The Impact of Educational Expenses of New Coders*

The Factors Affecting Education Costs of Beginner Programmers to Learn to Code aside from University Tuition

Bongju Yoo

25 April 2021

1 Introduction

With an increasing interest in computer programming and the growth of coding related jobs, more and more people invest their time and money in taking programming courses (Hughes 2020). In this paper, I examine some of the factors that may affect educational expenses of new coders or entry-level programmers aside from university tuition, mainly focusing on the level of education, the size of city, commuting time, income, and learning time. Approximately 20,000 beginner programmers had participated in the survey (freeCodeCamp 2017).

Firstly, I processed correlation analysis to identify the relationship between each independent variable and dependent variable (education costs) pair and found that all the pairs have a positive linear relationship except for commuting time. Secondly, I used multiple linear regression to determine the statistical significance of the factors. The result of the analysis shows that commuting time is not statistically significant. Interestingly, education background is also not statistically significant associated with educational expenses of beginner programmers. Other variables, such as the size of city, income, and learning time, have a statistical significance on the costs.

This analysis shows that beginner programmers living in small towns and/or low-income households are likely to spend less money on learning to code aside from university tuition than those who live in bigger cities and/or high-income households. It is assumed that there is still a lack of educational resources for new coders living in rural areas and/or low-income households. Free online learning platforms such as freeCodeCamp can be a great resource of education for the coders.

The paper is organized as follows. The Data section describes features of the original survey data and how the data is preprocessed. The Model section explains the multiple linear regression model used to assess the association between each factor and educational costs of beginner programmers, and evaluates the model using its residual standard error, multiple R-squared, and F-statistic. The Results section summarises the results of the regression model and model evaluation processes in the Model section. Lastly, the Discussion section discusses the findings and potential limitations of the paper, and suggests directions for future research related to this data.

2 Data

*https://github.com/bonjwow/new-coders

Table 1: A partial view of the cleaned dataset

Gender	Age	City Population	Commute Time	Income	Months Programming	School Degree	Money For Learning
0	21	2	2	13000	5	0	1000
0	26	1	0	24000	5	1	0
1	29	1	3	40000	12	1	0
1	23	0	0	30000	29	1	700
0	20	0	1	20000	5	2	0
0	18	0	1	10000	3	0	0
0	32	0	3	20000	1	3	0
0	29	1	3	60000	12	3	200
1	46	2	0	76000	14	2	500
0	31	0	3	34000	28	1	500

2.1 Data Collection

The survey was run by `freeCodeCamp`, a non-profit organization that helps people learn to code through their free online courses and build a network for their alumni (Larson 2019). The purpose of the survey was to examine how the users learn to code (Larson 2017). The survey was conducted over the internet, and the survey respondents were limited to those persons with less than 5 years learning programming; all the respondents were asked whether they have already been coding for more than 5 years or not before starting the survey (Larson 2017). The survey is composed of 48 questions and takes about five minutes to complete, and the survey results are under the Open Data Common License, which can be freely distributed through the organization' GitHub repository (Larson 2017).[1]

2.2 Description of Dataset

The original dataset used for this paper is obtained from `freeCodeCamp`'s GitHub page. The format of the dataset is a comma-separated values (CSV) file, which contains 136 columns and 18,175 observations. I selected 8 variables and cleaned the data using the R programming language (R Core Team 2020) and the `tidyverse` package (Wickham et al. 2019) and the `dplyr` package (Wickham et al. 2021). The selected variables are: Age, CityPopulation, CommuteTime, Gender, Income, MoneyForLearning, MonthsProgramming, and SchoolDegree. The CityPopulation is the estimated number of city population of the recipient; the question asked was "About how many people live in your city?", and there are three options to choose: "less than 100,000", "between 100,000 and 1 million", and "more than 1 million". Since these answers were coded as strings, I converted them into numeric variables using the `recode` function of `dplyr` (Wickham et al. 2021). The answers for CommuteTime and SchoolDegree were also coded as strings in the original dataset, so I applied the same data cleaning process to the variables as I did with CityPopulation. Also, I omitted observations which have a missing value. Table 1 is a partial view of the cleaned dataset after the preprocessing.

Table 2 displays descriptive statistics for the cleaned data after the preprocessing. The total number of observations is 7,022 and the data type for all variables are numeric variables. The median age of the respondents is 30 years and the median income of them is about $43,000 in U.S. dollars. The average of the respondents has spent about 2 years and about $1,000 in U.S. dollars learning to code.

[1] https://github.com/freeCodeCamp/2017-new-coder-survey

Table 2: Descriptive statistics for the cleaned dataset

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Gender	7,022	0.174	0.379	0	0	0	1
Age	7,022	29.774	7.684	13	25	33	71
CityPopulation	7,022	1.221	0.777	0	1	2	2
CommuteTime	7,022	2.199	1.451	0	1	3	5
Income	7,022	42,966.890	59,162.290	6,000	17,000	55,000	1,000,000
MonthsProgramming	7,022	23.986	46.496	0	3	26	744
SchoolDegree	7,022	1.676	0.968	0	1	2	4
MoneyForLearning	7,022	1,032.273	4,030.722	0	0	399	170,000

Table 3: Correlation matrix for variables

	1	2	3	4	5	6	7	8
1. Gender	-							
2. Age	0.03**	-						
3. CityPopulation	-0.02*	-0.08***	-					
4. CommuteTime	-0.01	-0.01	0.17***	-				
5. Income	0.01	0.16***	0.03**	0.03*	-			
6. MonthsProgramming	-0.08***	0.12***	0.02	0.01	0.11***	-		
7. SchoolDegree	0.12***	0.18***	0.12***	0.07***	0.06***	0.07***	-	
8. MoneyForLearning	0.04***	0.05***	0.05***	0.01	0.05***	0.06***	0.04**	-

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2.3 Correlation Analysis

I ran a correlation matrix to check correlation coefficients between the variables. To create the correlation matrix, I used a custom function written by Stefan Eng (Eng 2018). His function contains the `Hmisc` package's `rcorr` function to calculate the correlations with the p-values and display starts based on the significance level of each cell when the value is lower than a specific level (Harrell Jr, Charles Dupont, and others. 2021). Along with his custom to compute the correlation matrix, I used the `kable` package to print out the result in a table format (Zhu 2020). As can be seen in Table 3, the result of the correlation matrix shows that all independent variables have a positive significant relationship with the dependent variable, MoneyForLearning, except for CommuteTime. This can be interpreted that a commuting time to work does not affect educational expenditure of new coders aside from their post-secondary education cost.

3 Model

3.1 Model Formulae

In the earlier section, I ran a correlation matrix to identify the correlation of each variable and the significance of the relationship between the dependent variable (MoneyForLearning) and other dependents variables. But, I was not able to estimate how much the independent variables impact the dependent variable from the correlation analysis. In this section, I will examine effects of the dependent variables on the dependent variable, focusing on one of the dependent variables: Income. The two main questions that I set out to answer in this section are: (1) How do beginner programmers' income affect the cost of their education to learn code? and (2) Will the effects change according to the population of the city where they live in.

To obtain the answer for the first question, I used a simple linear regression model (Equation (1)) and multiple linear regression models (Equation (2) and (3)). All these three models use MoneyForLearning (the cost of their education to learn code) as a dependent variable. Model 1 contains only one independent variable, which is Income (see Equation (1)); otherwise, Model 3 uses all the independent options except for CommuteTime: Gender, Age, CityPopulation, Income, MonthsProgramming, and SchoolDegree (see Equation (3)). As mentioned in the Correlation Analysis section, the CommuteTime (a commuting time) variable does not significantly affect the response variable, so it is excluded from the regression models. And, Model 2 uses the same independent variables, excluding Gender and Age. For the second question about the effects of the population of the city where the programmers reside, I used four simple linear regression models (see Equation (2)). Like Model 1, all these models use MoneyForLearning as a dependent variable and Income as an independent variable (see Equation (1)). But, the observations were grouped by the population size of the programmers. Model 4 contains any size of population; and Model 5,6, and 7 are the small (less than 100,000), medium (between 100,000 and 1 million), and large (more than 1 million) size population.

$$Y \sim \beta_0 + \beta_1 X_1 + \epsilon \tag{1}$$

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon \tag{2}$$

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon \tag{3}$$

3.2 Data Preprocessing

As explained in the Data section, categorical variables were converted to continuous variables for the linear regression models. For example, the SchoolDegree variable (the level of education) was a categorical variable with character values. So, I converted the variable into a variable with numeric values, using the `dplyr` package. A complete list of survey questions and response options for each question can be found in the Appendix. Since the MoneyForLearning variable (the level of education) was not only categorical but also ordered and ranked in the original dataset, I converted the variable into a variable with numeric values. Also, I merged some of the response options of the variable. For instance, it is somewhat hard to say that a high school diploma or equivalent (GED) is higher than trade, technical, or vocational training, and a professional degree, such as MBA, MD, or JD, is higher than a non-professional master's degree. So, I merged the options, using the `recode` function of the `dplyr` package (Wickham et al. 2021).

3.3 Model Validation

I will use Mean Square Error (MSE) and Root Mean Square Error (RMSE) to evaluate the accuracy of the model. As the name implies, MSE is defined as the mean of the square of the residuals obtained from a regression model, and RMSE is the square root of MSE (Holmes 2000). I will use the residuals which are the results of the `lm` function to calculate MSE and RMSE of the models. Firstly, I will calculate the square of the residuals, and then computed the mean of the squared value for MSE (see Equation (4)). Secondly, for RMSE, I will use the `sqrt` function to calculate the square root of MSE (see Equation (5)). Once I obtain the results of MSE and RMSE, I will compare them to see how accurately the models predict the response.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(residual_i)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(residual_i)^2} \tag{5}$$

Since two multiple linear regression models (Model 2 and 3) are used in the first group of regression models, I will check if multicollinearity exists in these regression models, using variance inflation factor (VIF) and tolerance. I will use the `car` package to calculate the values of VIF and tolerance for each model (Fox and Weisberg 2019). Also, I will identify how much each of the independent variables affects the dependent variable with the standardized regression coefficients (beta). the `QuantPsyc` package will be used to compute the beta value (Fletcher 2012).

4 Results

4.1 Results of Regression Model

All the regression models in the first group of the models had p-values of F-statistic lower than a significance level of 0.05 (see Table 4). And, each of the independent variables also had a p-value of less than 0.05 except for the SchoolDegree variable in Model 3. This can be interpreted that all the independent variables excluding the SchoolDegree variable have an impact on the response variable (MoneyForLearning), and the level of education (SchoolDegree) is not statistically significant on the response variable with holding all other variables. However, the SchoolDegree variable becomes statistically significant if the Gender and Age variables are excluded from the regression model.

The second group of regression models to examine the effects of the population of the city where the programmers reside shows that all the models have statistically significant p-values of F-statistic except for Model 5, the model for the small city with less than 100,000 residents (see Table 5). In other words, beginner programmers' income affects the cost of their education to learn code; however, their income does not affect the cost when the model contains only observations from the small city.

4.2 Results of Model Validation

Table 6 is the results of MSE and RMSE of the first group consisting of Model 1, 2, and 3. Among these three models, Model 3 containing all independent variable options has the lowest value of RMSE. So, it can be said that Model 3 predicts the response variable more accurately with holding all independent variables, compared to Model 1 and 2 which contains the Income variable or excludes the Age and Gender variables.

Table 7 is the results of MSE and RMSE of the second group consisting of Model 4, 5, 6, and 7. Among these four models, Model 5 (the regression model for the small city) has the lowest MSE and RMSE. However, as explained earlier, the p-values of F-statistic for the model was higher than a significance level of 0.05, which means the model is not statistically significant. The most accurate regression model excluding Model 5 among the four models was Model 4 which includes cities of all sizes, followed by Model 7.

Model 3 and 4 are the most accurate model in each of the groups. However, as can be seen in Table 6 and 7, all the models have an extremely low R-squared value, which is very close to 0. In other words, the accuracy of the regression models is intensively low because the residuals (the difference between the measured values and the predicted values of the models) are high. Thus, it is hard to say that those six variables (Gender, Age, CityPopulation, Income, MonthsProgramming, and SchoolDegree) have a significant effect on education costs of beginner programmers to learn to code.

Table 8 is the results of the multicollinearity diagnostic test for Model (2). The results show that the MonthsProgramming variable has the highest beta value, which means the independent variable has the most significant effect on the dependent variable (MoneyForLearning) among those four independent variables included in the regression model. However, all the beta values of the independent variables are extremely low and close to 0, meaning that their effects are not significant. All the VIF and tolerance of each variable are very close to 1, which means there is almost no multicollinearity; a VIF of 10 or higher, or tolerance close 0 indicates that the model might have multicollinearity (Williams 2015).

Table 4: Predictions of money for learning

	Dependent variable:		
	MoneyForLearning		
	Model (1)	Model (2)	Model (3)
Gender			452.228***
			(127.718)
Age			21.376***
			(6.473)
CityPopulation		249.369***	279.850***
		(62.105)	(62.414)
Income	0.003***	0.003***	0.002***
	(0.001)	(0.001)	(0.001)
MonthsProgramming		4.469***	4.427***
		(1.038)	(1.046)
SchoolDegree		114.783**	62.563
		(50.043)	(51.152)
Constant	895.500***	317.123***	−328.825
	(59.389)	(120.613)	(215.362)
Observations	7,022	7,022	7,022
R^2	0.002	0.008	0.012
Adjusted R^2	0.002	0.008	0.011
Residual Std. Error	4,026.607 (df = 7020)	4,014.821 (df = 7017)	4,008.569 (df = 7015)
F Statistic	15.358*** (df = 1; 7020)	14.932*** (df = 4; 7017)	13.970*** (df = 6; 7015)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5: Predictions of money for learning by city population

	Dependent variable:			
	MoneyForLearning			
	Model (4): Any	Model (5): Small	Model (6): Medium	Model (7): Large
Income	0.003***	0.001	0.007***	0.002**
	(0.001)	(0.001)	(0.002)	(0.001)
Constant	895.500***	551.187***	792.127***	1,096.501***
	(59.389)	(71.138)	(133.695)	(87.232)
Observations	7,022	1,516	2,436	3,070
R^2	0.002	0.001	0.004	0.002
Adjusted R^2	0.002	0.001	0.003	0.002
Residual Std. Error	4,026.607 (df = 7020)	2,254.242 (df = 1514)	4,755.348 (df = 2434)	4,063.395 (df = 3068)
F Statistic	15.358*** (df = 1; 7020)	2.139 (df = 1; 1514)	9.074*** (df = 1; 2434)	5.671** (df = 1; 3068)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: Results of model validation metrics for model group 1

Model	R-squared	MSE	RMSE
Model (1)	0.002	16,208,945.702	4,026.033
Model (2)	0.008	16,107,307.380	4,013.391
Model (3)	0.012	16,052,605.147	4,006.570

Table 7: Results of model validation metrics for model group 2

Model	R-squared	MSE	RMSE
Model (4)	0.002	16,208,945.702	4,026.033
Model (5)	0.001	5,074,901.583	2,252.754
Model (6)	0.004	22,594,770.721	4,753.396
Model (7)	0.002	16,500,421.499	4,062.071

Table 8: Results of multicollinearity diagnostic metrics for Model (2)

	CityPopulation	Income	MonthsProgramming	SchoolDegree
Beta	0.048	0.038	0.052	0.028
VIF	1.015	1.015	1.015	1.021
Tolerance	0.985	0.985	0.985	0.979

Table 9 is the results of the multicollinearity diagnostic test for Model (3). The variable which has the highest beta value is the CityPopulation variable, which means the size of the city has the most significant effect on education costs of beginner programmers to learn to code among those six independent variables used in the regression model. Like the results of the test for Model (2), Model (3) also does not have significant multicollinearity; the VIF and tolerance of each variable of the model are in an acceptable range. In concussion, there is no significant multicollinearity found in both Model (2) and (3). However, all the regression models have an extremely low R-squared value, meaning the regression models have poor accuracy of prediction.

Table 9: Results of multicollinearity diagnostic metrics for Model (3)

	Gender	Age	CityPopulation	Income	MonthsProgramming	SchoolDegree
Beta	0.042	0.041	0.054	0.032	0.051	0.015
VIF	1.023	1.081	1.028	1.036	1.034	1.071
Tolerance	0.978	0.925	0.972	0.965	0.967	0.934

5 Discussion

5.1 Summary

In this paper, I examine potential factors that may affect costs of beginner programmers to learn to code aside from post secondary education based on the online survey run by freeCodeCamp in 2017. Firstly, I chose eight variables, including the dependent variable (educational expenses of the programmers) from the original dataset. Secondly, I ran a correlation matrix for all the independent variables to examine how closely they are related with the dependent variable. Thirdly, I used linear regression models to determine influence factors on the dependent variable from the independent variables excluding a commuting time which was confirmed not to have the correlation with the dependent variable. Fourthly, I ran two groups of linear regression models to examine effects of the dependent variables on the dependent variable: the first group was for identifying how programmers' income affects their educational costs to learn to code and the second group was for looking into how the effect of their income changes according to the population of the city where the programmers live in. Finally, I evaluated the models using R-squared, MSE and RMSE scores of each model. Also, I used beta, VIF, and tolerance scores to check if there is multicollinearity in the multiple linear regression models.

Summary of the findings are as follows: (1) The statistical relationship exists between all the independent variables excluding a commuting time and new coders' expenses to learn to code aside from their college or university tuition, (2) The level of education is not statistically significant on the new coders' expenses to learn to code, (3) Their income is statistically significant on the costs; but, the prediction model can become less accurate if the model includes only response data of the programmers from the small city, (4) There is no multicollinearity found in the multiple linear regression models, and lastly (5) All the models used in this paper have an extremely low R-squared score, which means that it is difficult to argue that those selected independent variable significantly affect the independent variable. In short, it can be argued that there is a correlation between each of the independent variables (the level of education, the size of city, a commuting time, income, and learning time) and the dependent variable (the cost of learning to code); however, there is no significant causal relationship between the variables.

5.2 Casuality

In addition to the low R-squared scores, there is another reason that causality cannot be explained from the regression models. In reality, there will be plenty of factors that actually affect beginner programmers' expenses to learn to code besides those selected seven variables. For example, there might be some institutions where the programmers can learn to code near their home, or some of them might be heavily exposed to online advertisements of some online-learning platforms that teach programming. In other words, realistically, it is almost impossible to measure all the variables happening in the large world and include them in the model. However, we can always measure which variables are significant and determine the most appropriate variables to make a model closer to the reality. For future studies, I suggest running the same analysis process, from the correlation analysis to linear regression models and model validation, with other variables that are not covered in this paper. Also, surveys of other years can be used to compare the results of this analysis.

Appendix

Survey questions

Q. [Age] How old are you?

- Type: integer

Q. [CityPopulation] About how many people live in your city?

- Type: string
- Options:

    1. less than 100,000
    2. between 100,000 and 1 million
    3. more than 1 million

Q. [CommuteTime] About how many minutes does it take you to get to work each day?

- Type: string
- Options:

    1. I work from home
    2. Less than 15 minutes
    3. 15 to 29 minutes
    4. 30 to 44 minutes
    5. 45 to 60 minutes
    6. More than 60 minutes

Q. [Gender] What's your gender?

- Type: string
- Options:

    1. male
    2. female

Q. [Income] About how much money did you make last year, in US dollars?

- Type: integer

Q. [MoneyForLearning] Aside from university tuition, about how much money have you spent on learning to code so far, in US dollars?

- Type: integer

Q. [MonthsProgramming] About how many months have you been programming for?

- Type: integer

Q. [SchoolDegree] What's the highest degree or level of school you have completed?

- Type: string
- Options:

1. no high school (secondary school)
2. some high school
3. high school diploma or equivalent (GED)
4. trade, technical, or vocational training
5. some college credit, no degree
6. associate's degree
7. bachelor's degree
8. master's degree (non-professional)
9. professional degree (MBA, MD, JD, etc.)
10. Ph.D.

References

Eng, Stefan. 2018. *Create an Apa Style Correlation Table with R.* https://stefaneng.github.io/apa_correlation_table/.

Fletcher, Thomas D. 2012. *QuantPsyc: Quantitative Psychology Tools.* https://CRAN.R-project.org/package=QuantPsyc.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression.* Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

freeCodeCamp. 2017. *2017 New Coder Survey.* https://github.com/freeCodeCamp/2017-new-coder-survey.

Harrell Jr, Frank E, with contributions from Charles Dupont, and many others. 2021. *Hmisc: Harrell Miscellaneous.* https://CRAN.R-project.org/package=Hmisc.

Holmes, Susan. 2000. *RMS Error.* https://statweb.stanford.edu/~susan/courses/s60/split/node60.html.

Hughes, Owen. 2020. *Developer Training Sees Spike in Demand as More People Learn to Code.* https://www.techrepublic.com/article/the-economic-outlook-is-uncertain-so-more-people-want-to-become-developers/.

Larson, Quincy. 2017. *We're Building a Massive Public Dataset About New Coders.* https://www.freecodecamp.org/news/take-the-2017-new-coder-survey-and-help-us-build-a-massive-public-dataset-8c808cbee7eb/.

———. 2019. *About freeCodeCamp - Frequently Asked Questions.* https://www.freecodecamp.org/news/about/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Williams, Richard. 2015. *Multicollinearity.* https://www3.nd.edu/~rwilliam/stats2/l11.pdf.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.