Hourly wages prediction using common wage determinants

Anais Ouedraogo

April 25, 2023

1 Introduction

- Introduce determinants of hourly wages that will be used in the regression

  E.g: Wage differences due to discrimination has been an issue that always existed and is present in many aspects of our daily life. We often hear about wage discrimination based on gender but also on race and age.

- Give general statistics

  E.g: Racial gaps in hourly wage are a real issue that needs to be addressed. Whites in the US earn around 30% more per hour than Blacks, and this difference is associated with large racial differences in occupational assignments.[2]

- Give goal of the paper

  E.g: The goal of this paper is to predict hourly wages using the Lasso prediction model and the Rigde prediction model using different income determinants.

2 Literature review

- Literature review on the relationship between each determinant and hourly wage

  For race: A paper by Petre Melinda examined the role of cognitive and non-cognitive skills in racial wage gaps. She came up with two main explanations for the persistence of wage gaps. Firstly, minority workers with the same ability and training receive lower wages, and secondly, minority workers bring less skill and ability to the market.[4] An article by Marjorie Baldwin and John A Bishop looked over the variation of black-white wage ratios on the wage distribution in 1991. They found that racial gap increases with wages for men and that there is nearly no wage gap for black women at all wage levels. [1] Using data from 1983, CPS Hirsch and Schumacher investigated the effect of racial composition of the labor market on wage rate. The results were that the wage rates of whites and blacks were lower in industry-occupation-regional groups with a high density of black workers. This study is different from most others because it studies the effect of black density on wage which is very interesting. This says that a white worker in a black dense industry will be affected by the discriminant wage rates.[3]

  Education Level:

  Sex:

  Age:

3 Data

In order to investigate the question, I am using two different datasets (atusresp and atuscps) from the American Time Use Survey(ATUS) available in the atus R package. Both data frames contain information collected in the CPS about all individuals who responded to the ATUS between 2003 and 2016. Atuscps dataset gives information about education and demographics while atusresp contains information about wages and employment for the same household id numbers. In total, both datasets have 35 variables with 181,335 observations. For this paper, we are only going to use the following 9 variables. Refer to table 1.

- Explore data
- Do barplots for categorical variables
- histogram for age
- graph correlation between age and hourly wage

4 Methods

1. Linear regression
    - Classic linear regression
    - y = $\beta_0 + \beta_1 x + u$

$$\begin{aligned}\text{logwage} = 1.81 &+ 0.004\text{Black only} - 0.004\text{Asian only} - 0.003\text{Other} + 0.074\text{hs diploma} \\ &+ 0.088\text{some college} + 0.119\text{associate degree} + 0.105\text{bachelor's degree} + 0.054\text{master's degree} \\ &- 0.130\text{prof degree} - 0.072\text{doctoral degree} - 0.004\text{midwest} - 0.029\text{south} + 0.009\text{west} \\ &- 0.022\text{female} + 0.002\text{age} - 0.098\text{PT} + 0.007\text{professional} - 0.117\text{service} \\ &- 0.085\text{sales} - 0.000\text{office admin} - 0.109\text{farming forestry fishing} \\ &+ 0.039\text{construction} + 0.037\text{install repair maint} - 0.003\text{production} \\ &- 0.039\text{transport} + u \end{aligned} \tag{1}$$

2. Lasso Regression model
    - 10-fold cross validation
3. Ridge Regression model

- 10-fold cross validation

5 Findings

- Discuss the result of classical linear regression and the coefficients in table 2
- Discuss lasso in sample and out of sample rmse using figure 1
- Discuss ridge in sample and out of sample rmse using figure 2

6 Conclusion

Both lasso and ridge model gave the same best rmse with the lowest penalty. There is not much difference in the prediction of the two model.The models are not overfitting and not underfitting. Race, age, sex, education level,region, part time/full time, and the working industry can be considered good determinants of hourly wages.

References

[1] M. Baldwin and J. Bishop. An analysis of racial differences in wage distributions. *Economics Letters*, 37(1):91, 1991.

[2] L. Golan and C. Sanders. Racial gaps, occupational matching, and skill uncertainty. *Federal Reserve Bank of St.Louis Review*, 101(2):135–153, 2019.

[3] B. T. Hirsch and E. J. Schumacher. Labor earnings, discrimination, and the racial composition of jobs. *Journal of Human Resources*, 27(4):602–628, 1992.

[4] M. Petre. Contributions of skills to the racial wage gap. *Journal of Human Capital*, 13(3):479–518, 2019.

[Table 1 about here.]

[Table 2 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

List of Figures

	penalty	.metric	.estimator	mean	n	std_err	.config
1	1.000000e-10	rmse	standard	0.1902874	10	0.002226106	Preprocessor1_Model01
2	1.599859e-10	rmse	standard	0.1902874	10	0.002226106	Preprocessor1_Model02
3	2.559548e-10	rmse	standard	0.1902874	10	0.002226106	Preprocessor1_Model03
4	4.094915e-10	rmse	standard	0.1902874	10	0.002226106	Preprocessor1_Model04
5	6.551286e-10	rmse	standard	0.1902874	10	0.002226106	Preprocessor1_Model05

Figure 1: Top Lasso RMSE.

	penalty	.metric	.estimator	mean	n	std_err	.config
1	1.000000e-10	rmse	standard	0.1947787	10	0.002224134	Preprocessor1_Model01
2	1.599859e-10	rmse	standard	0.1947787	10	0.002224134	Preprocessor1_Model02
3	2.559548e-10	rmse	standard	0.1947787	10	0.002224134	Preprocessor1_Model03
4	4.094915e-10	rmse	standard	0.1947787	10	0.002224134	Preprocessor1_Model04
5	6.551286e-10	rmse	standard	0.1947787	10	0.002224134	Preprocessor1_Model05

Figure 2: Top Ridge RMSE.

List of Tables

Variable	Description
region	region of household
sex	respondent sex age
respondent age edu	respondent education level race
respondent race occup$_c ode$	occupational code ptft
whether the respondent works part-time or full-time hourly$_w age$	hourly earnings at main job in dollars

Table 1: Variable description.

	(1)
(Intercept)	1.816
	(0.005)
regionmidwest	-0.004
	(0.003)
regionsouth	-0.029
	(0.002)
regionwest	0.009
	(0.003)
sexfemale	-0.022
	(0.002)
eduhs diploma	0.074
	(0.003)
edusome college	0.088
	(0.003)
eduassociate degree	0.119
	(0.003)
edubachelor's degree	0.105
	(0.003)
edumaster's degree	0.054
	(0.005)
eduprof degree	-0.130
	(0.011)
edudoctoral degree	-0.072
	(0.014)
raceBlack only	0.004
	(0.002)
raceOther	-0.003
	(0.005)
raceAsian only	-0.004
	(0.005)
age	0.002
	(0.000)
occup_codeprofessional	0.007
	(0.004)
occup_codeservice	-0.117
	(0.004)
occup_codesales	-0.085
	(0.004)
occup_codeoffice_admin	0.000
	(0.004)
occup_codefarming_forestry_fishing	-0.109
	(0.010)
occup_codeconstruction	0.039
	(0.005)
occup_codeinstall_repair_maint	0.037
	(0.005)
occup_codeproduction	-0.003
	(0.004)
occup_codetransport	-0.039
	(0.005)
ptftPT	-0.098
	(0.002)
hourly_wage	0.045
	(0.000)
Num.Obs.	56217
R2	0.870
R2 Adj.	0.870
AIC	-26490.1
BIC	-26239.9
Log.Lik.	13273.075
F	14508.256