Assignment 1

Balint Parragi, 2023-01-20

Predicting hourly wages of cooks

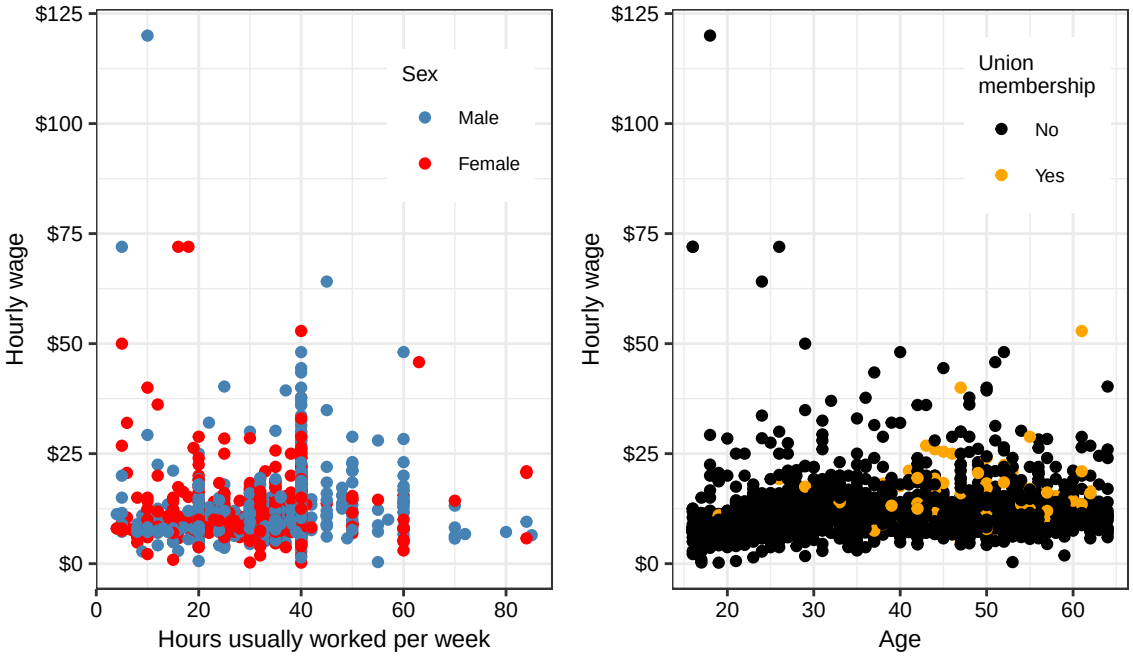
Data inspection

- Target variable is `earn_hour`: hourly wage from the **CPS earnings dataset**
- Selected occupation: cooks (id: 4020)
- Predictors can be either:
  - Quantitative: `age`, number of `own children`
  - Categorical: `sex`, `race` and `ethnicity`, `state` of residence, highest `grade` completed, `marital status`, `presence of children`, `citizenship`, `industry classificaton`, `employment class`, `union membership`, and `employment status`
- Some variables can be redundant (i.e the month of the interview), some have a lot of `NA`-values (i.e. `ethnicity`), and some categorical variables have many categories that increases the complexity of the models rapidly (i.e `state` of residence with 51 unique values, but it is not clear either - it has both character and numeric encoded values)
  - Also, `unionmme` and `unioncov` are fairly the same, there is no need for both: all *Yes* values in the former are `NA` in the latter, and the latter has only very a few (13) *Yes* values (which are all *No*-s in the former)

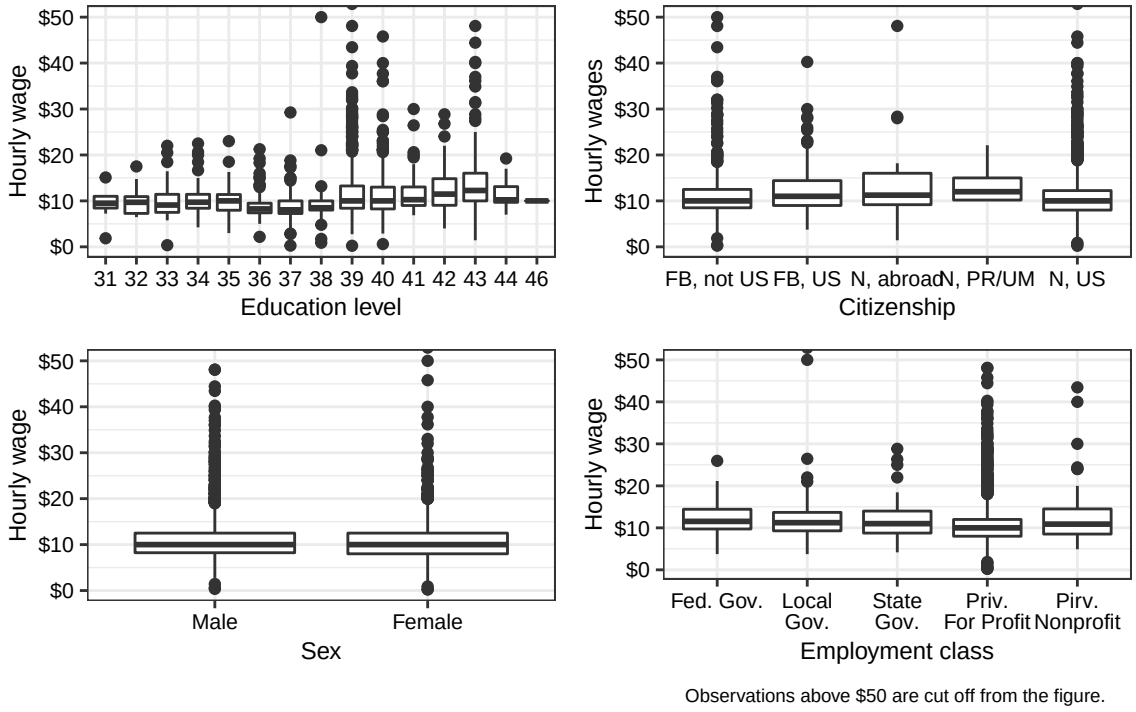
Exploratory figures

Most variables are categorical, so illustrating their relationship with hourly wages is easiest via boxplots. Main takeaways:

- Number of hours worked per week might have a negative quadratic relationship with earnings. The spikes at 20,30, and especially at 40 hours are not surprising.
- Age might be a good predictor of income, but the effect is small.
- There very few union members.



- People among the highest education levels seem to earn more on average.
- Native, US citizens tend to have the lowest average earnings, which might be surprising.
- Men earn more than women, but the difference is very small.
- There are subtle differences between employment classes.



Observations above $50 are cut off from the figure.

Models and explanatory variables

Dependent variable is hourly earnings in all specifications, the models are all OLS.

1. Model 1: $age + sex$
2. Model 2: Model 1 $+age^2 + uhours + grade92$
3. Model 3: Model 2 $+sex \times grade92 + race + prcitshp + ownchild + state + lfsr94$
4. Model 4: Model 3 $+uhours^2 + marital + unionmme + class + chldpres$

The primary principles of my choice of predictors:

- Wage generally relates to the person's age, sex (if there is discrimination among sexes), education and number of hours worked (if the firm appreciates more work/full-time employees more)
- Additionally, the marginal effect of education can differ between men and women. Moreover, race, citizenship and geographic location (state), employment status of the worker and number of children can all matter to some extent when predicting income
- All other variables might be less relevant, or not especially important in the case of **cooks**

Sanity check of coefficients

Quick note on the significant coefficients:

- Positive: age, higher education level (baseline: lowest level), union membership (baseline: not a member), employed-at-work (baseline: employed-absent), some states i.e New York, DC (compared to baseline: Alabama)
- Negative: sex (baseline: male), race (white-asian, white-hawaiian, white/black/american-indian compared to baseline: white), many industries (compared to baseline: alcoholic beverages, merchant wholesalers)

Verdict: (significant) coefficients meet the general economic expectations, there is no counterintuitive relationship

Model comparison

Model	N coeff	RMSE full	RMSE cv	BIC full
Model1	3	6.1683	6.1755	13926
Model2	19	6.0627	6.0938	13958
Model3	103	6.0725	6.2014	14523
Model4	117	6.0548	6.2210	14603

Model complexity and performance

Complexity can be either measured by the number of coefficients estimated or the number of variables appearing in the regression. I use the former metrics, and compare the models by *RMSE in the full sample*, *cross-validated RMSE* (cv) and *BIC in the full sample*.

1. Full RMSE obviously decreases (non-increasing) as we add more variables, overall this is not a good measure of model performance.
2. Cross-validated (trained-tested-averaged) RMSE now tells more, as it is higher for Model 3 and 4 - during the training period (due to the many variables), the fitting procedure led to an overfit in the data, so the test data does not fit the model well. But Model 2 now outperforms Model 1 as it produces a better fit for the test data.
3. BIC, on the other hand, suggests that the simplest model (Model 1) has the lowest BIC value, but the difference between this and the BIC value of Model 2 is negligible.

All in all, it appears that Model 1 and Model 2 outperform the more complex ones (those including interactions, squared elements, factors with many categories) with their simplicity. However, income can be affected by many aspects and dimensions, so it is hard to believe that only `age` and `sex` would matter, this is why I choose **Model 2** - with the lowest cross-validated RMSE and with one of the lowest BIC values - as the best possible predictor of hourly income among these four models. Simple model, but also has some explanatory variables.

Nevertheless, many extra things could be done to improve the analysis (unite different factor levels that are statistically identical, search for other similar occupations - i.e head chef - to merge with **cooks** so there would be more data, dealing with some unique outlier: 120$/hour income at age 18), but those are out of the scope of this analysis.