Assignment 1

Course: ECBS6067 - Prediction with Machine Learning for Economists

Anne Valder

2023-11-10

Task1: In order to construct 4 different models to predict earnings per hour, I first filter for the occupation '2310' which includes elementary and middle school teachers. Predictions for this occupation might help us when we have to decide what salary to put in a job advertisement for teachers, for example. Or if we want to investigate differences due to gender, education level, experience etc. The first step is to conduct the following data set preparation steps:

Sample design: In order to match the business or policy question, I start by constructing the target variable earnings per hour from 'weekly earnings' and 'usual work hours. Next, I drop extreme values where earnings per hour is below 1 or larger than 300. Since I am interested in predicting "standard" teaching earnings per hour, earnings smaller than 1 would be considered errors. Keeping an extreme value, like '583' that is likely to be an error, would have a high-cost due to the quadratic errors in the loss function. In addition, I select an age range between 18 and 65, i.e. the working population age and an education level (= highest grade attended) that includes only high school graduates or higher. The reasons for this is that for teachers under 18 teaching is likely not their main occupation. Moreover, including only teachers with a certain degree assures that we do not include for example tutoring from high school students in the sample.

Label engineering: In this step I transformed the target variable into logs since it might be interesting to analyse relative changes rather than changes per monetary unit. Moreover, relative price differences are often more stable and the distribution of log prices is often close to normal, which makes linear regressions give better approximation to average differences (Békés and Kézdi, 2021). As we can see from the histograms in the appendix, transforming the target variable to logs does not alter the shape of the distribution significantly. Therefore, I continue to use the target variable in levels.

Feature engineering: Next, I will define the list and functional form of variables that I later use as predictors. First, I look at missing values. Here I find that the two variables 'ethnic' and 'unioncov' contain missing values. Rather than imputing the data, I make use of related variables like 'race' and 'unionmme'. Second, after some visual inspections (see exemplary boxplots for 'sex' or 'prcitshp' in the appendix) I transform all relevant ordered categorical values or text variables to binary variables. I end up with 19 "new" binary variables: female, native, private, white, union, edu_HighS, edu_BA, edu_MA, edu_Prof, edu_PhD, married, divorced, wirowed, nevermar, child0, child1, child2, child3, child4plus. Third, I look at the functional form of the predictors. Since I suspect a nonlinear relation between earnings per hour and age. For example, the effect of age could be positive up until, the age of 50, and then negative thereafter. Therefore, I construct the variables 'agesq'. After inspecting the graphic relation, I also added 'lnage'. As indicated by the graphs, the variable 'lnage' captures the relationship the best. For all other continuous predictors, I assume linear relationships. Last, I consider interactions of variables. One is the interaction between 'female' and 'age' (and agesq, lnage). Age being here a proxy for years of experience. The interaction follows the belief, that the potential work experience for males and females is different, that is, females experience more career interruptions than males. Other interactions could have been occupation*state since industry and labour market structures that impact earnings per hour differ between regions. However, I am only considering one occupation here. To circumvent this another option is to include state fixed effects. In the end the sample consists of 3623 observations and 49 variables (some redundant for the following analysis like household ID). Table 3 in the appendix presents the descriptive statistics of all relevant variables.

Models: As specified above, the target variable is earnings per hour. AI suggests including the following predictor variables when predicting teacher's earning per hour: education, year of experience, additional certifications, geographic location, school type, class size, employment status, gender and many more (ChatGPT, 2023). Unfortunately, not all of these variables are available in the data. I end up with the following four models:

- **Model 1:** edu_BA + edu_MA + edu_Prof + edu_PhD
- **Model 2:** edu_BA + edu_MA + edu_Prof + edu_PhD + age + lnage
- **Model 3:** edu_BA + edu_MA + edu_Prof + edu_PhD + age + lnage + female + stfips
- **Model 4:** edu_BA + edu_MA + edu_Prof + edu_PhD + age + lnage + female + stfips + union+ native + private + white + married + divorced + wirowed + nevermar + child0+ child1 + child2 + child3 + child4plus + female_age + female_lnage

In model 1 (the simplest) I try to explain earnings per hour by the "most obvious" factor education-, i.e. skill-level. I assume that the higher the education, the higher should be the earnings. In the second model, I add 'age' and 'lnage' as a proxy for experience. I assume that the older the teacher, the more experience she/ he has in teaching. Since the visualizations showed a nonlinear relation between age and earnings per hour, I also add the log of age. It seems that the gain of age (i.e. experience) is especially high in the beginning and then diminishes quickly. In the third model I include the variable 'female' since past research indicates disparities in earnings due to gender. Moreover, I add the geographic location, since the labour market structures that impact earnings per hour might differ between regions. Furthermore. in model 4, I allow for more demographic variables I deem important for prediction: union membership, the origin of the teacher, their employment class and their race and some social variables concerning family status. On top, I add interaction terms as described above.

Task 2: Model Evaluation In order to analyse the performance of the four different models, I calculate the R-squared, the RMSE and the BIC in the full sample. In addition, I calculate the cross-validated RMSE. In Table 1 we can see that as the models become more complex (i.e. include more predictors) the R-squared increases, the RMSE of the full sample decreases until model 3 and then stays constant and the BIC first decreases and then increases again for model 3 and 4. This in line with the theory that the BIC penalizes adding additional predictors more heavily (Békés and Kézdi, 2021). Looking at table 2, we can see the cross validated RMSE, which is again the lowest for model 3 and 4.

Table 1: Earnings per hour - BIC, in-sampe RMSE

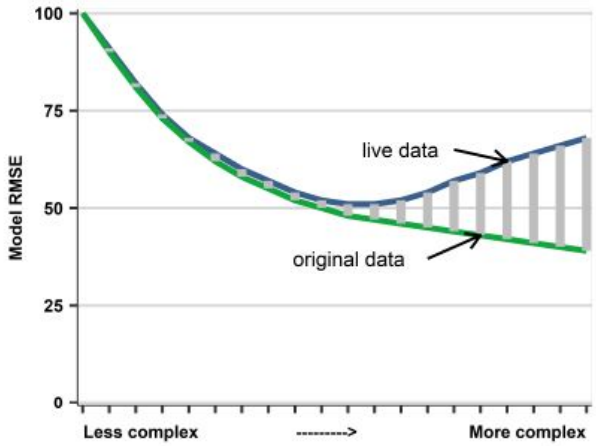
Models	N predictors	R-squared	RMSE (full)	BIC
model_1	4	0.06	12.89	28855.08
model_2	6	0.09	12.67	28746.69
model_3	57	0.16	12.20	28893.19
model_4	70	0.16	12.15	28968.00

Table 2: Earnings per hour - CV RMSE

Resample	Model1	Model2	Model3	Model4
Fold1	13.54	12.94	12.20	11.92
Fold2	12.65	12.31	13.26	12.45
Fold3	13.35	13.45	11.63	12.56
Fold4	12.05	11.98	12.46	12.82
Average	12.91	12.68	12.40	12.44

Task 3: Relationship between model complexity and performance: As can be seen from the results of the model evaluation above when the model complexity increases (due to additional features, interactions,

polynomials etc.) the performance of the model at first increases but once a certain complexity is reached the model performance stagnates or decreases. The reason metrics like the BIC or the RMSE (test and cross validated) begin to raise again is overfitting. Overfitting means to fit a model to the data set that is too complex. It can happen for example that adding more and more features does not increase the predictive power because the "new" features are strongly correlated with predictor variables already in the model. Metrics like BIC and AIC penalize this compared to metrics like R-squared, which increases with the number of predictors. An option to avoid overfitting in the training data is to evaluate it on test data (cross validation) (Békés and Kézdi, 2021). The graph below visualizes the relationship between model complexity and performance. At first, the model RMSE decreases for both the live and original data when the complexity increases. After a certain threshold (i.e. overfitting) the RMSE of the live and the original data start to diverge with the RMSE of the original data falling monotonously and the RMSE for target observations rising.



Source: Békés and Kézdi, (2021), p.379

References

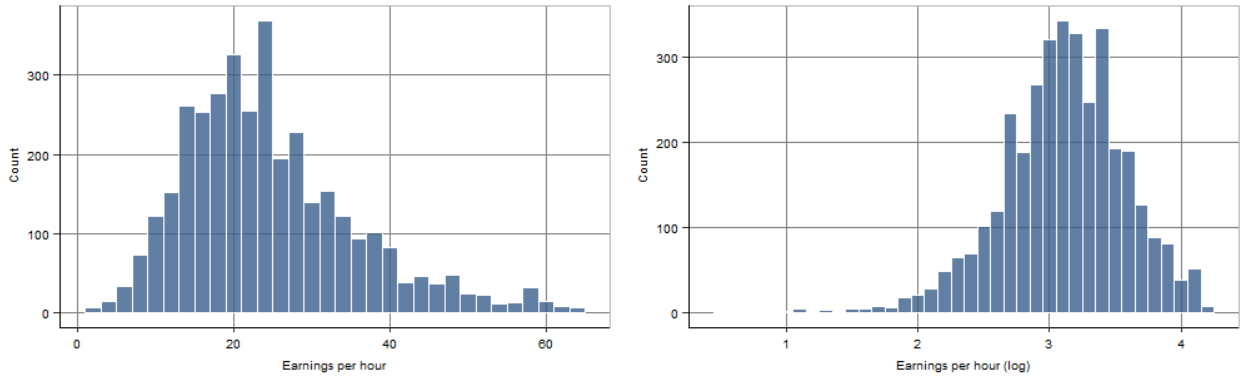
ChatGPT, Teacher Earnings Predictors, November 9, 2023. Retrieved from: https://chat.openai.com/share/44e6cfb2-6a2a-4c1d-a824-1618ae741cd7

Békés,G., Kézdi,G.(2021).R, Python and Stata code for Data Analysis for Business, Economics, and Policy, ch13-used-cars-reg, GitHub repository, https://github.com/gabors-data-analysis/da_case_studies/tree/master/ch14-used-cars-log

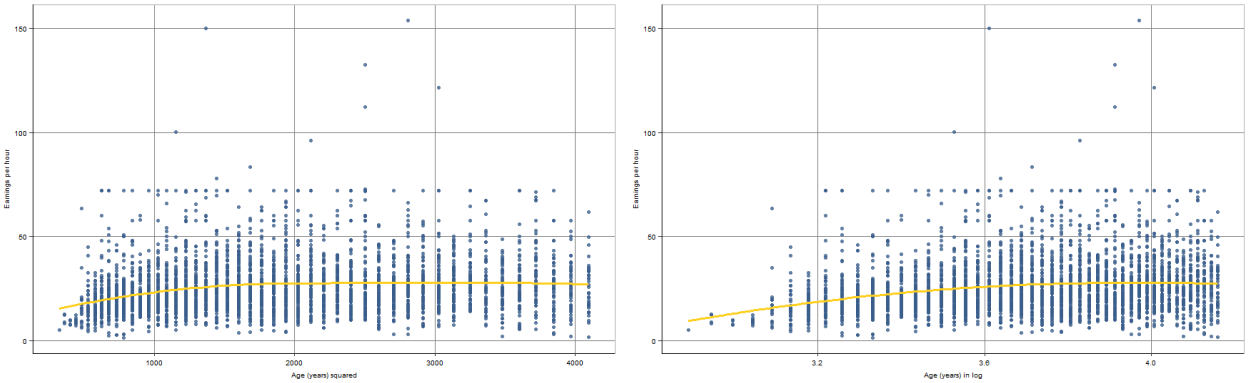
Békés, G., & Kézdi, G. (2021). Data analysis for business, economics, and policy. Cambridge University Press, chapter 13.

Appendix

Histogram of earnings per hour (left) and log earnings per hour (right)



Relation between agesq (left) and log of age (right) each with earnings per hour



Box plots for gender and citizenship status

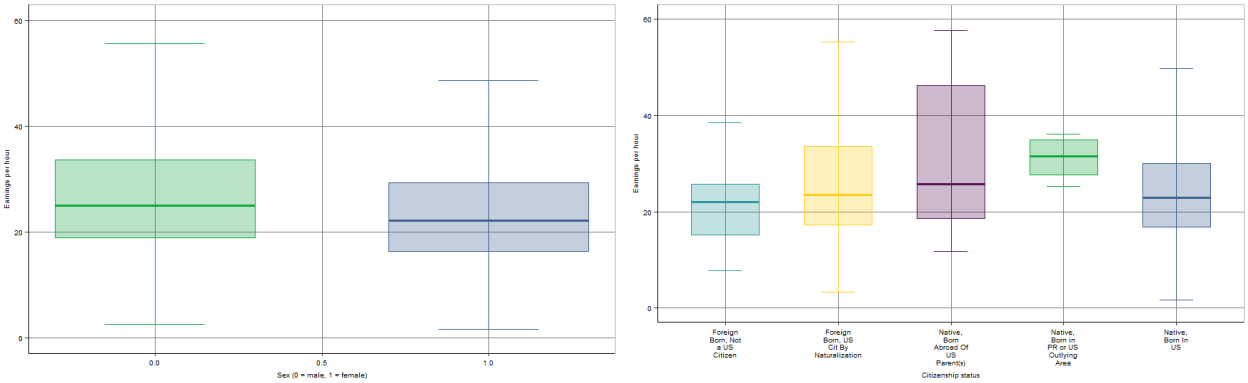


Table 3: Table 1A: Data summary

	Mean	Median	Min	Max	P25	P75	N
earnh	25.67	23.08	1.60	153.67	16.90	31.25	3623
learnh	3.13	3.14	0.47	5.03	2.83	3.44	3623
age	42.19	42.00	18.00	64.00	33.00	51.00	3623
agesq	1904.02	1764.00	324.00	4096.00	1089.00	2601.00	3623
lnage	3.71	3.74	2.89	4.16	3.50	3.93	3623
female	0.82	1.00	0.00	1.00	1.00	1.00	3623
female_age	34.61	38.00	0.00	64.00	26.00	49.00	3623
female_agesq	1563.67	1444.00	0.00	4096.00	676.00	2401.00	3623
female_lnage	3.04	3.64	0.00	4.16	3.26	3.89	3623
native	0.96	1.00	0.00	1.00	1.00	1.00	3623
private	0.20	0.00	0.00	1.00	0.00	0.00	3623
white	0.88	1.00	0.00	1.00	1.00	1.00	3623
union	0.56	1.00	0.00	1.00	0.00	1.00	3623
married	0.71	1.00	0.00	1.00	0.00	1.00	3623
divorced	0.10	0.00	0.00	1.00	0.00	0.00	3623
wirowed	0.01	0.00	0.00	1.00	0.00	0.00	3623
nevermar	0.18	0.00	0.00	1.00	0.00	0.00	3623
child0	0.52	1.00	0.00	1.00	0.00	1.00	3623
child1	0.05	0.00	0.00	1.00	0.00	0.00	3623
child2	0.03	0.00	0.00	1.00	0.00	0.00	3623
child3	0.13	0.00	0.00	1.00	0.00	0.00	3623
child4plus	0.26	0.00	0.00	1.00	0.00	1.00	3623
edu_HighS	0.00	0.00	0.00	1.00	0.00	0.00	3623
edu_BA	0.43	0.00	0.00	1.00	0.00	1.00	3623
edu_MA	0.47	0.00	0.00	1.00	0.00	1.00	3623
edu_Prof	0.01	0.00	0.00	1.00	0.00	0.00	3623
edu_PhD	0.01	0.00	0.00	1.00	0.00	0.00	3623