Report wage prediction in legal jobs

Péter ENDES-NAGY

Introduction

This is a report on *predicting hourly wages in legal jobs*. 4 models were built for predicting hourly (log) wages in the legal field, based on 2014's CPS data. The first 3 models are simple OLS regressions, gradually more complex, the predictors were chosen based on field knowledge and availability of variables in the dataset. The 4th Model was derived from the 3rd Model with LASSO.

Data

I used data from the 2014's Current Population Survey (CPS), that is a monthly survey of about 60,000 households in the US. A detailed description of the Survey and the methodology is available here: https://osf.io/uqe8z/

I was interested interested in legal occupations, so I narrowed down my sample to this field, using `2100` (Lawyers, Judges, magistrates, and other judicial workers), `2105` (Judicial law clerks), `2145` (Paralegals and legal assistant) and `2160` (Miscellaneous legal support workers) census codes.

Data cleaning and EDA

Without getting into too many technical details, the potential predictor variables were inspected on by one, some simplyfied and recoded into other variables.

There is a hierarchy among the 4 included occupations, some seems to have a rather supporting role, so a binary `occupation` variable was created. `race` was simplified into white and POC - POC usually face discrimination on the labor market. `marital` status was simplified into married and never married, `chldpres` (presence of a child) and having a child under 18yo or not. Man are usually rewarded and women are penalized for being married or having a child on the labor market, so these variables are going to be used in interaction with `gender`. `sector` was also simplified into a binary variable, differentiating between government and private sector as the hourly wages can be very different in them. Level of education was recoded, treating differently the graduated, those who have some or finished higher education degrees and those who have high school diploma or less - as most jobs in the legal field require MA or MBA, everyone under a BA degree could have been discarded, but it turns out they predominantly work in the supporting roles, so it makes sense keeping them in the sample.

Furthermore, those who marked "Employed - but absent" were filtered out (2% of the original sample). Imputation wasn't necessary as none of the chosen predictors had missing values.

By inspecting y, the hourly wage, a log transformation was decided as the hourly wage distribution is close to lognormal distribution - after the transformation, the distribution is still strongly skewed, the hourly wages seem to be capped. An extreme value of almost 500 USD/hour was discarded - the person worked only 1 hour.

The only continuous numeric predictor is age. Plotting against hourly wage, a spline at age 35 is recommended in the models.

At the end, I arrived to a sample with 1708 observations.

Model building

First of all, the sample was split into a working and hold-out sample. The model building was carried out on the working sample.

3 Models were built by adding more and more variables. A 4th Model was built with LASSO, based on Model 3.
Model 1: `sex` , `education`
Model 2: Model1 , `occ` (supporting roles) , `age` (spline at age 35)
Model 3: Model2 , `race` , `sector` (govt/private) , `chldpres` (child under 18yo or not) , `marital` , `stfips` (State) , `chldpres*sex` (interaction: having child under 18yo and gender) , `marital*sex` (interaction: marital status and gender)

Model 1-3 were run with 5-fold cross-validation.

Regarding Model 4, LASSO ended up with a lambda value of 0.05, the minimum that I set during the tuning. LASSO was also run with other tuning parameters (lambda starting from 0.01, increased by 0.01). It preferred a lambda as small as possible and not surprisingly, the variables kept are very sensitive to the parameters. In any case, I use the original lambda parameter starting from 0.05.
LASSO narrowed down the predictor pool to 7 variables (plus intercept) that are actually very similar to Model 2: sex, education1, education2, occ, lspline(age, 35)1, stfipsDC, age.

Diagnostics and comparing model results

The Models were run both on the hold-out set both on the full sample. The following table also includes the original model results on the working set.

Model	Coef	R_squared	BIC	Training_RMSE	Test_RMSE	Hold_RMSE	Full_RMSE
M1	4	0.289	1776	0.458	0.459	0.485	0.464
M2	7	0.339	1699	0.442	0.444	0.467	0.447
M3	63	0.411	1945	0.417	0.437	0.485	0.464
M4	8	0.334	NA	NA	0.450	0.472	0.454

R2 isn't much relevant in prediction, but it is worth mentioning that as expected, more variables we included in the model, the variables explained more of hourly wage's deviation: from 28.9% it increased to 41.09%.
BIC is lowest for Model 2, the very high number of coefficients in Model 3 inflated the BIC value.
As of RMSE values, Model 3 performed best on the test sample (5-fold cross-validation), but Model 2 was the best on the hold-out sample and on the full sample - better RMSE on the full sample as the sample size is 5-times larger. LASSO shrunk the 63 coefficients of Model 3 to 8 coefficient. These coefficients are quite similar to Model 2, the RMSE values are also very close. On the hold-out and full sample, Model 1 and Model 3 performed equally poorly.

Conclusion

In this report, I built 4 models to predict hourly wages in legal fields, using the 2014 CPS dataset. The first 3 model gradually became more complex, the 3rd model included 63 variables. The 4th Model was derived from Model 3 with LASSO.
Both based on BIC both based on RMSE (hold-out and full sample), Model 2 performed the best. LASSO was almost as good as Model 2 and choose quite similar values as I did for Model 2 using domain knowledge.