Data Analysis 3: Assignment 1 - Analytical Report

Github Repo: https://github.com/MarcellM01/Data-Analysis-3/tree/main/Assignment_1

This analysis delves into the CPS-Earnings dataset, focusing on generating predictive models for hourly wages within the banking sector. The dataset was specifically filtered to examine business operations specialists, considering occupation codes ranging from 0500 through 0740. The data was meticulously cleaned, replacing missing values where necessary, ensuring the integrity and reliability of the subsequent analysis.

Data Preparation and Exploratory Analysis

For building the predictive models, the selected variables include age, gender, number of children, marital status, employment status, and level of education. The choice of these variables was guided by several hypotheses: Gender, included to assess potential wage disparities that could arise due to gender discrimination. Marital Status, considered under the premise that marital commitments might influence an employer's perception of an employee's time allocation, potentially affecting wages. Number of Children, analysed to understand if parental responsibilities might correlate with wage differentials, reflecting a possible bias similar to the challenges faced by pregnant women in the workforce. Employment Status and Level of Education, directly relevant to an individual's role and responsibilities in the banking sector, these variables are hypothesized to have a strong correlation with earnings. Age, A standard variable in wage analysis, included to capture the experience-related aspects of wage determination. In addition to these, other factors like union membership, type of employer (government or private sector), and citizenship status were also considered for their potential impact on wages.

Dummy variables were created for each categorical variable to facilitate the analysis. The wage variable ('earnwke') underwent a logarithmic transformation to better represent relative differences and to normalize the distribution for statistical analysis. Lowess plots were generated for each variable against log-transformed wages, offering a visual insight into their respective correlations. These graphical representations play a crucial role in understanding the nuanced relationships between the selected variables and hourly earnings in the banking sector.

Model Development

Each model progressively incorporates more variables to capture the complexity of wage determination. Each categorical has a dummy variable dropped during model creation. While Model 1 provides a baseline understanding of gender-based wage disparities, Models 2 to 4 add layers of detail, considering education, union membership, age, parental status, employment type, and citizenship. The increasing complexity is aimed at achieving a more nuanced and accurate prediction of hourly wages.

Model Performance Comparison

In the comparative analysis of the four regression models, Model 3 emerges as the most balanced in terms of complexity and performance. It boasts the lowest Bayesian Information Criterion (BIC) at 5482.38, suggesting it has the best trade-off between model fit and complexity. Additionally, Model 3's Root Mean Square Error (RMSE) is impressively low, both for

the full sample (0.4831) and in the cross-validated context (average RMSE of 0.4829), underscoring its predictive accuracy. Its R-Squared value of 15.5% further illustrates a strong explanatory power. While Model 4 slightly surpasses Model 3 in predictive accuracy with a marginally lower RMSE (0.4794 for full sample, 0.4791 average cross-validated) and a higher R-Squared of 16.8%, it falls behind with a slightly higher BIC of 5488.74, indicating greater complexity. Models 1 and 2, with BICs of 5985.23 and 5777.35, RMSEs of 0.5184 and 0.5027 (full sample), and R-Squared values of 2.7% and 8.5% respectively, lag behind in both predictive power and complexity efficiency. This analysis clearly positions Model 3 as the preferable choice for balancing model simplicity with robust predictive capability.

Relationship Between Complexity and Performance

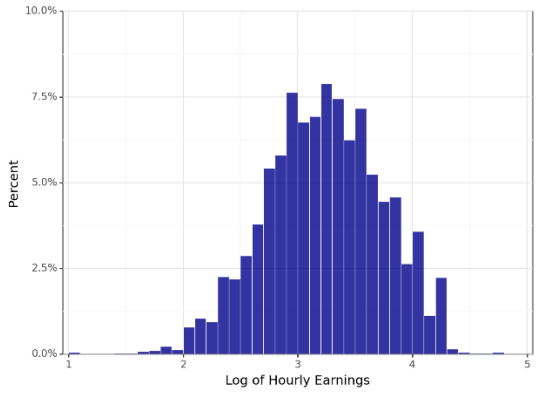
A discernible trend emerges when scrutinizing the relationship between model complexity and performance. Generally, an increase in complexity (incorporating more predictors) enhances the model's explanatory power and predictive accuracy, as evidenced by the improvement in RMSE and adjusted R-squared values. However, this enhancement is not linear, as indicated by the BIC values, where Model 3 (albeit very slightly) outperforms the more complex Model 4. This phenomenon underscores the diminishing returns of overcomplicating models, highlighting the need for a balanced approach in model construction.

Appendix

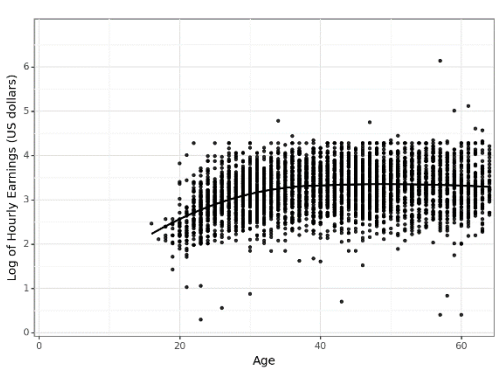
Regression Table

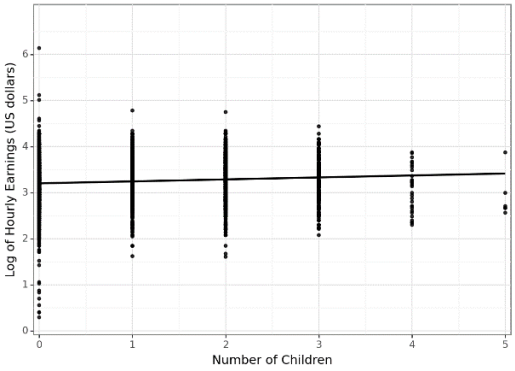
	Dependent variable: ln_earnings_per_hour			
	(1)	(2)	(3)	(4)
female	-0.174***	-0.160***	-0.155***	-0.149***
	(0.017)	(0.016)	(0.016)	(0.016)
union_member		0.043	-0.000	0.035
		(0.031)	(0.030)	(0.032)
MA_degree		0.305***	0.294***	0.297***
		(0.021)	(0.020)	(0.020)
Prof_degree		0.350***	0.323***	0.334***
		(0.061)	(0.060)	(0.059)
PhD_degree		0.522***	0.455***	0.432***
		(0.066)	(0.066)	(0.066)
age			0.011***	0.011***
			(0.001)	(0.001)
ownchild			0.045***	0.044***
			(0.008)	(0.008)
Private_For_Profit				0.072***
				(0.027)
Government_Federal				0.194***
				(0.037)
Government_State				-0.111***
				(0.042)
Government_Local				0.021
				(0.044)
Native_Born_In_US				0.032
				(0.143)
Foreign_Born_Not_US_Citizen				-0.039
				(0.153)
Foreign_Born_US_Citizen_By_Naturalization				0.069
				(0.147)
Native_Born_Abroad_Of_US_Parents				-0.028
				(0.155)
Intercept	3.330***	3.261***	2.764***	2.671***
	(0.013)	(0.014)	(0.032)	(0.149)
Observations	3917	3917	3917	3917
R^2	0.027	0.085	0.155	0.168
Adjusted R^2	0.026	0.084	0.153	0.164
Residual Std. Error	0.519 (df=3915)	0.503 (df=3911)	0.484 (df=3909)	0.480 (df=3901)
F Statistic	105.405*** (df=1; 3915)	82.348*** (df=5; 3911)	103.665*** (df=7; 3909)	55.843*** (df=15; 3901)
Note:				*p<0.1; **p<0.05; ***p<0.01

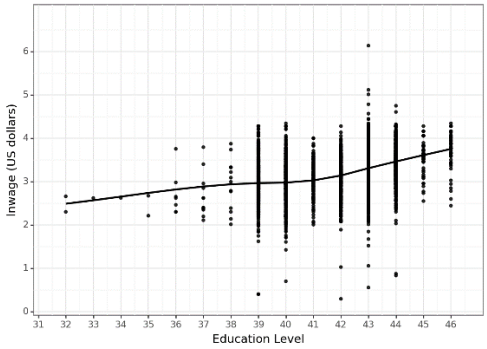
Histogram



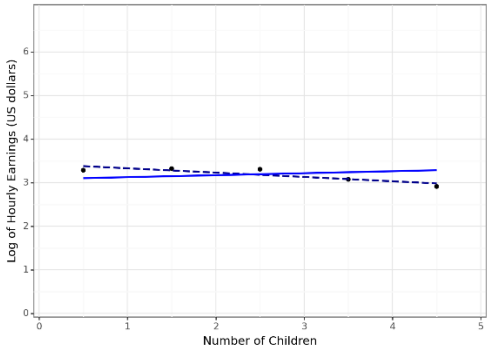
Lowess Plot







Predictive Plot



Prediction Tables

	Model1	Model3
Predicted	3.156150	3.425596
PI_low(95%)	2.139669	2.475455
PI_high(95%)	4.172630	4.375738

	Model1	Model3
Predicted	3.156150	3.425596
PI_low(80%)	2.491509	2.804332
PI_high(80%)	3.820790	4.046860