GitHub: https://github.com/szilvasipeter2000/Data-Analysis-3/tree/main

DA3 Assignment 1

Peter Szilvasi

2024-01-21

Introduction

This report examines hourly earnings for Bookkeeping, accounting, and auditing clerks in the cps-earnings dataset. Utilizing Ordinary Least Squares (OLS) regression, four models are constructed with increasing complexity to predict earnings. The analysis focuses on identifying key predictors and assessing the trade-off between model intricacy and predictive accuracy using metrics like RMSE and BIC. The GitHub repository contains code and documentation, showcasing a systematic approach to regression modeling.

Data

For this occupation, the dataset contains 1229 observations. Missing values are only present in the 'ethnic' variable, which makes it not feasible for using.

In the target variable, I filtered for values smaller than 50, as there are some outliers. This decision was based on the 99% percentile which is 48.8. This excludes 12 observations, resulting in 1217 total observations for our models.

Predictor Variables

Before choosing my predictor variables, I carefully examined, distributions and frequency tables for all variables in the dataset. Decision to exclude or include was based on the variance and distribution in the values. Other variables were transformed into binary variables to help the performance of the models.

Predictor variables are age, agesq, female, white, ed_max_high_school_graduate, ed_some_college, ed_college_degree, federal_gov, local_gov, state_gov, profit_priv, nonprofit_priv, has_children

Variables federal_gov, local_gov, state_gov, profit_priv, and nonprofit_priv, derive from the original class column.

The grade92 variable was transformed into 3 binary variables. The ed_max_high_school takes a 1 value when grade92 is less than or equal to 39. Next, the ed_some_college refers to values 40-42, referring to started but not finished college studies, while ed_college_degree refers to people with an obtained college degree.

The has_children varible takes 1 when the person has 1 or more children according to the ownchild column.

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
age	1217.0	45.24	02.dec	16.0	36.0	47.0	55.0	64.0
agesq	1217.0	2191.42	1045.38	256.0	1296.0	2209.0	3025.0	4096.0
female	1217.0	0.91	0.29	0.0	1.0	1.0	1.0	1.0
white	1217.0	0.84	0.37	0.0	1.0	1.0	1.0	1.0
ed_max_high_school_graduate	1217.0	0.34	0.47	0.0	0.0	0.0	1.0	1.0
ed_some_college	1217.0	0.46	0.50	0.0	0.0	0.0	1.0	1.0
ed_college_degree	1217.0	0.20	0.40	0.0	0.0	0.0	0.0	1.0
federal_gov	1217.0	0.02	0.13	0.0	0.0	0.0	0.0	1.0
local_gov	1217.0	0.05	0.22	0.0	0.0	0.0	0.0	1.0
state_gov	1217.0	0.03	0.18	0.0	0.0	0.0	0.0	1.0
profit_priv	1217.0	0.83	0.38	0.0	1.0	1.0	1.0	1.0
nonprofit_priv	1217.0	0.07	0.25	0.0	0.0	0.0	0.0	1.0
has_children	1217.0	0.30	0.46	0.0	0.0	0.0	1.0	1.0

Regression Models

Using these predictor variables I built 4 linear regression models, each being more complex than the previous. For the whole Regression table, see Stargazer Table 1. in the Appendix. The four models include 2,3,7, 13 variables respectively. To compare the performance of each, I will look at the 3 different indicators:

a) RMSE in the full sample
b) Cross Validated RMSE
c) BIC in the full sample

RMSE comparison

Table 1.: RMSE Table

	Model	N_Variables	RMSE
0	Model 1	2	7.453
1	Model 2	3	7.420
2	Model 3	7	7.340
3	Model 4	13	7.226

The provided table showcases the RMSE values for the four models, offering insights into their respective predictive accuracies. Model 4 exhibits the lowest RMSE at 7.226, indicating superior performance in estimating hourly earnings. The incremental decrease in RMSE from Model 1 to Model 4 suggests that increasing model complexity contributes to enhanced predictive capabilities in our case.

Table 2.: Cross Validated RMSE

	Model1	Model2	Model3	Model4
Fold1	7.370064	7.329544	7.210816	7.072308
Fold2	7.687463	7.679463	7.617978	7.474266
Fold3	7.506197	7.488669	7.359756	7.295743
Fold4	7.229666	7.136231	7.099168	6.975240
Average	7.448347	7.408477	7.321929	7.204389

Table 2 presents Cross Validated Root Mean Squared Error (CV-RMSE) values for the four models across different folds. In each fold, Model4 consistently demonstrates the lowest CV-RMSE, indicating superior predictive accuracy compared to the other models. The average CV-RMSE values follow a decreasing trend from Model1 to Model4, suggesting that increasing model complexity contributes to enhanced generalization performance for my analysis.

BIC Comparison

Table 3.: BIC Comparison

	Model	N_Variables	BIC
0	Model 1	2	8363.85
1	Model 2	3	8360.28
2	Model 3	7	8355.21
3	Model 4	13	8352.47

Table 3 provides insights into the model selection process based on the BIC, Notably, Model 4, despite, having the highest number of variables (13), shows the lowest BIC at 8352.47. However, there are really small differences between the other BIC values. Based just on the BIC values, Model 1 is good enough as it is very simple.

Conclusion: Choosing the Model

Overall Model 4 always demonstrated the lowest RMSE value, which is logical as it had the most features. However, we do need to take BIC into consideration as it penalizes the models for their complexity and helps to avoid overfitting. I would continue with testing live, unseen data and see how the models compare there. If Model 4 continues to outperform others in predicting new data, it may be a suitable choice. However, if its performance does not significantly improve on new data, a simpler model with fewer variables (e.g., Model 3) might be preferred to avoid overfitting.

Appendix

Stargazer Table 1.

Dependent variable: hourly wages

	Model 1	**Model 2**	**Model 3**	**Model 4**
Intercept	3.372	4.042*	3.598**	4.404***
	(2.350)	(2.354)	(1.796)	(1.661)
age	0.618***	0.684***	0.663***	0.644***
	(0.120)	(0.120)	(0.120)	(0.126)
agesq	-0.006***	-0.007***	-0.007***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.002)
female		-2.515***	-2.296***	-2.239**
		(0.893)	(0.879)	(0.890)
white			-0.583	-0.159
			(0.612)	(0.599)
ed_max_high_school_graduate			-0.128	0.102
			(0.685)	(0.651)
ed_some_college			0.896	1.198*
			(0.657)	(0.623)
ed_college_degree			2.830***	3.105***
			(0.716)	(0.665)
federal_gov				5.775***
				(1.549)
local_gov				2.400**
				(1.030)
state_gov				-3.144***
				(0.845)
profit_priv				-1.115**
				(0.538)
nonprofit_priv				0.488
				(0.904)
has_children				-0.309
				(0.516)
Observations	1217	1217	1217	1217
R^2	0.029	0.038	0.058	0.087
Adjusted R^2	0.028	0.035	0.054	0.079
Residual Std. Error	7.462 (df=1214)	7.432 (df=1213)	7.361 (df=1210)	7.261 (df=1205)
F Statistic	27.466*** (df=2; 1214)	24.215*** (df=3; 1213)	1063.832*** (df=6; 1210)	647.596*** (df=11; 1205)
BIC	8363.85	8360.28	8355.21	8352.47
Note:				*p<0.1; **p<0.05; ***p<0.01