Salary Data Analysis and Regression

Lorenzo, Hayden, Sophia

2024-03-13

Authors

Hayden Zhong [henryz3@uw.edu] – First author of Job Title Categorization sections

Lorenzo Wahaus [ldw539@uw.edu] – First author of Linearity, collinearity, interactions, and outliers sections

Sophia Chiesa [sc933@uw.edu] – First author of Introduction, Education Level Impact sections

Introduction

The world of career salaries, especially from the perspective of undergraduate students on the cusp of joining the full-time workforce, is complex and enigmatic. Due to legal and cultural confidentiality, and the fact that so many hidden factors seem to impact a job's salary, it can be difficult to develop a clear idea of how salaries of different jobs compare. In addition, we were curious how salaries vary over demographic groups such as gender and education level. These analyses can help reveal potential salary inequity trends in the case of gender, and help advise and inform educational choices in the case of education level. Our overall goal is to develop a better understanding of the salary landscape of the current workforce.

The questions our group sought to answer include, first and primarily, which factors are the strongest predictors of salary. Secondly, as both age and years of experience are included in the data, we wanted to see how collinear these columns are. It makes logical sense that they would be, as you would not expect a person fresh out of college to have much experience. But, there could be other factors that cause them to not be related, such as changing jobs or fields, and thus being older with less experience. Additionally, we want to investigate how big of an effect education level has on salary, especially with added years of experience. This question was motivated by the idea that if having a Master's degree or PhD provides only a slight increase in salary over having a Bachelor's at the end of a career, then the degree may not be financially worthwhile.

Data Description

The data used for this project is the Salary_Data dataset from Kaggle, found here

The description of the data set does not say where the data is from geographically. However, the data set creator did respond to a comment asking what unit the salary data was in, saying that it was in Indian Rupees(INR). Given this, it is most likely that this data set is from India.

There are 6698 complete rows and six columns: age (in years), gender (male, female, or other), education level (Bachelor's, Master's, or PhD), job title, years of experience, and salary in USD. The numeric variables are age, years of experience, and salary, while gender, education level, and job title are categorical. While education level and gender have predetermined levels, the job title column is freeform; some job titles come up hundreds of times while others only appear once. Additionally, there are only 14 rows with gender as other, so there is relatively less data for that point.

Due to the job title column having a large number of categories, we will be grouping categories together to have a more reasonable number of categorical levels for our analysis. Given the arbitrary nature of this, it could potentially have an impact on the significance of our findings of if job title is a significant predictor of salary.

Data Processing

The dataset as sourced from Kaggle is generally clean and ready-to-use. However, some data processing was needed for the job title and education level columns.

Job Title Categorizing

For our analysis of the Salary Dataframe, we wanted to do linear regression of the salary based on the type of job someone was employed in. The first instinct would be to use the `Job.Title` column of our dataframe. The problem with this approach is that there are way too many job titles that were listed in the data there were 192 different job titles listed in the dataset, which would make any categorical regression nigh impossible.

So we decided to categorize our job titles into different job categories which correspond to different kinds of jobs. The 5 levels for our `Job.Category` variable are `STEM, BUSINESS, ADMIN, SERVICE, and OTHER`.

- `STEM` describes scientific and technical jobs, like computer programming, scientific research, and engineering.
- `BUSINESS` refers to any jobs involving the financial and marketing aspects of running businesses and companies.
- `ADMIN` refers to managerial and administrative jobs, involves project management, training, and human resource management.
- `SERVICE` refers to jobs the involve mainly interacting with customers and helping them in some way
- `OTHER` refers to jobs that don't fit any of the categories listed above

To sort the majority of job titles, we did a content analysis of all the words contained in all the job titles to see which words were recurring. These key words were then coded and sorted into 1 of 5 job categories. We then created a function to check if a job title contained a key word from a list, and if it did it was sorted into that job category, if the job title didn't contain any of the keyword it was sorted into `OTHER` if the job title contained keywords from 2 or more categories it was temporarily sorted into the `OVERLAP` category. These jobs with overlap were then manually sorted based on which job category fit the best.

The new variable can now be used in our linear regression analysis. One potential issue with the validity of this statistical analysis, is that the job categorizations we designated are somewhat arbitrary particularly in the case that we had overlap between job titles. This could potentially compromise the statistical power of any analysis of the job category variable.

Education Levels Normalization

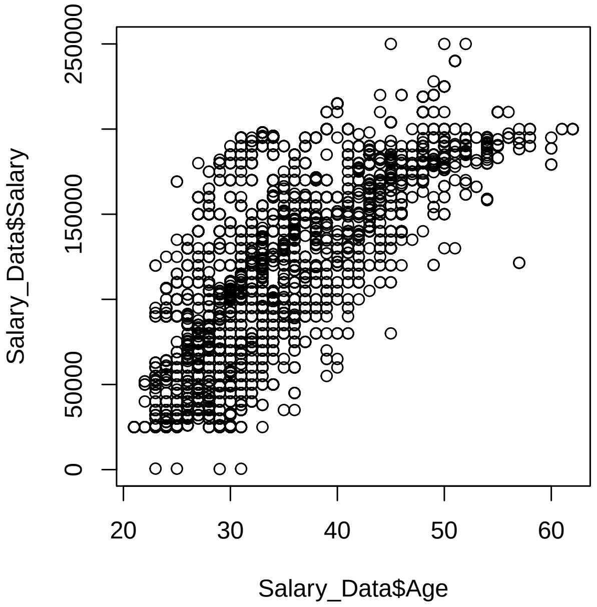
We also noticed later on that the Education.Level column had categories that should be put together, such as phD and PhD, and Bachelor's and Bachelor's Degree, where they denoted the same level of education. So, we fixed that to the levels of High School, Bachelor's, Master's, and PhD.

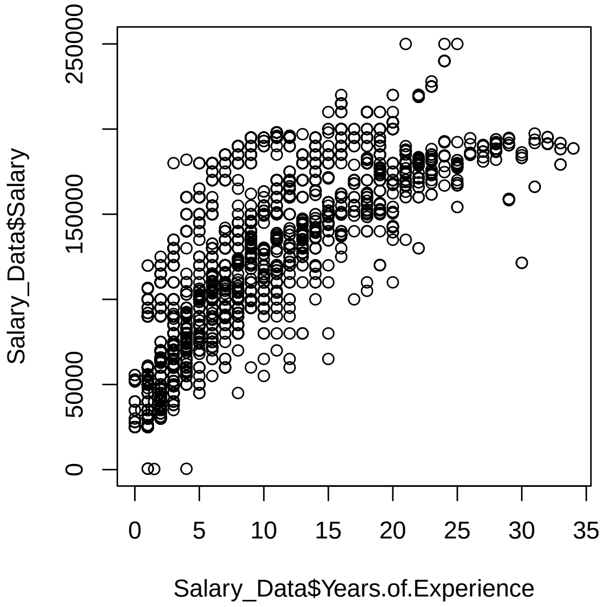
NAs

Finally, there were only 6 rows of our 6704 with N/A or missing values, two of these being completely empty rows, one missing any education value, and the remaining three with N/A values in the salary and years of experience columns. These 6 rows were filtered out of our processed dataset.

Linearity check

Before doing a linear model, we want to check that the predictors and response variables have a linear relationship. So, as age and experience are our numerical predictors, we will make a scatterplot of each of them compared to salary. Looking at the scaterplots, we see that they both appear to have a linear relationship, so a linear model does make sense in this case.



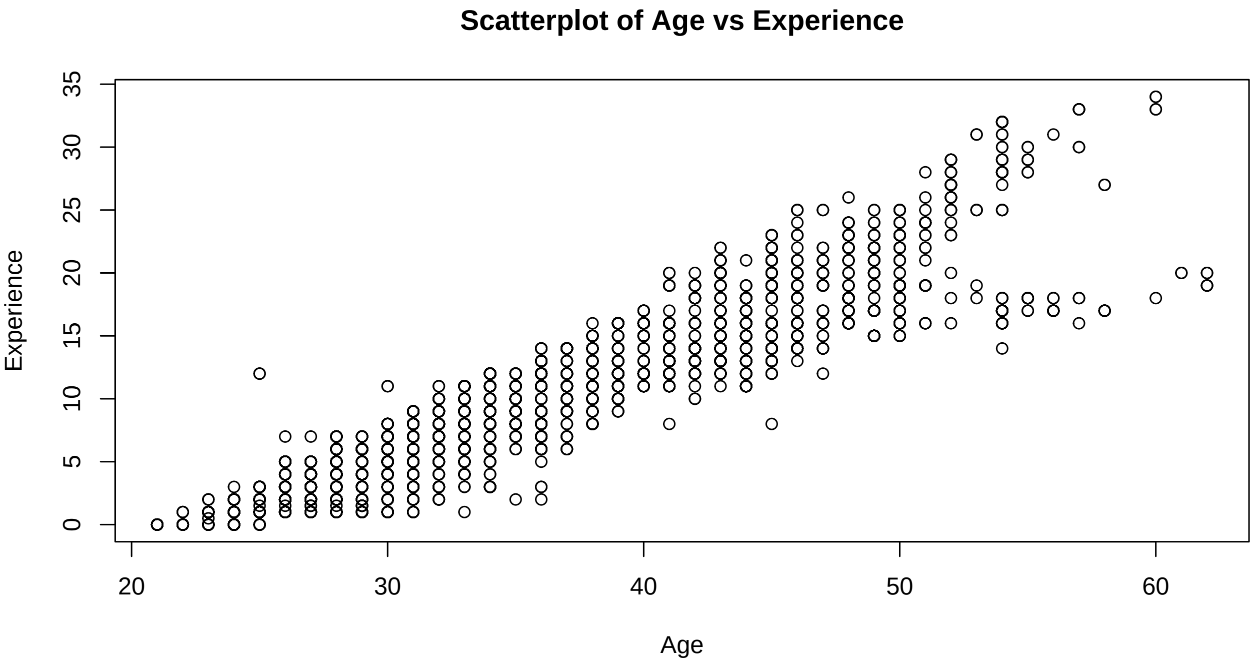


Starting Model

Collinearity of Age and Experience

We want to check if age and years of experience are collinear. It makes logical sense that they would be, as you would not expect a person fresh out of college to have much experience. And a person who is older could have been is a job for years. But, there could be other factors that cause them to not be related, such as changing jobs or fields, and thus being older with less experience.

We will use three methods to check the collinearity of age and experience. The first will be a scatterplot of the two variables. If they are unrelated we will see an even spread, and if they are related, there will be a more linear spread. We will also use two functions in R, to find two numbers that would hint towards collinearity, the correlation coefficient and the variance inflation factor. If the correlation coefficient is greater than 0.8 or the variance inflation factor is greater than 10, there is evidence of collinearity. The third test will be fitting a linear regression to just age and experience, with age as the predictor and experience as the response. If the two are collinear, then there will be statistically significant evidence of a linear relationship between the two variables from a t-test.



Looking at the scatterplot of age versus experience, we see that there appears to be a linear relationship between the two.

```
## [1] 0.9377253
```

Using the correlation function in R gives us a correlation value of 0.9377, which is greater than what we said was needed to consider a collinear relationship possible.

```
##                  Age Years.of.Experience
##             8.286979            8.286979
```

Creating a quick linear model of Salary predicted by Age and Experience, we can find the Variance Inflation Factor of the two predictors. We see that each of them have a VIF of greater than 8. This again shows us that there is a strong case for collinearity between the two factors.

```
##
## Call:
## lm(formula = Years.of.Experience ~ Age, data = AgeVsExp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2700  -1.1227  -0.1531   1.1159  10.3393
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.994278   0.116459  -145.9   <2e-16 ***
## Age           0.746199   0.003378   220.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.105 on 6696 degrees of freedom
## Multiple R-squared:  0.8793, Adjusted R-squared:  0.8793
## F-statistic: 4.879e+04 on 1 and 6696 DF,  p-value: < 2.2e-16
```

Using R to create a linear regression between Age and Years of Experience we see that the t-test gives a very low p-value, less than 0.001. This means that there is strong evidence of a linear relationship between the two variables, matching all of our previous tests. We see that if Age increases by one, years of experience will increase by roughly 0.746. Thus, as we have statistically significant evidence of Age and Years of Experience being collinear, we will not have both of them in our final model.

Checking for Interactions

To check for interactions, we fit a linear model including all predictors, and all of their interactions. From those, we selected only the interactions that were statistically significant at the 0.05 level of probability. This gives us the following interactions as significant. Education has multiple significant interactions with both age and experience. Education also has multiple significant interactions with job category. And, looking at rows 16 to 19, gender appears to have a significant interaction with education, and years of experience.

So, to check that these interactions are significant, we will use the AIC and BIC functions. We will start with a fit of salary based on gender, experience, job category and education. I am leaving out age, as we know it to be collinear with experience based on work done above. We will compare this fit to one that also includes the interactions of education with experience, job category, and gender individually; it will also include the interaction between gender and experience. If the AIC and BIC for the fit including interactions is lower, we will conclude that the interactions are significant and should be included in our final model. Looking at our results below, this does hold, so we will add the interactions to our final model.

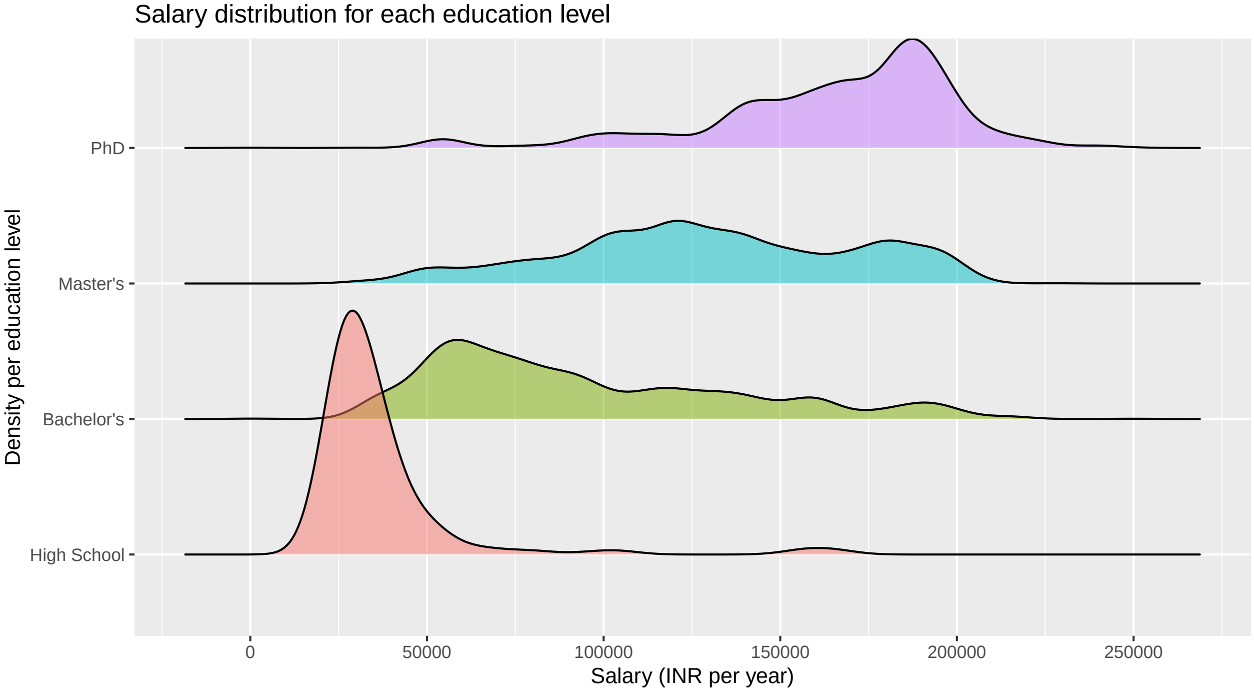
```
##                  df      AIC
## fit_basic        12 155763.2
## fit_interactions 31 155170.1
```

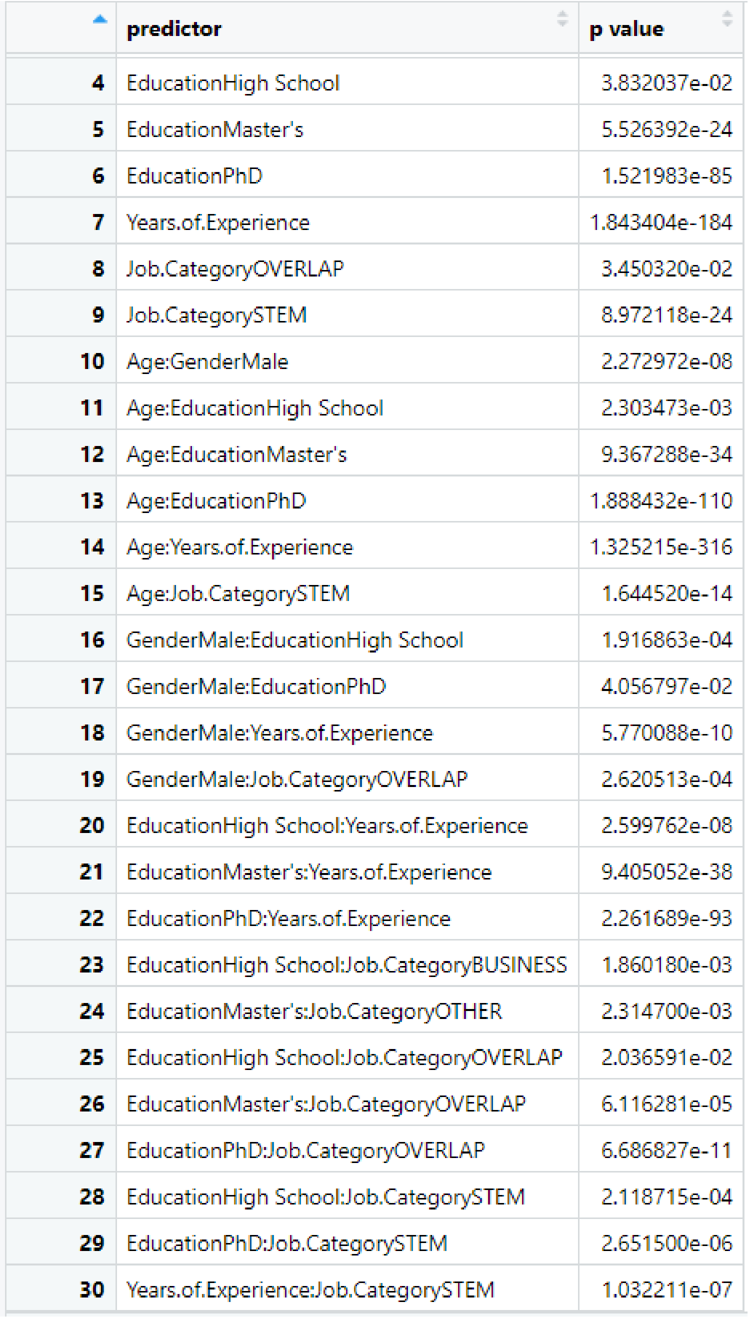
```
##                  df      BIC
## fit_basic        12 155844.9
## fit_interactions 31 155381.2
```

Education Level Impact

[rewrite analysis since it was deleted]

```
## Picking joint bandwidth of 6250
```



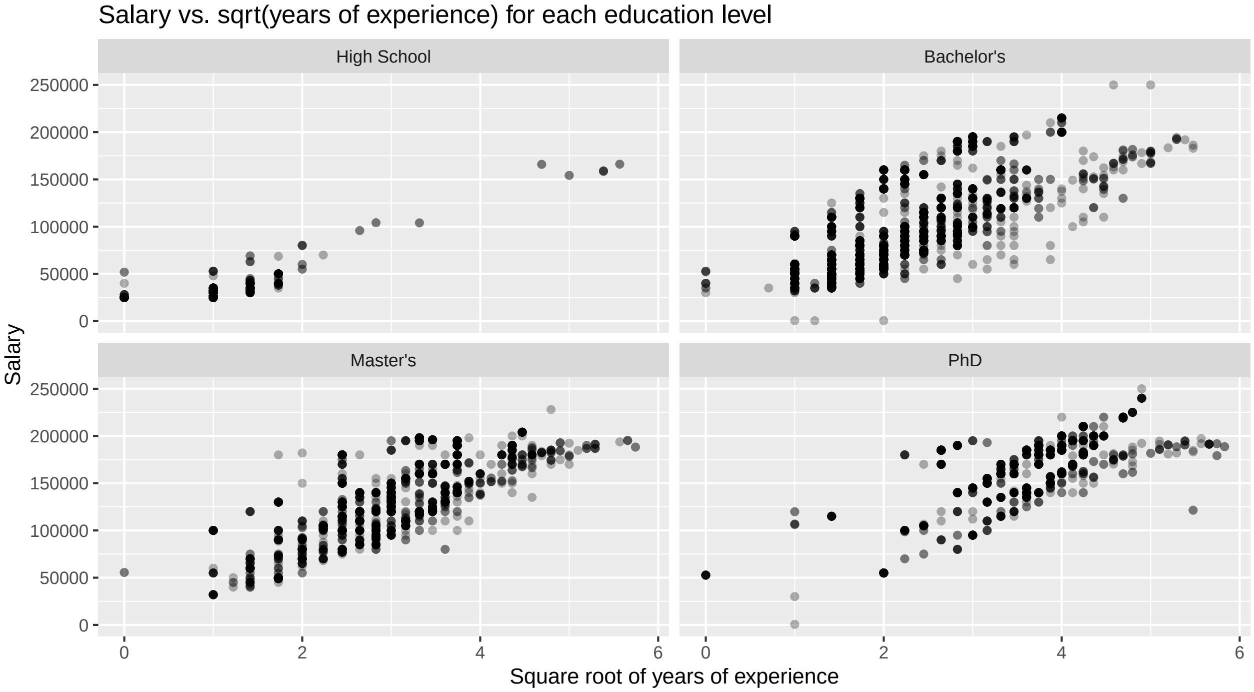


	predictor	p value
4	EducationHigh School	3.832037e-02
5	EducationMaster's	5.526392e-24
6	EducationPhD	1.521983e-85
7	Years.of.Experience	1.843404e-184
8	Job.CategoryOVERLAP	3.450320e-02
9	Job.CategorySTEM	8.972118e-24
10	Age:GenderMale	2.272972e-08
11	Age:EducationHigh School	2.303473e-03
12	Age:EducationMaster's	9.367288e-34
13	Age:EducationPhD	1.888432e-110
14	Age:Years.of.Experience	1.325215e-316
15	Age:Job.CategorySTEM	1.644520e-14
16	GenderMale:EducationHigh School	1.916863e-04
17	GenderMale:EducationPhD	4.056797e-02
18	GenderMale:Years.of.Experience	5.770088e-10
19	GenderMale:Job.CategoryOVERLAP	2.620513e-04
20	EducationHigh School:Years.of.Experience	2.599762e-08
21	EducationMaster's:Years.of.Experience	9.405052e-38
22	EducationPhD:Years.of.Experience	2.261689e-93
23	EducationHigh School:Job.CategoryBUSINESS	1.860180e-03
24	EducationMaster's:Job.CategoryOTHER	2.314700e-03
25	EducationHigh School:Job.CategoryOVERLAP	2.036591e-02
26	EducationMaster's:Job.CategoryOVERLAP	6.116281e-05
27	EducationPhD:Job.CategoryOVERLAP	6.686827e-11
28	EducationHigh School:Job.CategorySTEM	2.118715e-04
29	EducationPhD:Job.CategorySTEM	2.651500e-06
30	Years.of.Experience:Job.CategorySTEM	1.032211e-07

Figure 1: Screenshot since knitting freaked out

[rewrite analysis since it was deleted]

Our next step is to examine how salary changes with years of experience relative to education level. As our initial scatterplots showed a curved relationship between years of experience and salary, we have taken a square root transformation of years of experience which is plotted below.



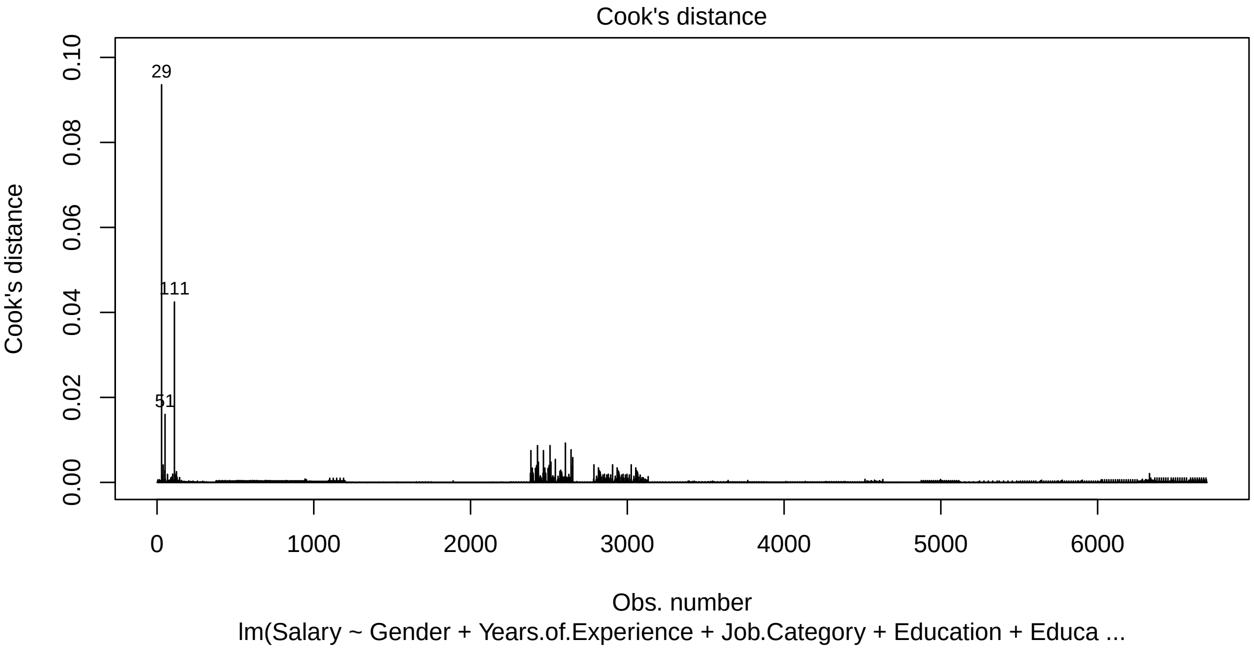
The scatterplot does not show any clear differences between the linear trends of salary vs. years of experience for each education group. To examine the relationship more closely we constructed a linear model using only square root of years of experience, education level, and their interaction as predictors of salary.

```
##
## Call:
## lm(formula = Salary ~ Education * sqrt_yoe, data = Salary_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -100995  -19518   -3514   13444   95023
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   11481.8     1346.4   8.528  < 2e-16 ***
## EducationHigh School           1441.1     2414.4   0.597  0.55061
## EducationMaster's             11028.7     2614.1   4.219 2.49e-05 ***
## EducationPhD                  40816.9     3743.5  10.903  < 2e-16 ***
## sqrt_yoe                      38535.5      578.4  66.628  < 2e-16 ***
## EducationHigh School:sqrt_yoe -16340.3     1559.4 -10.479  < 2e-16 ***
## EducationMaster's:sqrt_yoe     -2470.3      924.7  -2.672  0.00757 **
## EducationPhD:sqrt_yoe          -7471.2     1100.6  -6.788 1.23e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26840 on 6690 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7414
## F-statistic:  2744 on 7 and 6690 DF,  p-value: < 2.2e-16
```

In agreement with the analysis of interactions above, the interaction between education level and square root of years of experience is significant at the $\alpha = 0.01$ level for all education levels tested. That is, the change in salary for increased years of experience is significantly different for different education levels. In order of high school, Bachelor's, Master's, and PhD groups, after adjusting for the square root transformation the expected increase in salary per one year increase in experience is 22195, 38535, 36065, and 31064 INR. We can see that the high school group has nearly half the expected slope as the Bachelor's group, but Master's and PhD groups actually have smaller slopes. In other words, the benefit of a Bachelor's over a high school diploma is strongest considering both the median salary and increase in salary over years of experience. Due to the higher median salaries of the Master's and PhD groups, there are still benefits to these higher degrees but the increase in salary over time is less.

Outliers

To find any possible outliers in our data, we will find the Cook's Distance for each of the data points. If any data point has a Cook's distance greater than 0.5, we will look into removing it as it would potentially be an outlier. The model that we will use to calculate this is the interactions fit model, as seen in the Interactions section, as it is our most complete model at this point. Looking at the plot, we see that all but one of the points have a cook's distance of less than 0.05. The 111th data point has a relatively high cook's distance between 0.30 and 0.35. But, it is less than 0.5, so we do not have enough evidence to consider any points an outlier to be removed.



Final model