Zsófia Rebeka Katona
Data Analysis 3 - Assignment 1

Predictive models of the earnings dataset

Introduction

The aim of this project is to carry out predictive analysis based on different evaluation criterias (BIC, RMSE and cross-evaluated RMSE) and compare the different regression models carried out on the CPS earnings dataset.

My chosen occupation was Human resource managers (Code: 0136) and Human resource workers (Code: 0630). To obtain the earning per hours, I divided the weekly earnings by the "uhours". Before carrying out the predictive analysis, I examined the distribution of the earnings per hour and log earnings per hour. (Figure 1.) The earnings per hour distribution was relatively normal, slightly skewed to the left, with some extreme values on the right tail. The log distribution was normal, slightly skewed to the right.

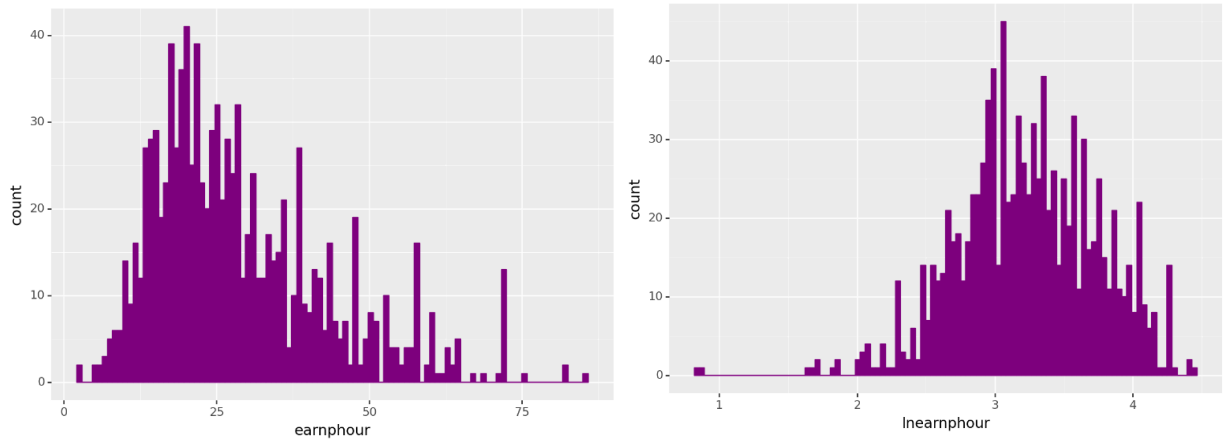

Figure 1: Distribution and log distribution of earnings per hour (wages) of HR professionals and workers

Modeling

In Model 1, I regressed earnings per hour on the age variable, as a larger value in age usually comes with more professional and educational experience, creating more disparities in earnings. As Figure 2 shows, the linear regression indicates a growing tendency. The lowess method shows a similar growing tendency in wage until the age of 40 which shifts into stagnation and a slight decrease after that.

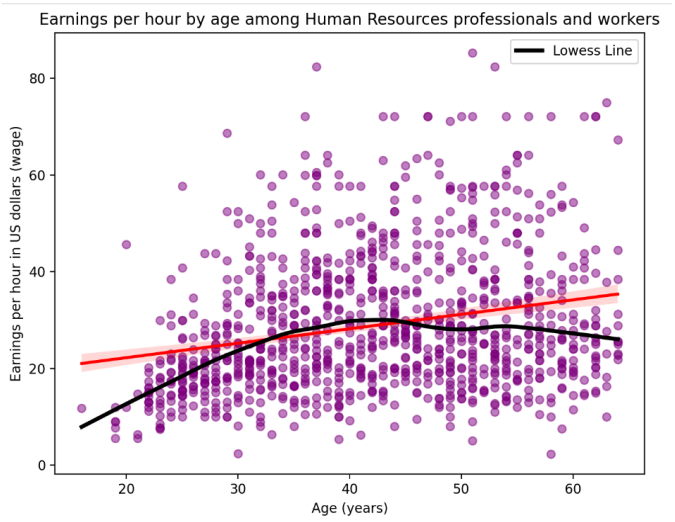

Figure 2: Model 1 (Earnings per hour by age among HR professionals and workers) regression line and Lowess line

Zsófia Rebeka Katona

As part of feature engineering, I expanded Model 2 by adding gender as the second variable as it shows a notable influence on earnings differences. I created a dummy variable and a new column for “female”. I added age squared to Model 3, allowing the predictive models to fit data more accurately in addition to capturing potential non-linearities. In the last part of feature engineering, the predictors in Model 4 were level of education and having children as both significantly impact earnings. I decided that it would yield a more pronounced contrast in earnings if I included the high school and higher education levels separately, the higher education including Bachelors to Professional degree and PhD. Moreover, having children might entail career interruptions, different workplace policies and further discrimination, creating further wage disparities and a more relevant prediction. I created dummy variables for level of education and having children, denoting the variables as “hsedu”, “higheredu” and “children”. Despite having information on the number of kids, I grouped them as one dummy variable to capture the general impact of having children.

Diagnostics

As for sanity check, I examined the coefficients in Figure 3: All covariates are significant at 99%. Individuals with a higher value in age, both level of education and individuals with children are expected to earn more. Females are expected to earn less. Even with the highest adjusted R-squared value, Model 4 only explains 13.7% of the variation in wages. The multicollinearity between age and age squared affects coefficient interpretability, though its significance diminishes in prediction where coefficients play a less crucial role.

	Dependent variable: earnphour			
	(1)	(2)	(3)	(4)
age	0.299***	2.124***	2.172***	1.988***
	(0.037)	(0.256)	(0.256)	(0.269)
agesq		-0.022***	-0.022***	-0.020***
		(0.003)	(0.003)	(0.003)
female[T.True]			-4.920***	-4.724***
			(1.022)	(1.013)
hsedu				7.850***
				(2.851)
higheredu				13.478***
				(2.691)
children				1.620
				(1.024)
Constant	16.251***	-19.419***	-16.657***	-26.941***
	(1.508)	(4.854)	(4.946)	(6.082)
Observations	1017	1017	1017	1017
R^2	0.056	0.094	0.116	0.142
Adjusted R^2	0.055	0.092	0.113	0.137
Residual Std. Error	13.969 (df=1015)	13.696 (df=1014)	13.534 (df=1013)	13.352 (df=1010)
F Statistic	65.407*** (df=1; 1015)	85.175*** (df=2; 1014)	66.617*** (df=3; 1013)	35.299*** (df=6; 1010)
BIC	8261.37	8227.03	8208.76	8199.06
Note:				*p<0.1; **p<0.05; ***p<0.01

Figure 3: Stargazer regression table of all 4 models

The BIC values of the four models have a range of 62 (8261 to 8199), Model 4 obtaining the lowest score (8199), suggesting the best fit and penalty. Similarly, the RMSE of the models suggest that Model 4 would produce the lowest RMSE for the target observations, possibly providing the best prediction. The RMSE values range from 13.95 to 13.30 with slight differences in between the models. To carry out the cross-validated RMSE, I split the data into 4 categories and calculated the OLS for each fold. Averaging the results of each fold for each model, it can be concluded that Model 4 still provides the best fit. The range of the CV RMSE results were from 13 to 14.

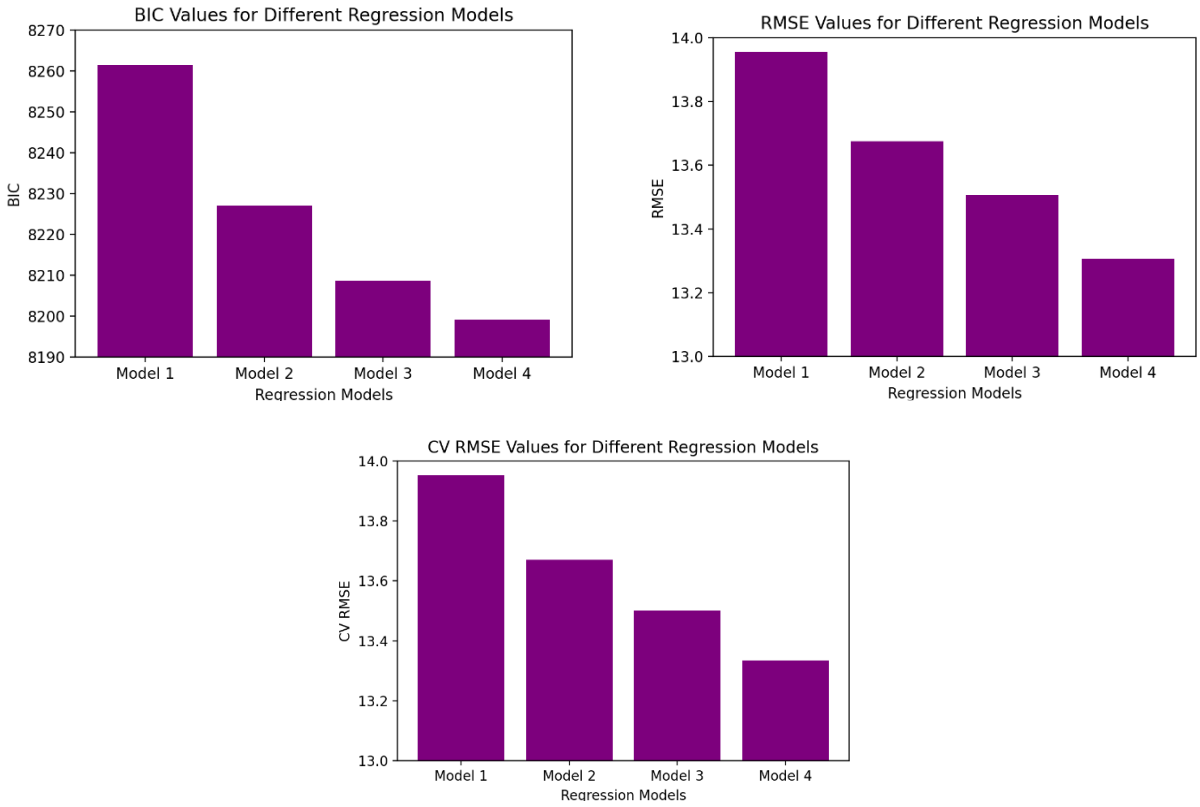

Figure 4-6: BIC, RMSE and cross-validated RMSE scores

Prediction

Before carrying out the prediction, I transformed the values in my “female” variable from boolean to numeric values as the get_prediction function couldn’t comprehend the boolean values in my “new” dataframe. In making my prediction, I selected the following values for the predictor variables: a 30-year-old female with higher education, no children, and a $28 hourly wage. I have examined the residual values and basic metrics of all models. In general, the mean is not close to 0, indicating that the models make biased predictions. The standard deviations are similar to the CV RMSE range (13-14) referring to a greater variability. This also means that the variability in the residuals is consistent with the overall variability observed during cross-validation. The range of residuals are around 80 for all of the models, denoting a wider spread of errors and potential model limitations.

Based on the 80% confidence interval, it can be concluded that even with the best model selection, Model 4’s prediction interval is extremely wide. As it can be seen in Figure 7, this suggests that our target variable's predicted earnings per hour (wage) falls between 14 and 49, which is still large. I have also calculated the predicted earnings with a 95% confidence interval which has shown similar results.

	Model1	Model2	Model3	Model4
Predicted	25.218	24.819	23.577	31.577
PI_low(80%)	7.301	7.256	6.220	14.067
PI_high(80%)	43.134	42.382	40.935	49.088

Figure 7: The prediction interval of the models with an 80% CI

Conclusion

According to all the evaluation criterias (BIC, RMSE and CV RMSE), Model 4 consistently outperforms other models in terms of fit and predictive accuracy. The impact of age, gender, level of education and having children is significant on earnings. On the other hand, there is a possibility that Model 4 would have become overfitted if more variables were introduced as the scores of evaluation criterias gradually decreased, suggesting a lower performance in external validity. The prediction confidence intervals of all the models are extremely wide, suggesting a potential uncertainty in predicting wages even with the best model selection. In terms of overall model validity, despite the wide prediction CI, all evaluation criterias reassured selecting Model 4 as it provides the most reliable prediction on the wage of HR professionals and workers based on the chosen predictors.