Income Related Factors in Data Science and Machine Learning

0. Header

James Gaxiola: 1, 2.1, 2.2, 2.3, 2.5

Veronica Morad: 2.4, 2.5, 3, 4

1. Introduction

The data comes from a 2020 Kaggle Machine Learning & Data Science Survey completed by 20,036 professionals in the fields of data science and machine learning from around the world. It ran for 3.5 weeks in October 2020 and recorded responses to 47 questions concerning things like country, age, gender, education level, income, and programming language, environment, and platform preferences. The data present interesting insight on the current state of data science and machine learning.

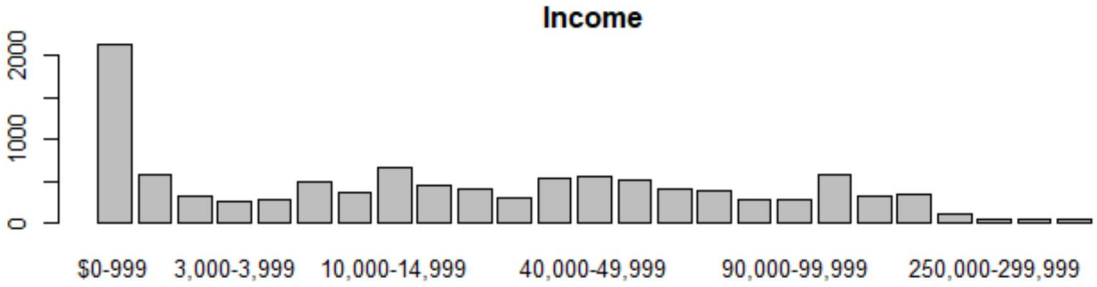
We decided to work with the data concerning income, education level, and programming language preference. We clean and visualize the data, provide point estimates and interpretations, transform the data and evaluate its fit, construct a linear regression model, and test for the differences between educational degrees. We determined that the income data was skewed and that the best model was found by cleaning the $0-$999 values out of the income data and taking the logarithm of it. Using point estimates and proportions we found that there is likely a relationship between income and education as well as income and programming language. We found the relationship between income and education was not significant via linear regression. Lastly, we used two proportion confidence intervals to determine that there is likely a difference pertaining to education and income.

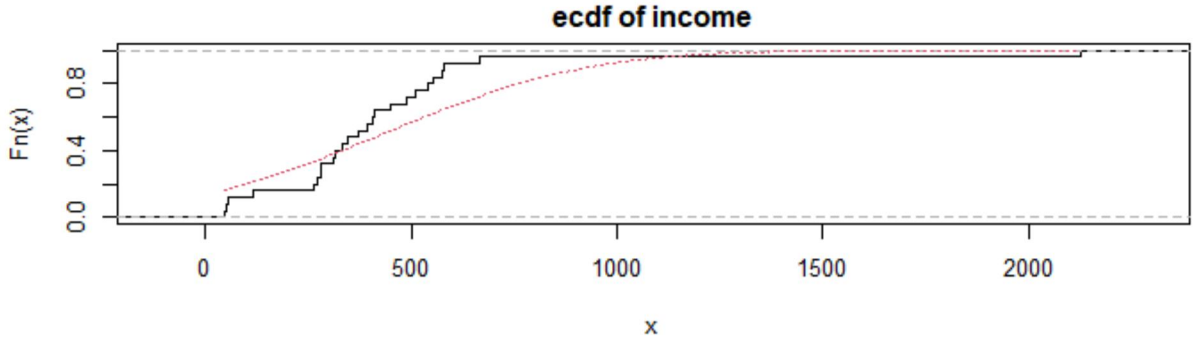
2.1 Data cleaning and initial visualization

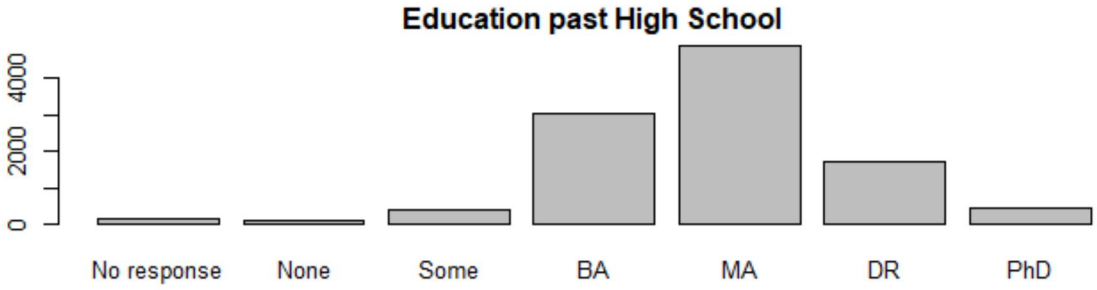
Before doing any analysis, we removed the data where responses to the questions of interest (salary and education) were N/A. This way we could work on discerning the relationship between the two variables.

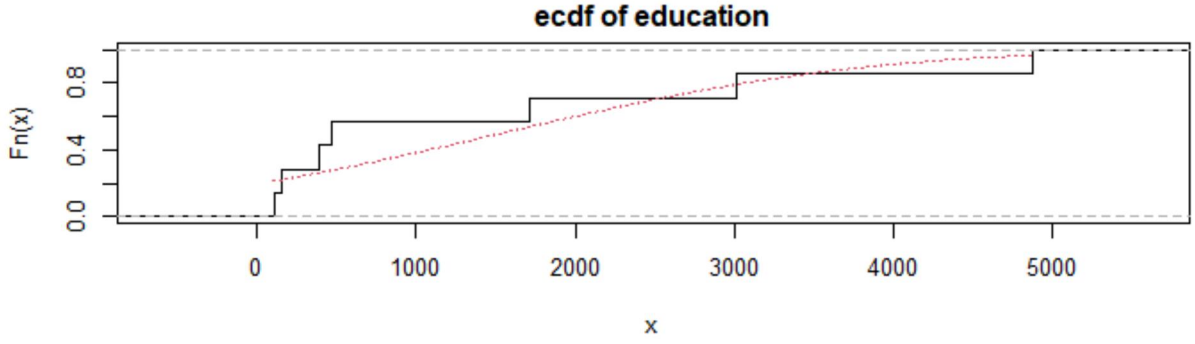
When plotting the bar plot of income, we see the data is strongly right skewed with many values between $0-$999. This may be due to error in data entry and the fact that some respondents live in developing countries where salaries and cost of living would generally be lower. Many of the respondents that reported a salary of $0-$999, for instance, live in India. The conversion rate of U.S. dollars may exacerbate this. As one would expect, the values above $150,000 are low, meaning that there are few respondents whose salary exceeds that. Otherwise the data seems pretty evenly distributed across income levels; again the differential cost of living and monetary exchange rates may play a role. Next we look at the ecdf of income and see that the data looks somewhat similar to what we would expect of a normal distribution. This will be further checked later on.

When plotting the bar plot of highest education level, we see that there is a peak around MA. This indicates that the most respondents have a Master's degree, followed by Bachelor's, and then Doctorate. The least respondents (ignoring those who did not respond) responded no education past high school. Next we look at the ecdf of education and see that the data looks somewhat similar to what we would expect of a normal distribution.


Income
0
1000
2000
$0-999
3,000-3,999
10,000-14,999
40,000-49,999
90,000-99,999
250,000-299,999


ecdf of income
Fn(x)
0.0
0.4
0.8
0
500
1000
1500
2000
x


Education past High School
0
2000
4000
No response
None
Some
BA
MA
DR
PhD


ecdf of education
Fn(x)
0.0
0.4
0.8
0
1000
2000
3000
4000
5000
x

2.2 Point estimates

Using proportions, we form a few conclusions about the relationships between income, education, and programming language.

We find that 15.6% of respondents have a Bachelor's degree and make below $100,000 while 26.3% of respondents have a Master's degree and make below $100,000. We find that 2.2% of respondents have a Bachelor's degree and make above $100,000 while 5.9% of respondents have a Master's degree and make above $100,000. Thus out of the respondents who make over $100,000, 26.6% of them have a Bachelor's degree and 73.3% of them have a Master's degree. With this difference, it seems that educational degree is related to salary.

We find that 8.6% of respondents primarily use R and make below $100,000 while 32.6% of respondents primarily use Python and make below $100,000. We find that 1.9% of respondents primarily use R and make above $100,000 while 6.9% of respondents primarily use Python and make above $100,000. Thus out of the respondents who make over $100,000, 21.4% of them prefer R and 78.6% of them prefer Python. With this difference, it seems that programming language is related to salary.

Income	Bachelor's degree	Master's degree	Total
Below $100,000	1,961	3,305	5,266
$100,000 and up	271	744	1,015
Total	2,232	4,049	12,562

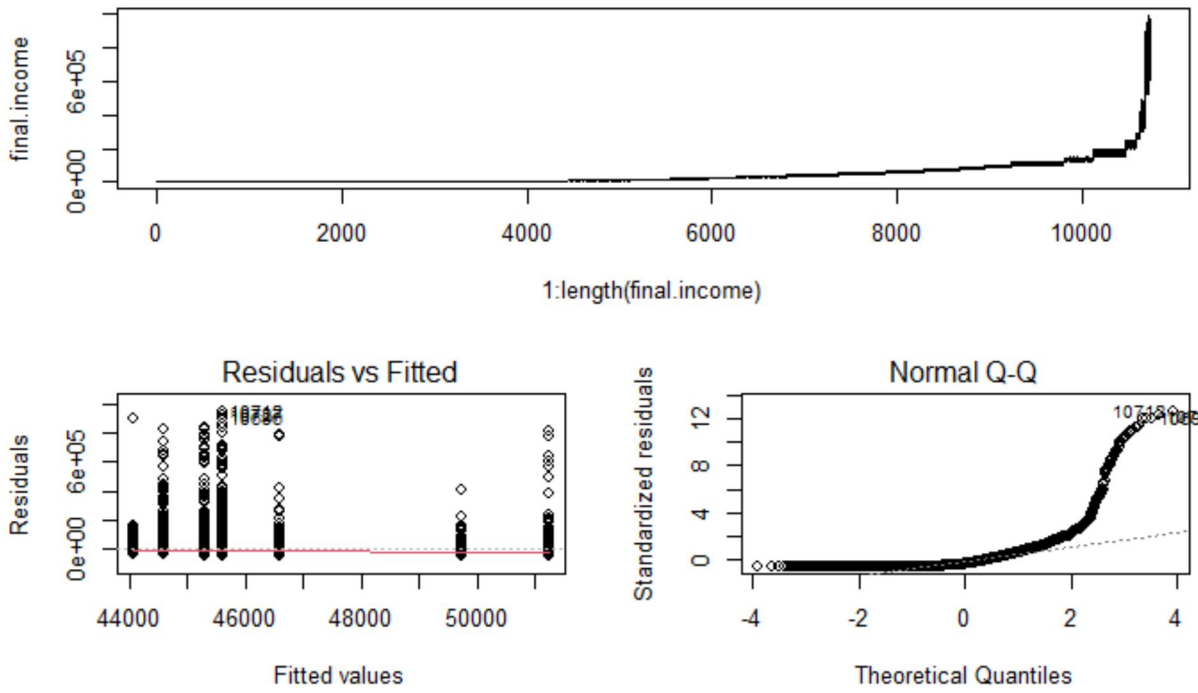
Income	R	Python	Total
Below $100,000	1,453	5,517	6,970
$100,000 and up	320	1,173	1,493
Total	1,773	6,690	16,926

Note: the cleaned data was used to construct the table for income vs. degree and the raw data was used to construct the table for income vs programming language

2.3 Graphical Analysis

We can see from the residuals plot and the Q-Q plot that the raw, cleaned data is not normal. The residuals are not randomly scattered and present a clear trend in the positive direction while the Q-Q plot is nonlinear and does not follow the expectation of the standardized residuals.

When plotting the length of the income with the income, the graph is exponential. Hence we chose to perform a log transformation on the cleaned data to make it more linear and to minimize the range of the y values for analytical purposes.

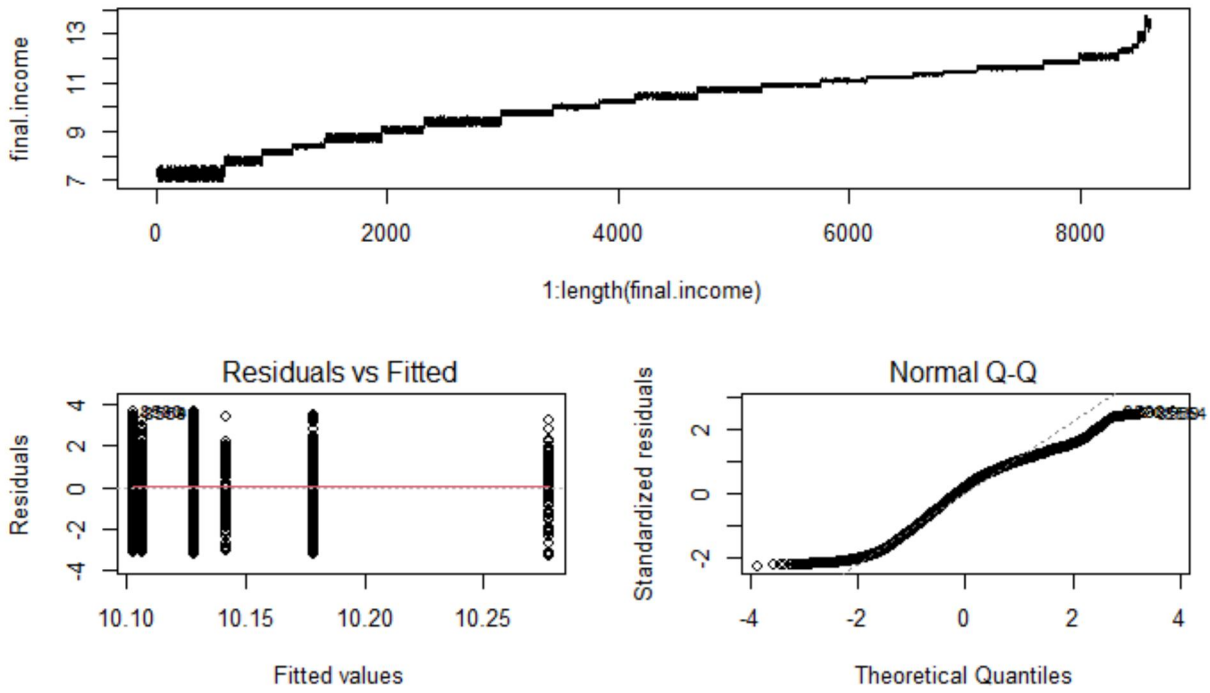


2.4 Transforming the data

Once we take the log transformation we plot the income with the length of the income, the residuals, and the Q-Q plot once again.

The residuals are more randomly distributed and the Q-Q plot follows the expected standardized residual values more closely however the plot of length of income with income looks unusual in the lower range. Thus we can conclude that the log transformation was successful in normalizing the data.

Because of the clustering at the lower end, we chose to remove the income values from $0-$999 to better perform linear regression. After removing these values, the plot of length of income with income looks more linear and balanced, the residuals are still randomly scattered, and the Q-Q plot follows the expected standardized residual values even more closely.



Log transformation, after removing $0-$999 values

2.5 Linear regression

After running linear regression on the cleaned data, we get slopes for each of the educational levels, relating education to income.

The model predicts that respondents with doctoral degrees make 0.025 less than respondents with Bachelor's degrees (the reference level). Those with Master's degrees are predicted to make 0.005 more than those with Bachelor's degrees. Those who have no formal education past high school make 0.149 more than those with Bachelor's degrees. Those with a professional degree make 0.0499 more than those with Bachelor's degrees. Lastly, those who have some college or university education without earning a Bachelor's degree make 0.021 less than those with Bachelor's degrees.

These unusual results may be influenced by the fact that Bachelor's and Master's degrees received the most responses and thus may have a wider range of income values while unusual income values in the lower tiers of education may be overrepresented.

Running an ANOVA test we see that the p-value is 0.9371, meaning that educational degree doesn't significantly affect the respondents' income.

```
Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                     10.1282516  0.0305591 331.432   <2e-16 ***
eduDoctoral degree                                              -0.0251928  0.0482599  -0.522    0.602
eduI prefer not to answer                                        0.0136365  0.1504861   0.091    0.928
eduMasterâ€™s degree                                              0.0005052  0.0380611   0.013    0.989
eduNo formal education past high school                          0.1489686  0.1663026   0.896    0.370
eduProfessional degree                                           0.0499733  0.0803909   0.622    0.534
eduSome college/university study without earning a bachelorâ€™s degree -0.0213168  0.0922680  -0.231    0.817
---
```

```
Analysis of Variance Table

Response: final.income
             Df  Sum Sq Mean Sq F value Pr(>F)
edu           6     3.8  0.6256  0.3001 0.9371
Residuals  8594 17913.1  2.0844
```

3. Advanced analysis

We conducted a 2 proportion z test on the proportion of people who make between $100,000-$124,999 and have a Master's degree as well as the proportion of people who make between $100,000-$124,999 and have a Bachelor's degree. The conditions for this test are met by the central limit theorem. We let p.m for the proportion for Masters and p.b be the proportion for Bachelors.

Constructing a confidence interval for the difference between these proportions, p.m-p.b, we get a confidence interval of (0.282, 0.405). Therefore, a confidence interval for p.m is (p.b + 0.282, p.b + 0.405) and we made two conclusions. It is very likely that there is a positive difference between the two as the interval does not contain zero and that a person having a

Master's degree instead of a Bachelor's degree increases the likelihood of making between $100,000-$124,999 by an amount between 28.2% and 40.5%

4. Conclusion/Discussion

In order to work with the data we had to remove sections the respondents left unanswered, for both question 4 and 24 simultaneously, since we did not have one to be answered while the other was not. It should be noted that this cut the data by approximately a half, but because the raw data was already so large we were still left with a large sample we do analysis on. In order to perform a linear regression, we had to transform salary from a categorical variable to numerical. We did this by constructing a simulation so that we can get a random uniform number between the categorical income variables. This number of random uniform numbers this simulation produced is equal to how many observations there are in each income bracket. We noticed that for the last income bracket it only said greater than $500k, so for an upper limit we choose a realistic inclusive upper limit of 1 million, however since we do not know have access to the respondents actual income for this response, it is possible that that number was higher. Finally we had to remove the values of $0-$999 because the distribution was so heavily right skewed, it affected the normality of the model, and was not suitable for linear regression. Future studies may focus on a data scientist's lifestyle as being supported by yearly compensation. For instance, in the report it was observed that a lot of the respondents who said they made between $0-$999 USD were from India and although looking at that compensation numerically seem low, in comparison to the cost of life in India they may actually be living a comfortable lifestyle. Other work may also compare variables between countries. For example, the number of women in the field may be different in countries with different social climates; or the level of education may reflect the abundance and accessibility of higher

education in different countries. It would also be interesting to compare the overall differences between developing and developed countries.