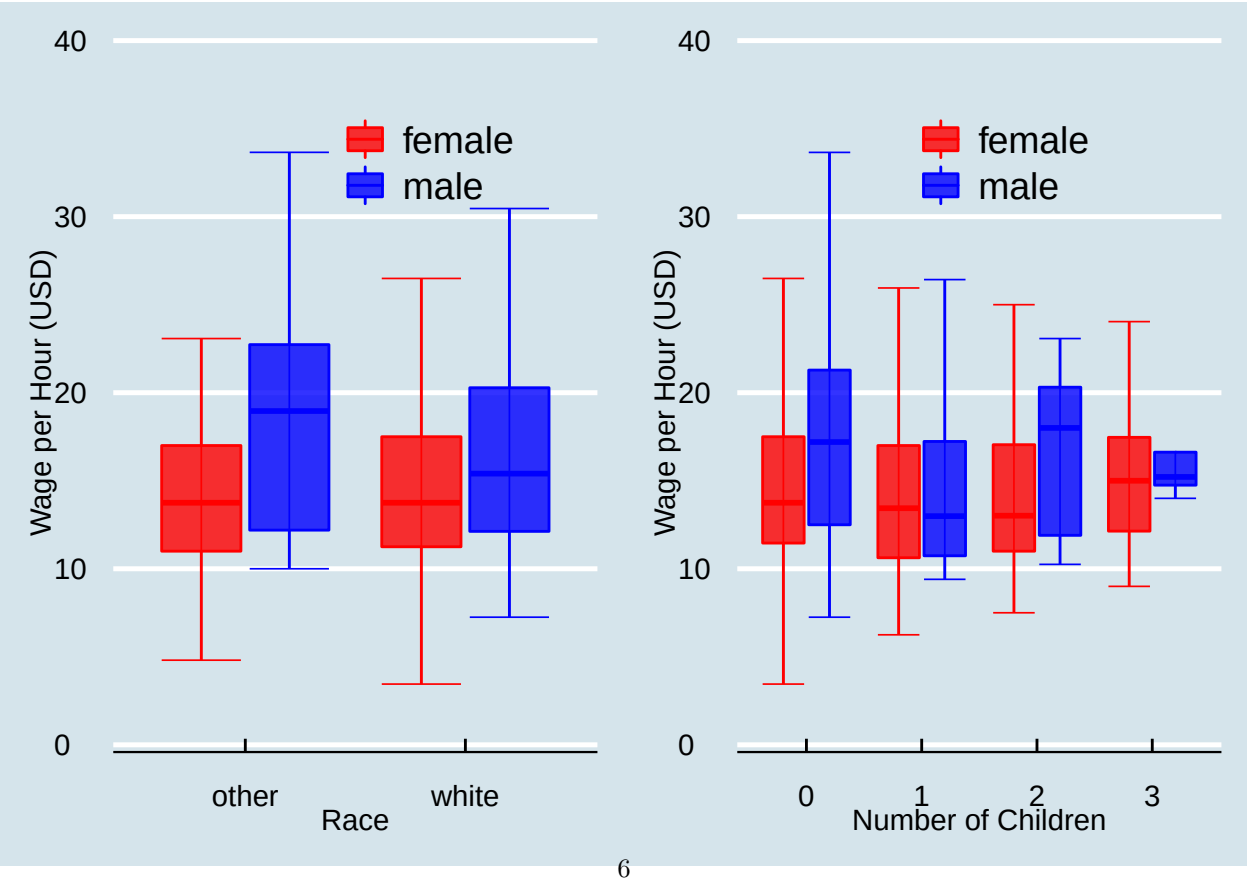


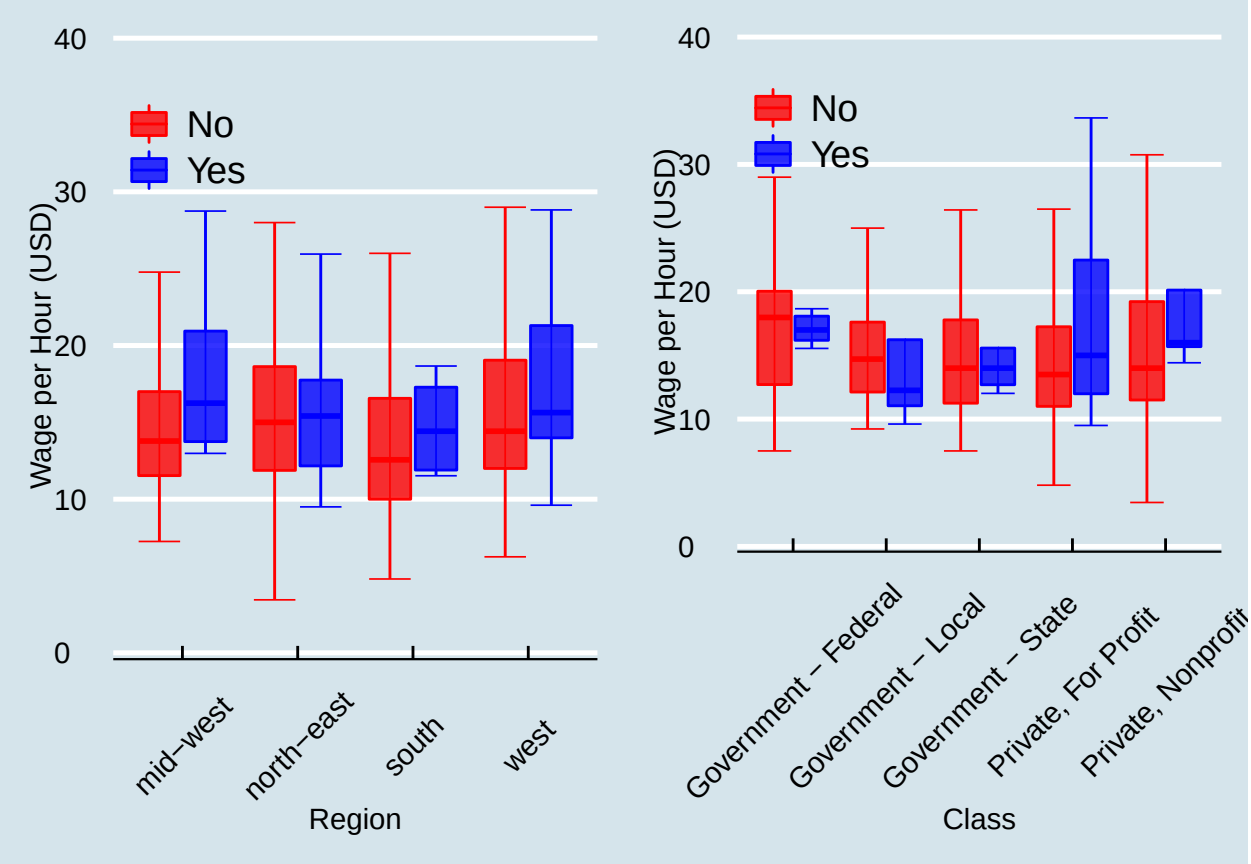
Distribution of variables

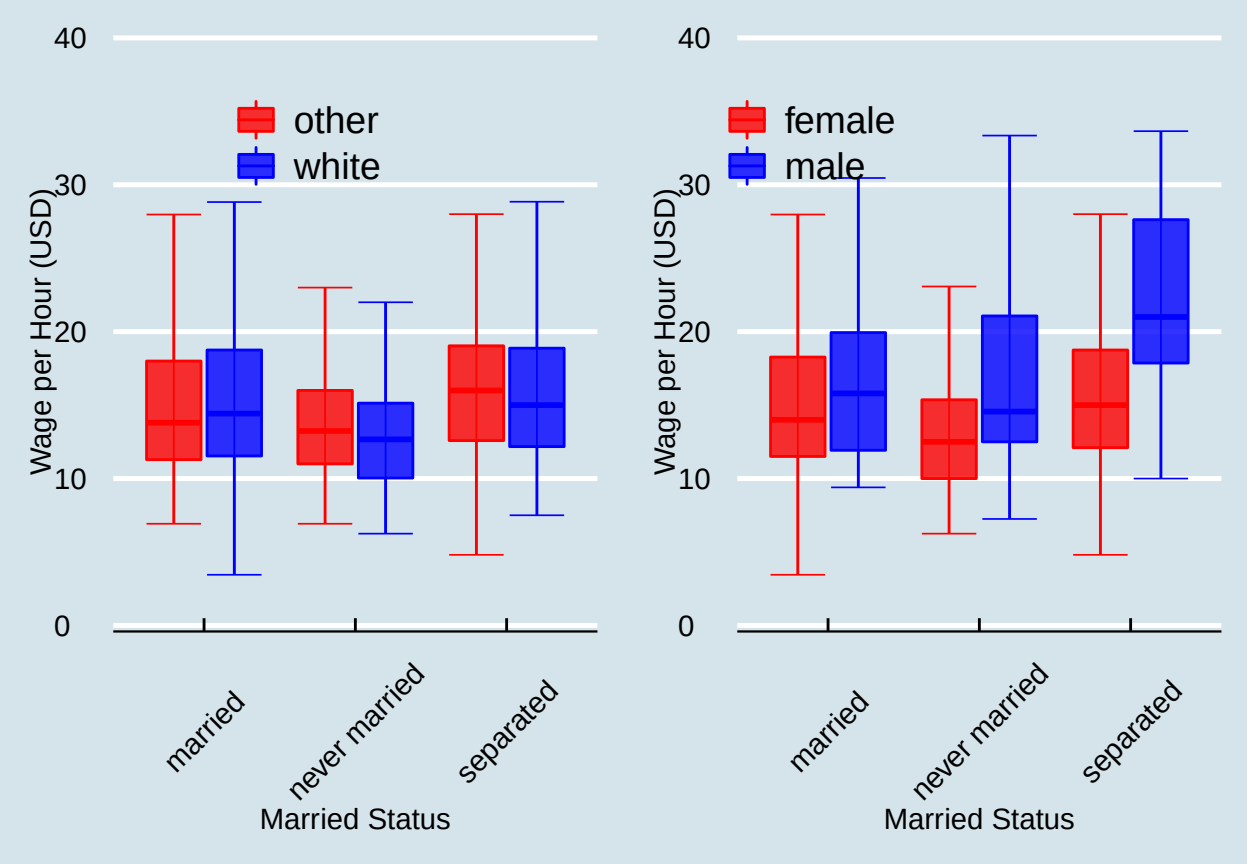


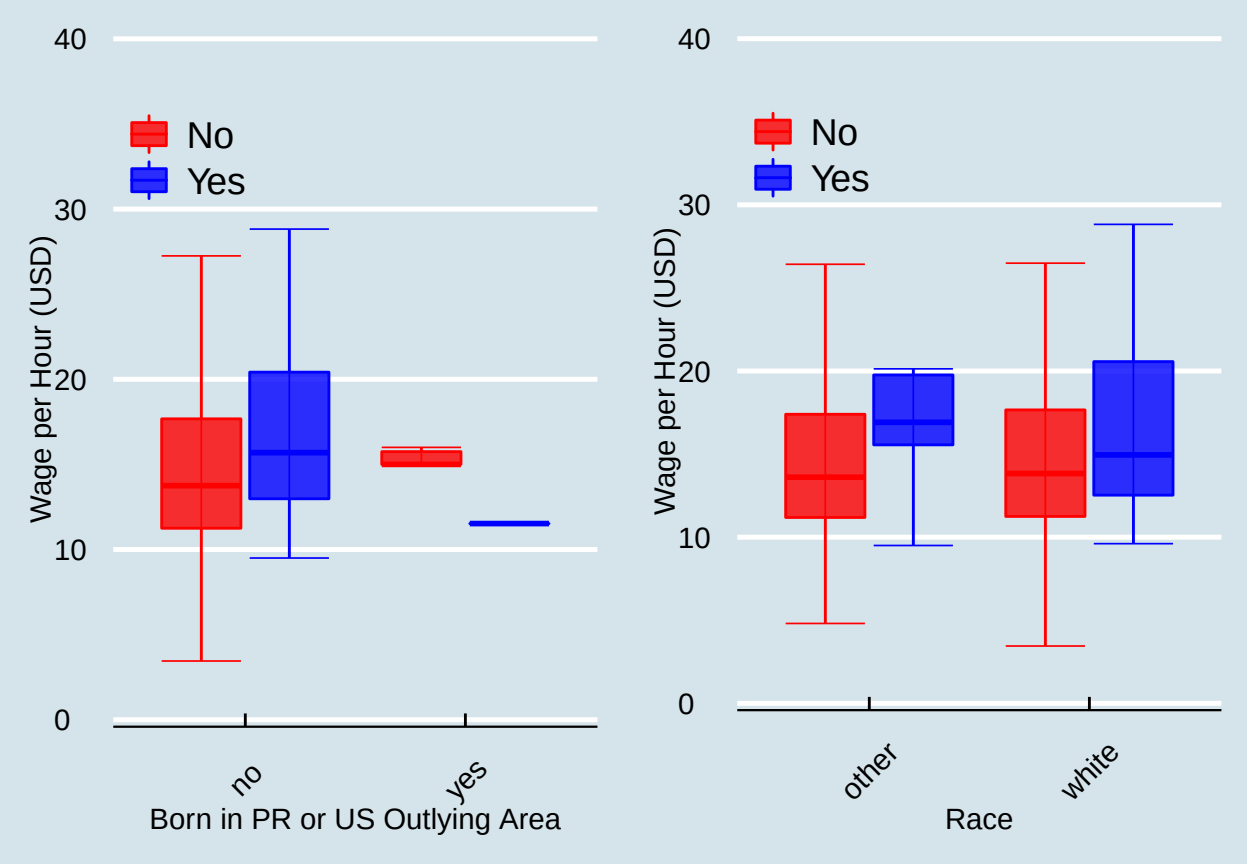
Box plots to identify interactions

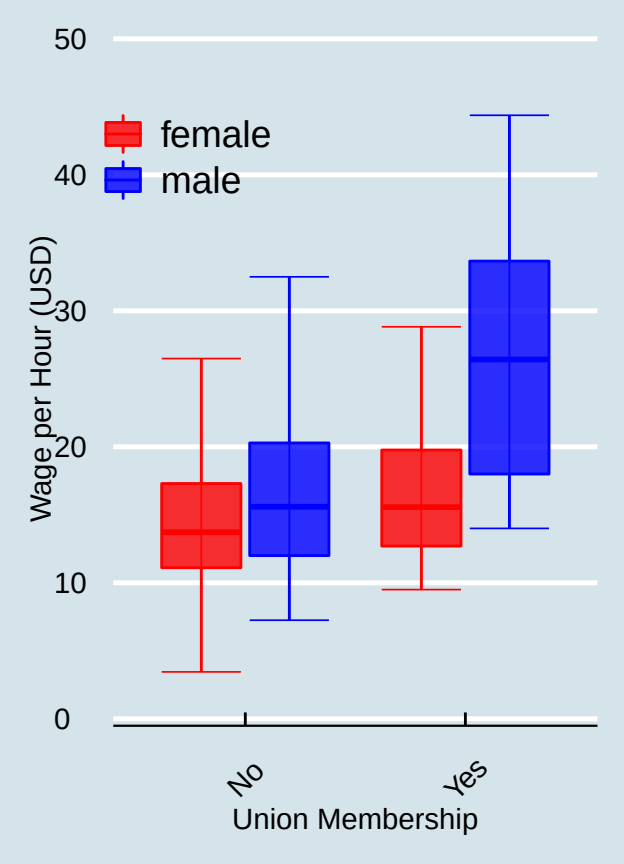





40
30
20
10
0
Wage per Hour (USD)
No
Yes
mid−west
north−east
south
west
Region
Government − Federal
Government − Local
Government − State
Private, For Profit
Private, Nonprofit
Class


40
30
20
10
0
Wage per Hour (USD)
other
white
married
never married
separated
Married Status
40
30
20
10
0
Wage per Hour (USD)
female
male
married
never married
separated
Married Status


40
30
20
10
0
No
Yes
Wage per Hour (USD)
no
yes
Born in PR or US Outlying Area
40
30
20
10
0
No
Yes
Wage per Hour (USD)
other
white
Race


50
40
30
20
10
0
female
male
Wage per Hour (USD)
No
Yes
Union Membership