Report - Assignment 1

Timur Naushirvanov

03/10/2021

Introduction

In this report, I am presenting the results of my analysis of hourly wages among Education, Training, and Library Occupations. The **dependent variable** is a natural logarithm of a consistent hourly wage series during entire period (weekly wages divided by weekly working hours), the **independent variables** are age, female, union membership, number of children, particular occupations, and their different modifications and interactions. Each of these variables has theoretical and empirical grounds for believing that they influence our dependent variable.[1]

Data Preparation and Description

I chose to log the dependent variable in order to reduce the effect of influentional observations (instead of deleting them). Some of the observations were deleted because of the lacking information about either working hours or weekly wages, and imputing this data would not give precise observations for this analysis. Female and union membership are binary variables, each particular occupation (occ2012) - factor, age and number of children - discrete. I also added age squared and number of children squared because visualisations showed light quadratic relations between those variables and ln(wage).

Building Models

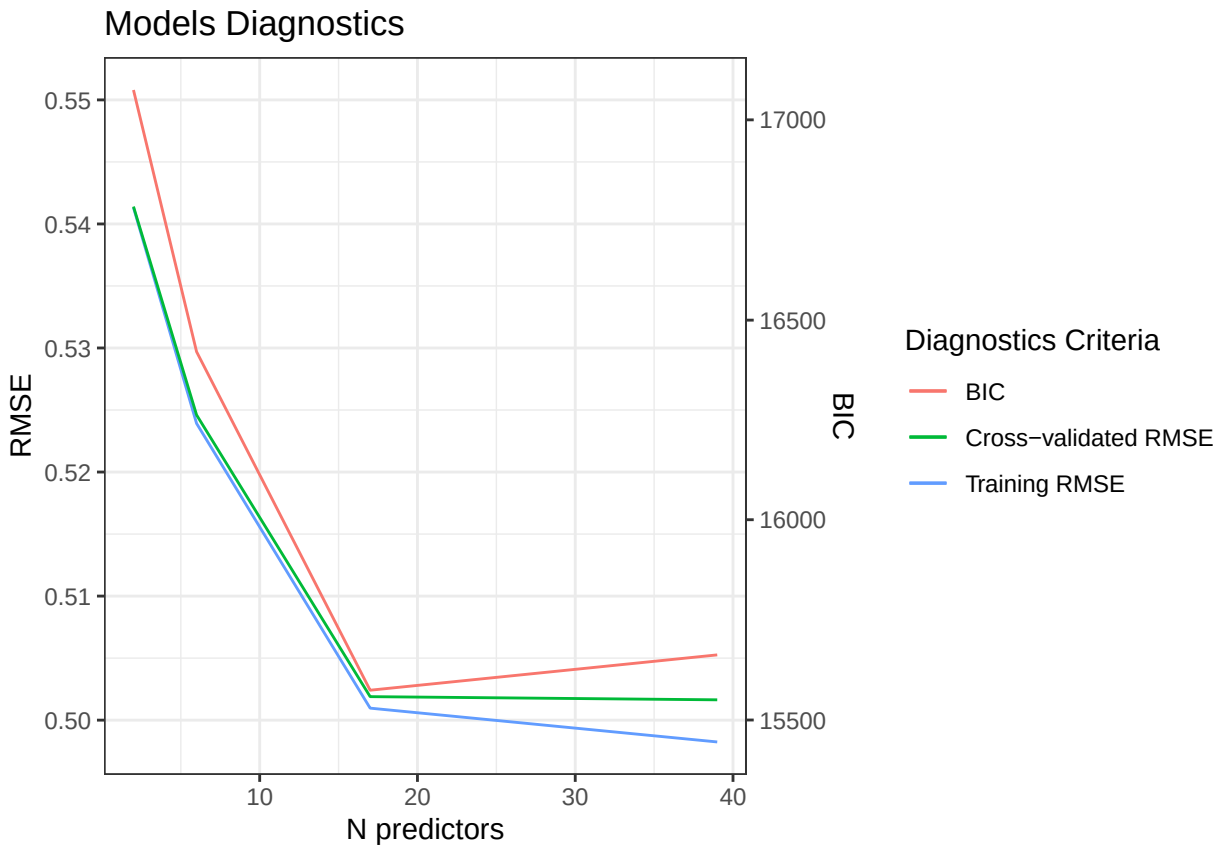
- Model 1: lnwage ~ age + agesq - one of the preliminary visualisations allowed to notice squared relationship between lnwage and age
- Model 2: lnwage ~ age + agesq + female + union_mem + ownchild + ownchildsq - adding variables with theoretical and empirical evidence of having impact on wages
- Model 3: lnwage ~ age + agesq + female + union_mem + ownchild + ownchildsq + occ2012 + female*union_mem - adding a fixed effect on different types of occupations within this cohort and an interaction term (do female union members have significantly different salaries from male non-union members?)
- Model 4: lnwage ~ age + agesq + female + union_mem + ownchild + ownchildsq + occ2012 + female∗union_mem + occ2012∗female + occ2012∗union_mem + age∗female + age∗union_mem - adding some extra interaction terms

Analysing Models Performance

As it is expected, increasing number of predictors (higher complexity) reduces BIC and RMSE; however, when the number of predictors is too big, the model overfits data, BIC increases, and RMSE continue to decrease much slower (in this particular example). Based on this information, I would choose model 3, because it has the lowest BIC (which tries to prevent overfitting).

	N predictors	R-squared	Training RMSE	BIC	Cross-validated RMSE
Model1	2	0.0765369	0.5413136	17074.72	0.5413998
Model2	6	0.1349282	0.5239203	16420.72	0.5246029
Model3	17	0.2090628	0.5009682	15574.75	0.5018935
Model4	39	0.2176525	0.4982405	15663.09	0.5016371

[1] Additional visualisations and tables, including descriptive statistics and the regression table, can be found in the Appendix document.


Models Diagnostics
RMSE
0.55
0.54
0.53
0.52
0.51
0.50
BIC
17000
16500
16000
15500
10
20
30
40
N predictors
Diagnostics Criteria
BIC
Cross−validated RMSE
Training RMSE