

Predictive Modeling on Hourly Wages

Ian Brandenburg (2304791)

GitHub Repo: https://github.com/Iandrewburg/Assignment_1

This analysis looks at the CPS-earnings dataset with the purpose of generating predictive modeling of hourly wages. The dataset was filtered to specifically look at Medical Practitioners, and thus, the occupation codes from 3000 through 3540 were analyzed. Furthermore, the data was cleaned by replacing missing values with “Missing”, but this did not seem to impact the data at all.

The variables that were selected for building the predictive model include: age, gender, children present, marital status, working status, and level of education. Gender and marital status were selected on the premise of possible wage discrimination based on gender, with the justification that if an employer learns you are married, they may want to pay you less because your time will be allocated outside of work to your family. Similarly, how in the United States, pregnant women face discrimination from employers, the number of children an individual has could be associated with wage differences. Furthermore, working status and level of education have a more direct association with the daily work of the individual, and could be associated with wages stronger. Finally, age tends to have an association with wages, because the older an individual gets, the more experience they accumulate.

Dummy variables were created for each of the variables, isolating the different levels of each variable. Furthermore, wages underwent a log transformation to display the relative differences. Lowess plots were created to visualize each variable with respect to log wages. These can be found in the index.

Four regression models were created, each incrementally becoming more complex. Each variable category has a dummy variable dropped. The following were not included in the regression models for the purpose of creating baseline categories: never married, 0 children, masters degree, and male. The final model includes interaction variables between age and education level, as age and education level visually seemed to have a more significant alteration association in hourly wages in the lowess plot.

Of the four models, model 4 has the highest R-Squared (16.1%), lowest BIC (15595.33), and lowest RSME (0.56); however, this does not mean it is the best model. This is because the difference between model 4 and model 3 is very miniscule. As Model 3 has much larger R-Squared than models 1 and 2, Model 3 would be the best model to go with as it has less variables than model 4, and is therefore less complex and easier to interpret.

The K-Fold Cross Validation of the RSME was performed to analyze the performance of the data on a more granulated level for the purpose of machine learning applicability. The results show that Model 3 and 4 have the lowest values. They are very close in value, so Model 3 is the best model here since it has the lower number of variables included in its model.

Prediction models were created for Model 1 and Model 3 for the purpose of comparison. The first prediction table represents the 95% prediction interval (PI) and shows that the predicted log hourly wage value from Model 1 is 3.229 and 3.454 for Model 3. This represents the expected log hourly wages according to each model, and Model 3 shows a higher value since

there are more variables included in the model. For Model 1, the low PI suggests that there is a 95% confidence that the log hourly wage will not fall below 2.055, and for Model 3, it will not fall below 2.361. The high PI for Model 1 suggests that the log hourly wage will not exceed 4.403, and for Model 3 will not exceed 4.548. Model 3 continues to give the higher values in the prediction modeling.

In the 80% prediction interval table, Model 3 performs similarly as it did in the 95% prediction interval table. Model 3 has consistently larger values than Model 1. Furthermore, the range between the high and low PI of Model 3 is smaller than Model 1, suggesting that Model 3 more accurately predicts log wages.

Appendix

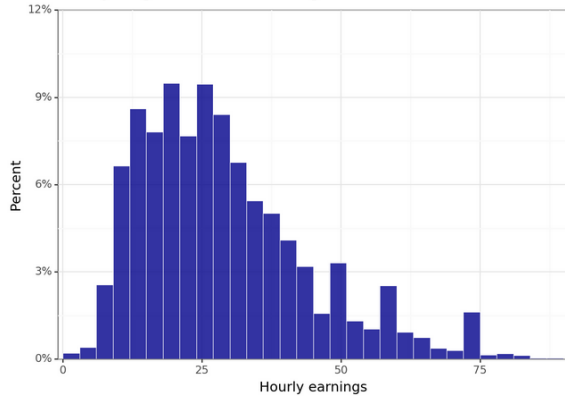
Regression table for log hourly wages

	<i>Dependent variable: log hourly wages</i>			
	(1)	(2)	(3)	(4)
Constant	2.788***	1.932***	1.779***	1.743***
	(0.023)	(0.103)	(0.098)	(0.102)
Age	0.011***	0.056***	0.044***	0.045***
	(0.001)	(0.005)	(0.004)	(0.004)
Age ²		-0.001***	-0.000***	-0.000***
		(0.000)	(0.000)	(0.000)
Age × University				-0.001
				(0.001)
Age × Prof Degree				0.003
				(0.003)
Age × PhD				0.003
				(0.002)
Female		-0.137***	-0.068***	-0.066***
		(0.016)	(0.016)	(0.016)
Working		-0.002	-0.006	-0.006
		(0.044)	(0.043)	(0.043)
Married		0.081***	0.045***	0.044***
		(0.013)	(0.012)	(0.012)
Divorced		-0.000***	-0.000***	0.000***
		(0.000)	(0.000)	(0.000)
Widowed		-0.050	-0.035	-0.034
		(0.050)	(0.048)	(0.048)
1 Child		0.164***	0.116***	0.116***
		(0.026)	(0.025)	(0.025)
2 Children		0.017	0.013	0.012
		(0.031)	(0.028)	(0.028)
3 Children		0.029	0.033*	0.032
		(0.020)	(0.020)	(0.020)
4+ Children		0.035**	0.039**	0.038**
		(0.017)	(0.016)	(0.016)
University			0.389***	0.445***
			(0.014)	(0.048)
Prof Degree			0.233***	0.120
			(0.031)	(0.112)
PhD			0.274***	0.158
			(0.028)	(0.100)
Observations	9208	9208	9208	9208
R ²	0.040	0.076	0.160	0.161
Adjusted R ²	0.040	0.075	0.159	0.159

Residual Std. Error	0.599 (df=9206)	0.588 (df=9197)	0.561 (df=9194)	0.561 (df=9191)
F Statistic	362.512*** (df=1; 9206)	87.954*** (df=10; 9197)	152.317*** (df=13; 9194)	130.950*** (df=16; 9191)
RSME	0.599	0.588	0.56	0.56
BIC	16713.16	16449.69	15595.33	15619.09
Note:	*p<0.1; **p<0.05; ***p<0.01			

Distribution Histograms

Hourly Wage Distribution Histogram

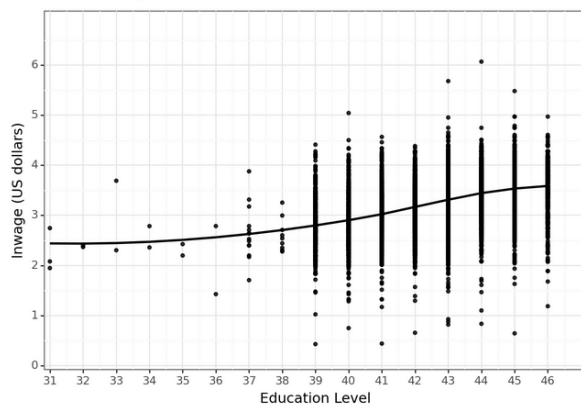


Distribution Histogram of Log Hourly Wages

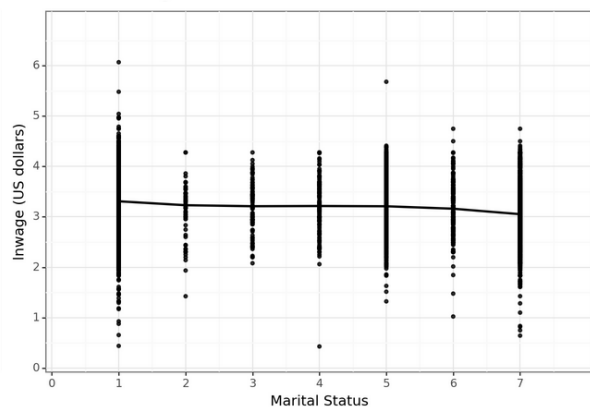


Lowess Plots

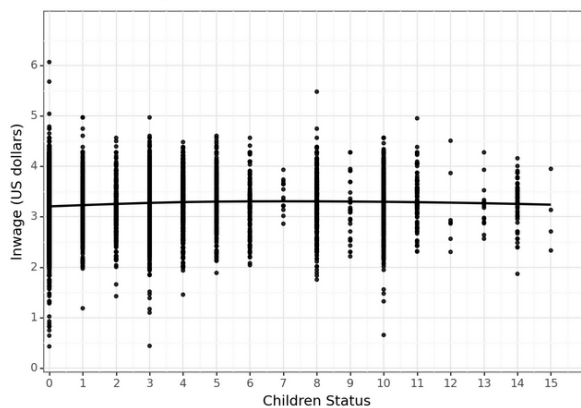
Lowess Wages Across Levels of Education



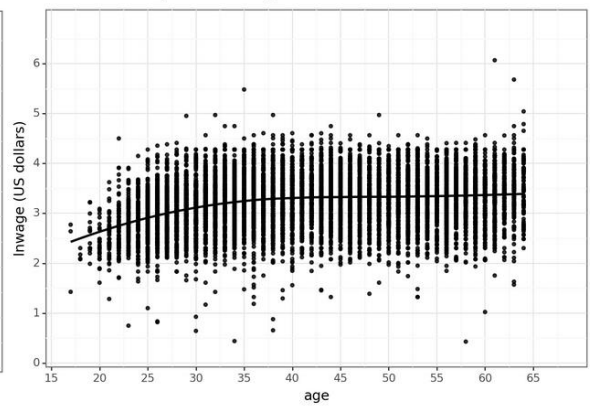
Lowess of Wages Across Marital Status



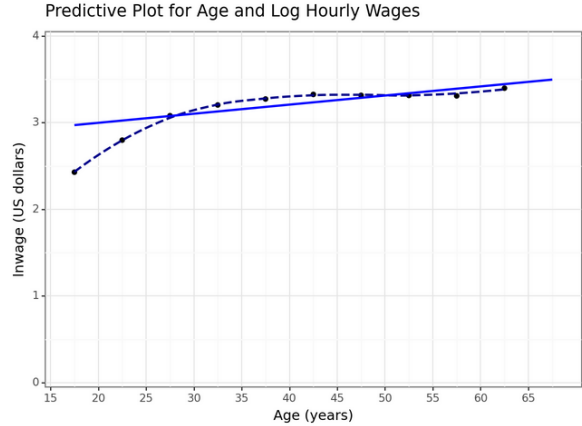
Lowess of Wages Across Children Present



Lowess of Wages Across Age



Predictive Plot



Prediction Interval Tables

	Model1	Model3		Model1	Model3
Predicted	3.229	3.454	Predicted	3.229	3.454
PI_low(95%)	2.055	2.361	PI_low(80%)	2.461	2.740
PI_high(95%)	4.403	4.548	PI_high(80%)	3.997	4.169