DA3 - Assignment 1

Viktória Kónya

26 January 2022

Introduction

The goal of this report is to predict the hourly wages of Financial managers using socio-demographical and job related information about the professionals. The estimated models are going to be evaluated with the RMSE and the BIC measures in the full sample as well as with the cross-validated RMSE in order to choose the model with the best predictive properties.

Sample selection

The source of the analysis is the CPS dataset and we are going to focus on financial managers with undergraduate or graduate degrees (BA or higher). Managers between the age of 21 and 64 and with the minimum of 20 working hours per week were considered only in order to focus on the active labor force with a degree. Observations with earnings below one dollar per hour were excluded as such low wage is unlikely among Financial managers. The final sample consists of 881 observations with 416 women and 446 men.

Predictors of hourly wages

First, let's consider the potential predictors of the hourly wages. **Fig.3.** in **Appendix A3.** shows the relationship of the outcome with each predictor separately. From the socio-demographic variables we are going to consider the gender, the age, the level of education, whether the manager has children, the marital status and the race. Age seems to be a strong predictor with 26.5% correlation with the outcome. The lowess smoother suggests concave pattern of association hence we can we can approximate the relationship with a quadratic function. The wage disadvantage of female managers is about 10 dollars per hour and we can expect earnings to increase with an MA degree but the returns of additional education are not clear. Also, married managers and managers with kids in the households are expected to earn somewhat higher. From the job related variables the the type of the job and union membership were considered, however as financial managers are mainly employed by private companies the predictive power of these variables are not expected to be strong.

We can also assume that the pattern of association between the age and the wages and the pattern of association between the education level and the wages are different among male and female financial managers. **Fig.4.** in **Appendix A4.** shows the relationship with the hourly wages by subgroups. With the same level of education, female financial managers can expect lower salaries in all education levels. However, in case of the professional degree and the PhD degree the confidence intervals show high uncertainty suggesting that we cannot reject the hypotheses that the average salaries are the same in the two subgroups. Regarding the age, the quadratic relationship seems fine among male managers, while among female managers the relationship seems more likely to be cubic. We can incorporate these patterns in our analysis by adding interaction terms to out prediction model.

Predictive models and performance comparison

We can start our model setup by estimating the simplest model and then enrich our prediction model by adding more explanatory variables. **Table 3** in **Appendix A5.** summarizes the results of our regressions with the measures of fit. In the first model only age with a quadratic form was added as the only predictor of the hourly wages. The second model was extended with more demographic information about the managers and in the third model job related information was added as well. Finally, the last specification adds the interactions suggested by **Fig.4.** in **Appendix A4.**.

If we look at the estimated coefficients, we can see that as we anticipated, female financial managers are expected to earn less then their male coworkers, wages expected to grow with the age at a diminishing rate and having an MA compared to having only a BA degree also contributes to the higher salaries. Regarding the marital status, widowed and never married managers are expected to earn somewhat less than their married coworkers, and managers with kids in their households can expect higher salaries compared to managers without kids. As for the job related variables, managers with union membership have lower salaries than managers without union membership.

Let's take a look at the measures of fit of the four models to find the best specification for our prediction problem. **Table 1** summarizes the RMSE and BIC calculated from the full sample. The RMSE continuously decreases as we

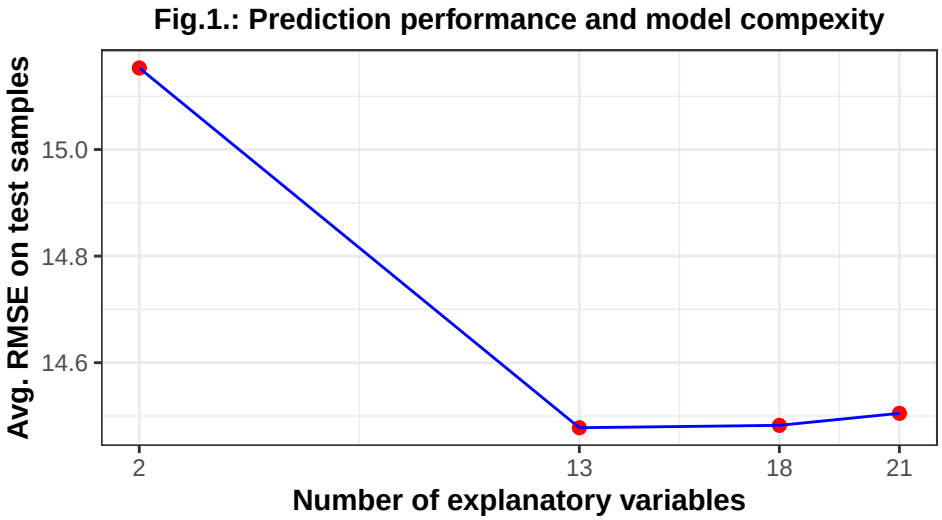
Table 1: Model evaluation based on full sample RMSE and BIC

Model	N coeff	R-squared	RMSE	BIC
Model 1	2	0.09323	15.117	7,305.8
Model 2	13	0.19342	14.257	7,277.2
Model 3	18	0.19895	14.208	7,305.1
Model 4	21	0.20489	14.156	7,318.9

add more predictors to the model, suggesting that the most complex model with the interaction terms capture the best the variation of the hourly wages. In contrast, if we rely on the BIC measure, we would chose a simpler model with only the socio-demographic information (Model 2).

Performance and model complexity

Let's also cross validate our results using 4-fold cross-validation. **Fig.1.** shows the average test RMSE of the four folds for each model specification, starting with the simplest model with only 2 predictors.



This approach also favors the second model with only the socio-demographic information over the last model with more complex patterns. The conclusion based on the cross-validated RMSE is in line with those of the BIC which also suggested that the second model had the best properties. If we take a look at the exact RMSE values in **Appendix A6.** we can see that the difference in the average RMSE between Model 2, 3 and 4 is so small that even if a latter two outperformed the simpler model, then would still prefer to keep the second model instead of adding 5 and 8 additional variables to the specification.

Appendix

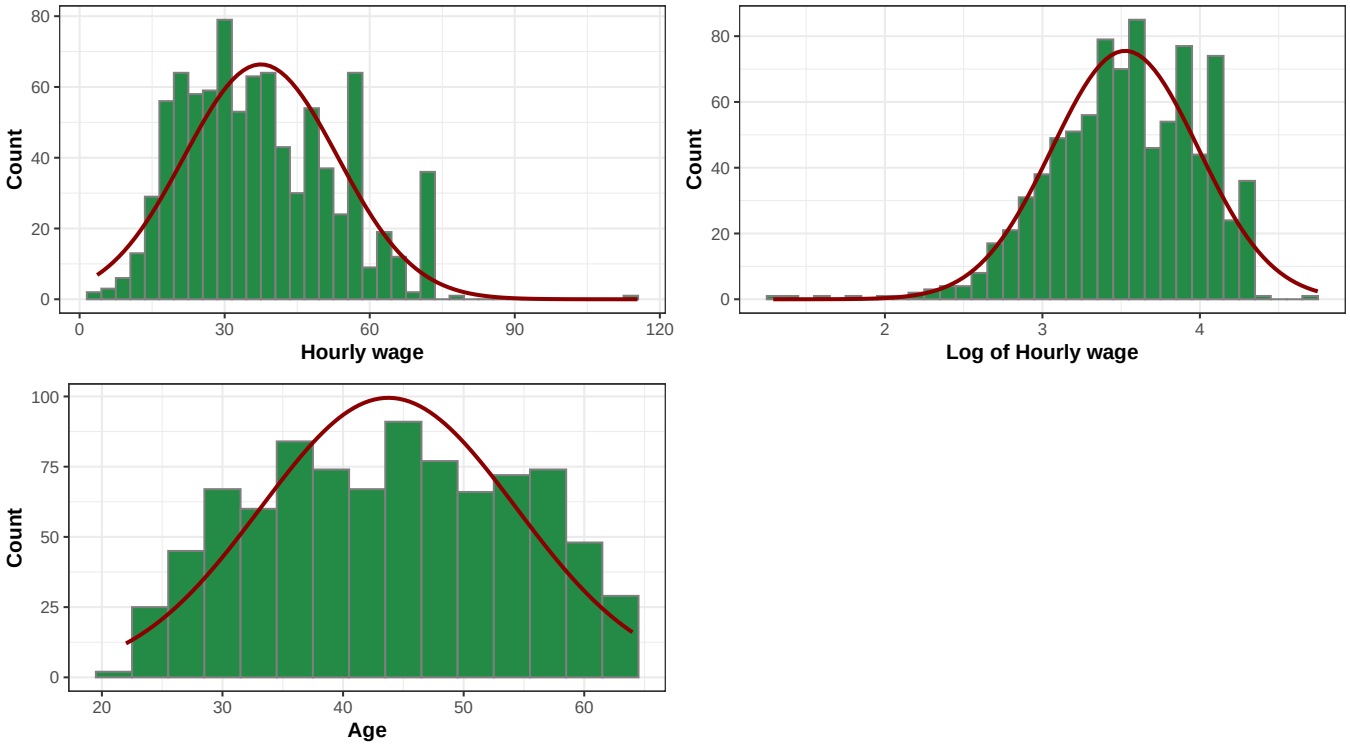
A1. Descriptive statistics

Table 2: Descriptive statistics

	N	Missing	Mean	SD	Min	Max	Range	P05	Median	P95
Hourly wage	881	0.00	37.48	15.88	3.64	115.38	111.75	15.68	34.96	67.30
Log of Hourly wage	881	0.00	3.53	0.47	1.29	4.75	3.46	2.75	3.55	4.21
Age	881	0.00	43.76	10.60	22.00	64.00	42.00	27.00	44.00	61.00
Female	881	0.00	0.47	0.50	0.00	1.00	1.00	0.00	0.00	1.00
BA Degree	881	0.00	0.70	0.46	0.00	1.00	1.00	0.00	1.00	1.00
MA Degree	881	0.00	0.29	0.45	0.00	1.00	1.00	0.00	0.00	1.00
Professional	881	0.00	0.01	0.09	0.00	1.00	1.00	0.00	0.00	0.00
PhD	881	0.00	0.01	0.09	0.00	1.00	1.00	0.00	0.00	0.00
Marital - Married	881	0.00	0.68	0.46	0.00	1.00	1.00	0.00	1.00	1.00
Marital - Divorced	881	0.00	0.10	0.30	0.00	1.00	1.00	0.00	0.00	1.00
Marital - Widowed	881	0.00	0.01	0.09	0.00	1.00	1.00	0.00	0.00	0.00
Marital - Never maried	881	0.00	0.20	0.40	0.00	1.00	1.00	0.00	0.00	1.00
Has child	881	0.00	0.44	0.50	0.00	1.00	1.00	0.00	0.00	1.00
Race - White	881	0.00	0.87	0.34	0.00	1.00	1.00	0.00	1.00	1.00
Race - African-American	881	0.00	0.05	0.22	0.00	1.00	1.00	0.00	0.00	1.00
Race - Asian	881	0.00	0.07	0.25	0.00	1.00	1.00	0.00	0.00	1.00
Race - Other non-white	881	0.00	0.01	0.09	0.00	1.00	1.00	0.00	0.00	0.00
Union member	881	0.00	0.03	0.16	0.00	1.00	1.00	0.00	0.00	0.00
Job type - Private - For-profit	881	0.00	0.84	0.36	0.00	1.00	1.00	0.00	1.00	1.00
Job type - Government - Federal	881	0.00	0.03	0.16	0.00	1.00	1.00	0.00	0.00	0.00
Job type - Government - State	881	0.00	0.03	0.18	0.00	1.00	1.00	0.00	0.00	0.00
Job type - Government - Local	881	0.00	0.04	0.20	0.00	1.00	1.00	0.00	0.00	0.00
Job type - Private, Nonprofit	881	0.00	0.06	0.23	0.00	1.00	1.00	0.00	0.00	1.00

A2. One variable analysis

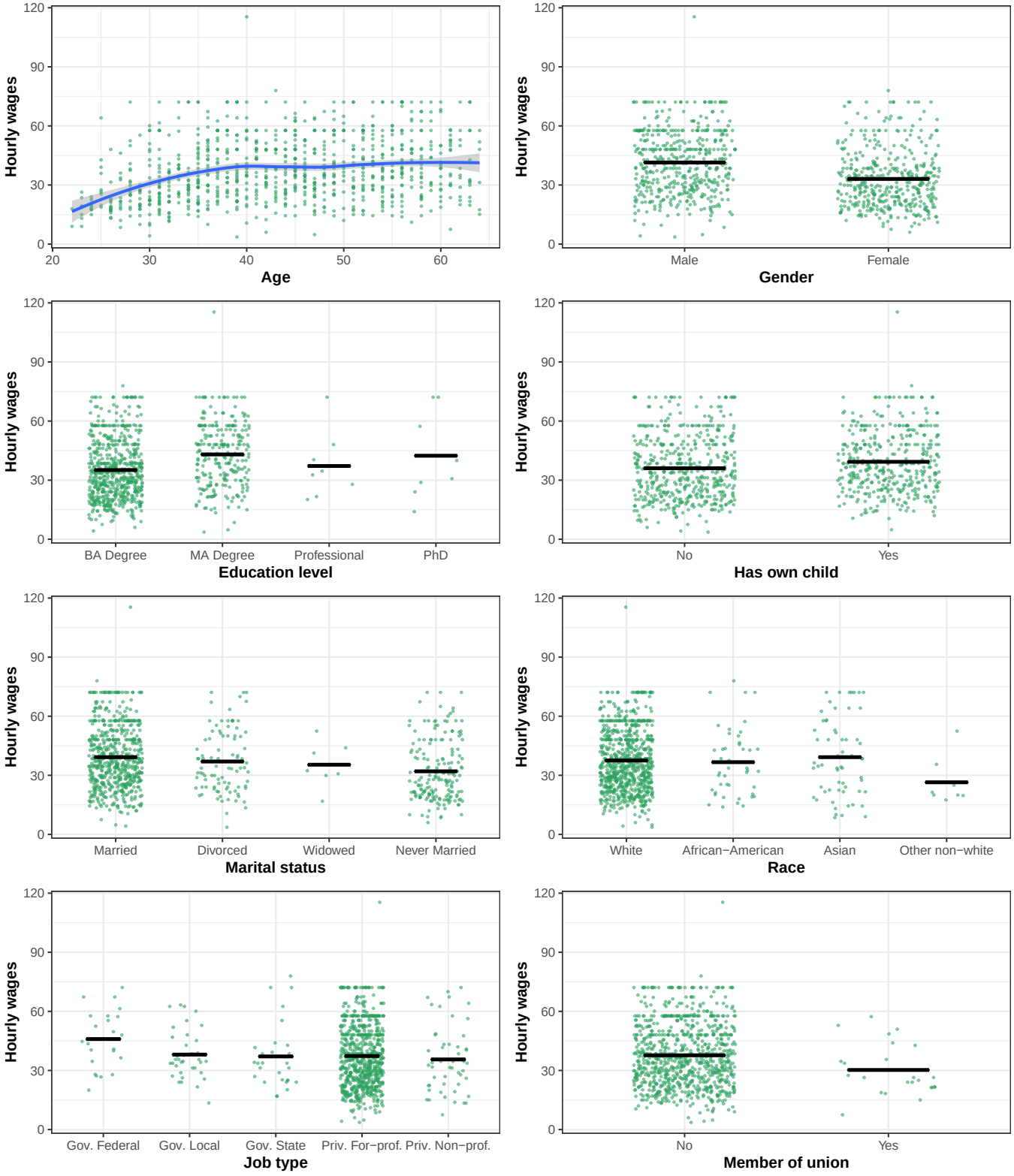
Fig2.: Histograms



Note: Normal curves were added to the distributions.

A3. Relationship with outcome variable

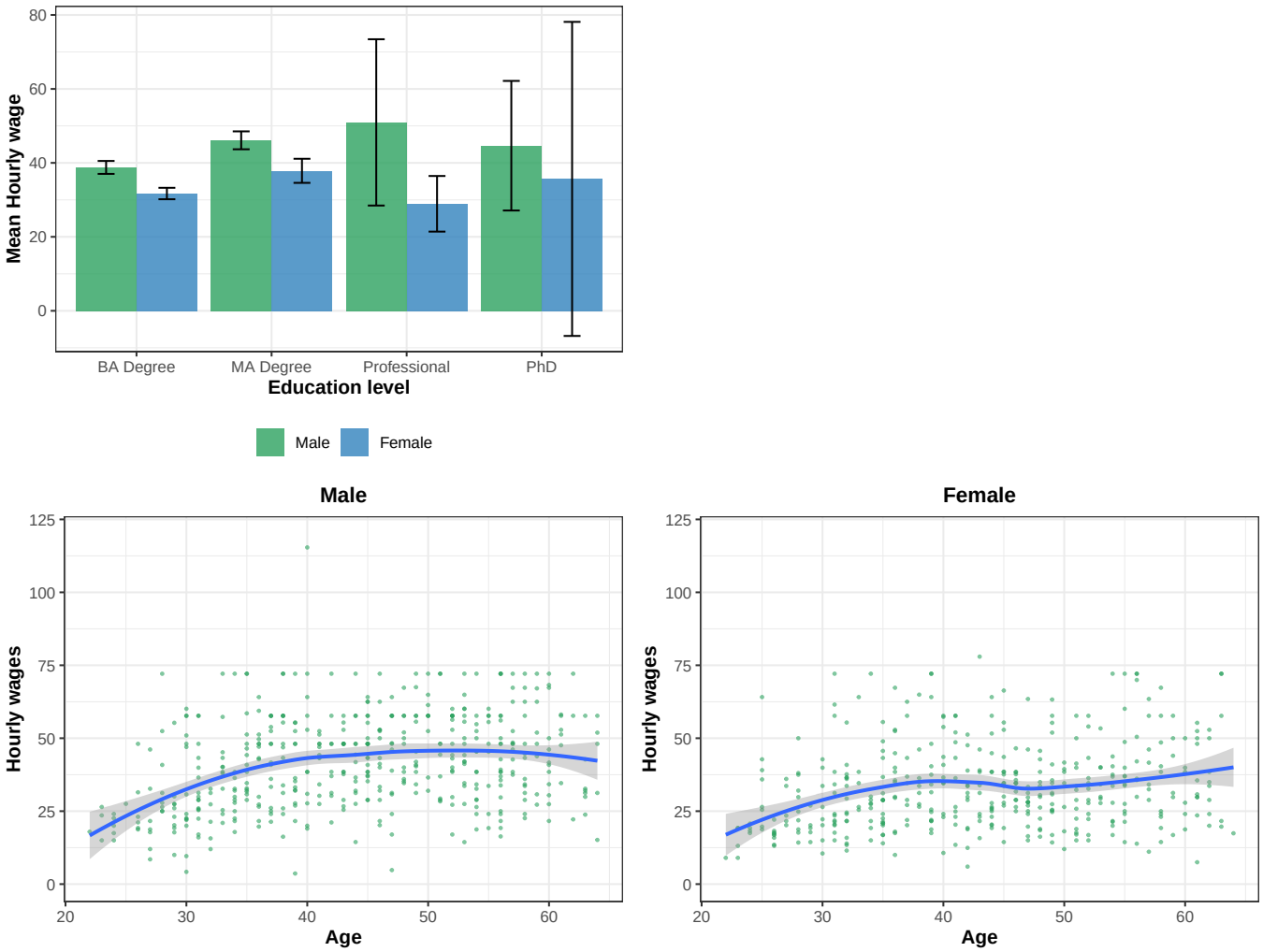
Fig3.: Relationship with outcome



Note: Horizontal lines show the average hourly wages within the category.

A4. Interactions

Fig4.: Interactions



Note: Confidence intervals were added to the averages.

A5. Regression

Table 3: Regression models for predicting horly wages

Dependent Variable:	Hourly wage			
Model:	(1)	(2)	(3)	(4)
Variables				
Intercept	-18.63**	-5.078	2.267	-18.14
	(7.587)	(8.003)	(9.129)	(11.85)
Age	2.283***	1.683***	1.582***	2.493***
	(0.3725)	(0.3975)	(0.4053)	(0.5424)
Age sqare	-0.0216***	-0.0149***	-0.0137***	-0.0233***
	(0.0044)	(0.0047)	(0.0047)	(0.0063)
Female		-7.158***	-6.912***	30.56**
		(1.004)	(1.022)	(15.03)
Education - MA		5.879***	5.693***	5.370***
		(1.131)	(1.142)	(1.468)
Education - Professional		2.421	1.604	1.662
		(4.582)	(4.557)	(4.458)
Education - PhD		3.379	2.626	2.476
		(7.934)	(8.200)	(8.367)
Marital - Divorced		0.5026	0.5342	0.6429
		(1.694)	(1.703)	(1.705)
Marital - Widowed		-1.709	-1.138	-0.8489
		(3.499)	(3.461)	(3.474)
Marital - Never married		-0.9027	-0.8697	-0.7833
		(1.443)	(1.445)	(1.437)
Race - African-American		0.2899	0.3523	0.5366
		(2.590)	(2.548)	(2.567)
Race - Asian		2.625	2.836	2.881
		(2.254)	(2.268)	(2.257)
Race - Other non-White		-5.795	-5.744	-6.017
		(4.570)	(4.922)	(4.956)
Has child		1.880	2.065*	1.895
		(1.199)	(1.200)	(1.206)
Union member			-3.558	-3.651
			(2.662)	(2.728)
Job - Government-Local			-5.189	-4.969
			(3.531)	(3.513)
Job - Government-State			-5.158	-4.800
			(4.154)	(4.147)
Job - Private,ForProfit			-5.497*	-5.407*
			(3.038)	(3.012)
Job - Private,Nonprofit			-7.911**	-7.533**
			(3.678)	(3.651)
Female × Education - MA				0.4845
				(2.344)
Age × Female				-1.670**
				(0.7322)
Age sqare × Female				0.0175**
				(0.0085)
Fit statistics				
AIC	7,291.4	7,210.3	7,214.2	7,213.7
BIC	7,305.8	7,277.2	7,305.1	7,318.9
RMSE	15.117	14.257	14.208	14.156
Observations	881	881	881	881
No. Variables	2	13	18	21

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

A6. Cross-validated test RMSE

Table 4: Model evaluation based on cross-validated test RMSE

Resample	Model1	Model2	Model3	Model4
Fold1	14.89635	14.47096	14.57288	14.34116
Fold2	13.64327	12.88888	12.89629	15.16412
Fold3	16.45031	15.74269	15.65437	14.14576
Fold4	15.48608	14.66537	14.67006	14.34743
Average	15.15315	14.47789	14.48235	14.50492