Data Analysis 3: Assignment 1

Muhammad Arbash Malik – 2202071
GitHub

Introduction:

The goal of this assignment is to build predictive models for our target variable, "earnings per hour" with our chosen predictor variables. The dataset used was the cps-earnings dataset that was read through using OSF website under Gabors Data Analysis datasets. It contains information about earnings of households for US in 2014.

Data Description:

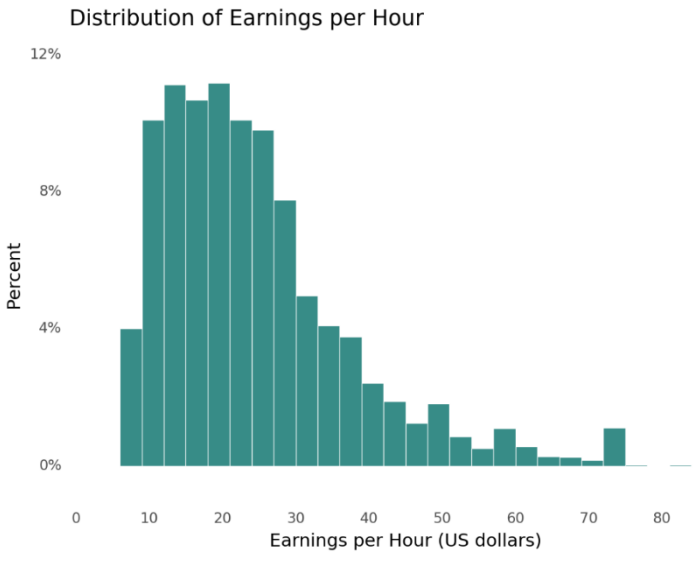
The dataset has 149,316 observations and 23 variables. I filtered the original dataset and focused only on the occupations that I chose i.e. Education, Training, and Library Occupations. This covers the education sector of the country. This resulted in 9,981 observations. From the 23 variables, I chose 9 predictor variables for my models:

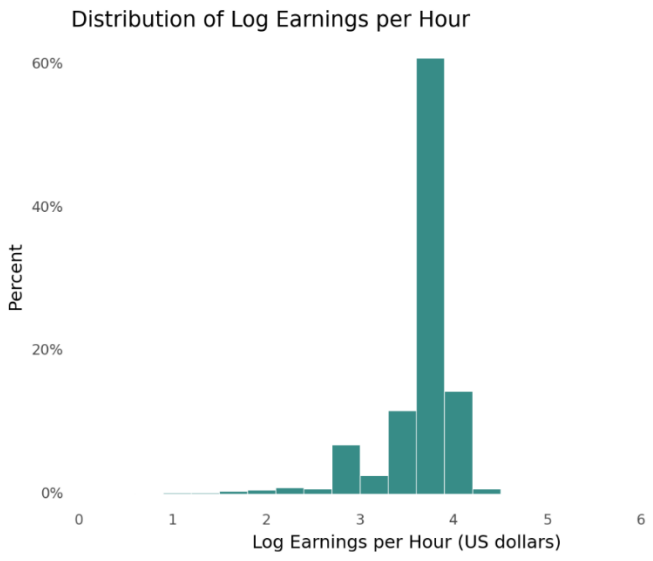
- **eph**: This variable was created to represent the average dollars earned per hour as the original data set only had weekly wages and weekly number of hours worked. The `earnwke` was divided by `uhours` to give our target variable
- **grade92:** This variable represents the level of education for an individual ranging from 31(less than 1st grade) to 46 (PhD) race –
- **age:** The age of each individual in the dataset.
- **sex:** The gender of each individual in the dataset taking 1 if male, and 2 if female.
- **marital**: The individual's marital status. Ranges from 1 (Married Civilian) to 7 (Never Married).
- **ownchild**: Whether the individual has children or not. Takes value 0 if not, else gives the number of children in primary family.
- **unionmme**: Whether the individual is a member of a union or not. Takes Yes or No values.
- **lsfr94:** Employment status at the time of survey, takes value of Employed-At Work or Employed-Absent
- **class:** Whether the individual is working in the government sector or a private sector. Takes 5 values for e.g Government – Local, Private – For Profit.
- **race:** The variable identifies the race of the individual ranges from 1 to 20. 1 being white.

Data Manipulation:

After choosing the predictor variables for my models, some basic data manipulation was done. Firstly, some of the observations were filtered out such as individuals having age less than 18, weekly earnings less than or equal to 0, and earnings per hour less than 7.5. A log transformation for earnings per hour `eph` was also done to compare which variable to use, which resulted in using the target variable without any transformations. However, I dropped extreme values that were greater than USD 54 (See Graph 1). A filter on education level was also done; the individuals should at least have high school diploma (grade92 >= 39).

From all the predictor variables, binary variables were created to represent each of the categorical values they had. For example, after filtering, grade92 had 8 values left (39-46), so 8 binary variables were created to represent the values. The same was done for every other predictor variable mentioned above. In total, 29 variables were created for the modelling.





As you can see from the graphs above the distribution is skewed for the variable `eph`. For the Log distribution, there are a lot of observations with values between 3.66 and 4, almost 60%. I decided to stick with simple target variable without the transformations, and after filtering out the extreme values, the number of observations were 9,136.

Models:

All models are constructed to make sure that standard errors are heteroscedasticity robust. As per the instructions, the first model was the simplest one where education level, and age was used (10 variables). The second model incorporated gender as well (11 variables). The third model then incorporated marital status, children status, and union membership (16 variables). The fourth model, which was the most complex one incorporated employment category, class/sector, race, and citizenship (30 variables). It also incorporated interactions as well.

	Dependent variable: eph			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Intercept	3.638***	4.794***	4.539***	1.407**
	(1.044)	(1.046)	(0.906)	(0.565)
Education + Age	Yes	Yes	Yes	Yes
Education + Age + Gender		Yes	Yes	Yes
Education + Age + Gender + Marital Status + Children + Union			Yes	Yes
Education + Age + Gender + Marital Status + Children + Union + Employment Category + Class/Sector + Race + Citizenship (Interactions)				Yes
Observations	9136	9136	9136	9136
R^2	0.204	0.217	0.235	0.243
Adjusted R^2	0.203	0.216	0.234	0.241
Residual Std. Error	9.177 (df=9126)	9.104 (df=9125)	9.001 (df=9121)	8.958 (df=9106)
F Statistic	5846.568*** (df=9; 9126)	5383.584*** (df=10; 9125)	4083.971*** (df=14; 9121)	2080.903*** (df=29; 9106)
RMSE	9.172	9.099	8.993	8.943
BIC	66511.94	66374.1	66197.77	66231.6
Note:				*p<0.1; **p<0.05; ***p<0.01

From the simplified stargazer table (see detailed stargazer in code), you can see that R^2 and Adjusted R^2 both increased from Model 1 to Model 4, a 4% increase in both, Model 4 has the highest of both.
Model 1 has the highest RMSE while Model 4 has the lowest RMSE of all the models, 9.172 and 8.943 respectively. However, Model 4 does not have the lowest BIC score, Model 3 takes the win here of a score of 66,197.

To further enhance our analysis, a k-fold cross-validation RMSE was done for all the models.

	Model 1	Model 2	Model 3	Model 4
Fold1	9.075392	9.014713	8.910633	8.867399
Fold2	9.165562	9.085607	8.989221	8.931774
Fold3	9.250644	9.175893	9.061422	9.005968
Fold4	9.191604	9.112324	9.003659	8.943017
Average	9.170801	9.097134	8.991234	8.937040

From the table above you can see that model 4 has the lowest score of 8.93. Although the RMSE score is slightly lower, I pick Model 3 as the best model compared to the others as the BIC score is lower and is lower in complexity meaning it will be easier to interpret as well.

Prediction:

Predictions were also done for our chosen model. It was done for both the 95% PI and the 80% PI. Two sets of values were chosen for the predictor variables are (variables not shown have value 0 – Set 2 is in the code):

Set 1:

- Age: 35
- Age Squared: 35^2
- Professional Degree: 1
- Female: 1
- Married: 1
- Ownchild: 2
- Employed-AtWork: 1
- Union Membership: 1
- Private – For Profit: 1
- White Person: 1
- US Native (Born in US): 1

	Model 3
Predicted	27.73
PI_low(95%)	10.00
PI_high(95%)	45.45

	Model 3
Predicted	27.73
PI_low(80%)	16.13
PI_high(80%)	39.32

From the tables it can be seen, the 95% prediction interval ranges from 10 (Low) to 45.45 (High), while for the 80% prediction interval it ranges from 16.13 (Low) to 39.32 (High).

Comparison of Models – Performance and Complexity

A comparison of models was also done to show the difference between the models predicted values and intervals. This was done for both sets of new observations for prediction, but I will show only the first one here.

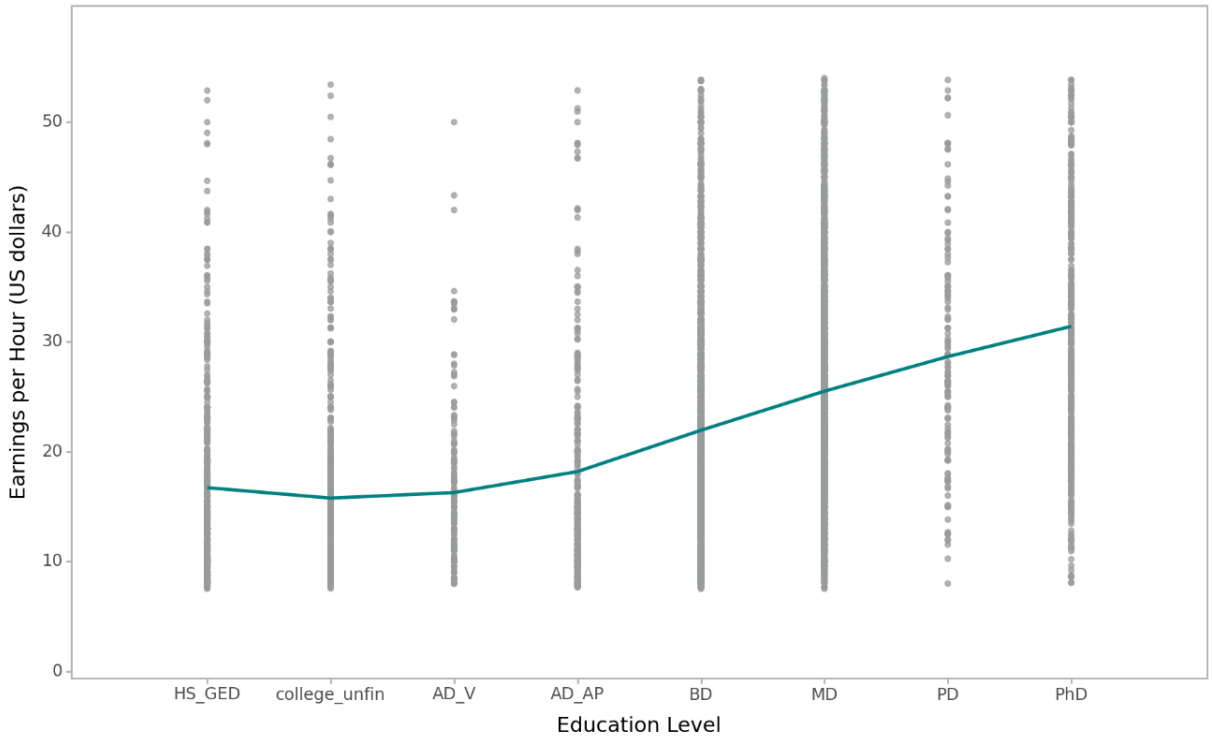
	Model 1	Model 2	Model 3	Model 4
Predicted	26.95	25.76	27.73	28.22
PI_low(95%)	8.88	7.83	10.00	10.57
PI_high(95%)	45.02	43.69	45.45	45.87

	Model 1	Model 2	Model 3	Model 4
Predicted	26.95	25.76	27.73	28.22
PI_low(80%)	15.14	14.04	16.13	16.68
PI_high(80%)	38.77	37.48	39.32	39.76

All the models were used to predict so a difference can be seen in the values. The overfitted models, Model 3 and Model 4 predict higher earnings per hour as compared to the relatively underfitted models, Model 1, and Model 2. But the difference is almost less than 1.5 USD. So, the increase in complexity has not related to much increase in performance. We can also see that the prediction intervals are also similar (not in values, but in the range).

Appendix

Graph 1: Lowess for Education Level



Graph 2: Lowess for Age