Project Proposal: Education Levels vs Salary

Hsiang-Hsuan Hung

October 31, 2015

Motivation

In this project proposal, I will discuss how education levels (EL) influence to employees' salary income. Many people are interested in this question, and are wondering, whether pursing a higher degree is useful for them to find a decent job and receive higher salary in their career. This is because pursing a higher degree, such as a doctoral degree, needs to spend more time, and we will take a risk without receiving any feedbacks (cannot graduate and receive degree) and waste time eventually. Here I will use the relevant data to answer the above questions and propose furture directions. The preliminary analysis suggests a positive relation, i.e. higher EL is associated with higher salary. However, there are other complicated reasons to influence the relation.

Data Source

The website "HigherEdJobs" (go to "https://www.higheredjobs.com/salary/") shows survey collections to reveal average income at various EL's and job positions. There are 1,227 participants in this survey, including 81% of all U.S. doctoral institutions, 52% of master's institutions and 34% of baccalaureate institutions. A total of 331 special-focus and associate's institutions also completed this year's survey. Public institutions comprised 48% (592) of respondents, and private institutions comprised 52% (635) of respondents. Therefore, the observation and conclusion are less relevant to whether the participants are educated from public, private schools, or even both.

In the survey collection, I am particularly interested in the item: "Professionals in Higher Education Salaries (Mid-Level Administrators)", which is relevant to my education background. Using R, we can retrieve the data from the website "https://www.higheredjobs.com/salary/salaryDisplay.cfm?SurveyID=33":

```r
library(XML)
library(bitops)
library(RCurl)
url <- "https://www.higheredjobs.com/salary/salaryDisplay.cfm?SurveyID=33"
tabs <- getURL(url)
df <- readHTMLTable(tabs, stringsAsFactors = F, header=TRUE)
## somehow the title name shows "Null", and now artifically give names
names(df) <- c("Academic Affairs", "Student Affairs", "Institutional Affairs", "Fiscal Affairs", "Exter
names(df)
```

```
##  [1] "Academic Affairs"
##  [2] "Student Affairs"
##  [3] "Institutional Affairs"
##  [4] "Fiscal Affairs"
##  [5] "External Affairs"
##  [6] "Facilities"
##  [7] "Information Technology"
##  [8] "Research Professionals"
##  [9] "Extension Programs, Technology Transfer, Health Science & Environmental Sustainability"
## [10] "Athletic Affairs"
## [11] "Exempt Office/Clerical, Skilled Craft, Services and Maintenance Personnel"
```

There are totally 11 groups in this category. In each group, we will have a data frame like

```
head(df[9])
```

```
## $`Extension Programs, Technology Transfer, Health Science & Environmental Sustainability`
##                               Job Title All Institutions Research
## 1             Head, Community Services          $63,372  $80,613
## 2  Senior Technology Licensing Officer         $100,666 $103,500
## 3                      Staff Physician         $148,722 $147,996
## 4                   Nurse Practitioner          $81,117  $84,218
## 5                          Staff Nurse          $52,115  $55,358
## 6              Clinical Research Nurse          $66,552  $66,708
## 7           Pharmacist, Student Health         $104,745 $106,000
## 8                         Veterinarian         $107,373 $107,373
## 9                  Animal Care Manager          $57,178  $65,436
## 10     Dietetic/Nutrition Professional          $53,954  $52,705
## 11  Head, Environmental Sustainability          $67,157  $79,000
##    Other Doc. Master's Baccalaureate Two-Year
## 1     $75,994  $56,821       $62,753  $63,391
## 2     $77,928        *             *        *
## 3    $150,709 $141,590      $193,494        *
## 4     $80,928  $78,062       $80,906  $81,399
## 5     $53,672  $49,754       $51,617  $55,557
## 6           *        *             *        *
## 7     $98,395  $99,432             *        *
## 8           *        *             *        *
## 9     $58,969  $48,152       $47,073        *
## 10    $57,367  $51,000       $56,903        *
## 11    $76,125  $57,543       $60,200  $65,000
```

The whole data “df” is the 2014-15 Professionals in Higher Education Salary Survey conducted by The College and University Professional Association for Human Resources. There are some missing data indicated as “∗” due to insufficient survey samples. For simplicity, here let us focus on observing the averaged salary over all job positions at various EL’s (without considering detail andgroups): “All Institutions” means the averaged salary for a fixed job without considering EL, so I am not going to use in the followings.

Our focus is on the relation between salary income vs EL, i.e. degree feature. “Research” and “Other Doc.” denote employees receiving PhD from a research university and others. “Master’s” and “Baccalaureate” means receiving the master and bachelor degrees relevant to the job positions. “Two-Year” means two-year college education without bachelor degree. There are totally 289 different positions (289 sample).

Data Analysis

First we need to reorganize the data for each EL, and then find out the statistics. The data in “df” are character-type with “$”" and “,” symbols, so we cannot directly use them. Define a function called “convert”:

```
convert <- function(data, index){
      nCagetory <- length(names(data))
      vectData <- c( )
      for (i in seq_along(1:nCagetory)){
            numberData <- data[[i]]
            vecArray <- data.frame(numberData[, index])
            vectData <- rbind(vectData,vecArray)
      }
      vect <- c( )
      for (i in seq_along(1:dim(vectData)[1])){
            string <- levels(droplevels(vectData[i,1]))
            string <- substr(string,2,nchar(string))
            vect[i] <- as.numeric(gsub(",", "", string, fixed = TRUE))
      }
      return(vect)
}
```

This function gives back a number data array (289 × 1) for each EL, i.e. removing the symbol “$” and “,”. For example, the salary list for all job types at “Research PhD”, “Other Doc.”, “Master”, “Bachelor” and “Two -Year” can be done by

```
resrhPhD <- convert(df,3)
otherPhD <- convert(df,4)
master   <- convert(df,5)
bachelor <- convert(df,6)
twoYear  <- convert(df,7)
```

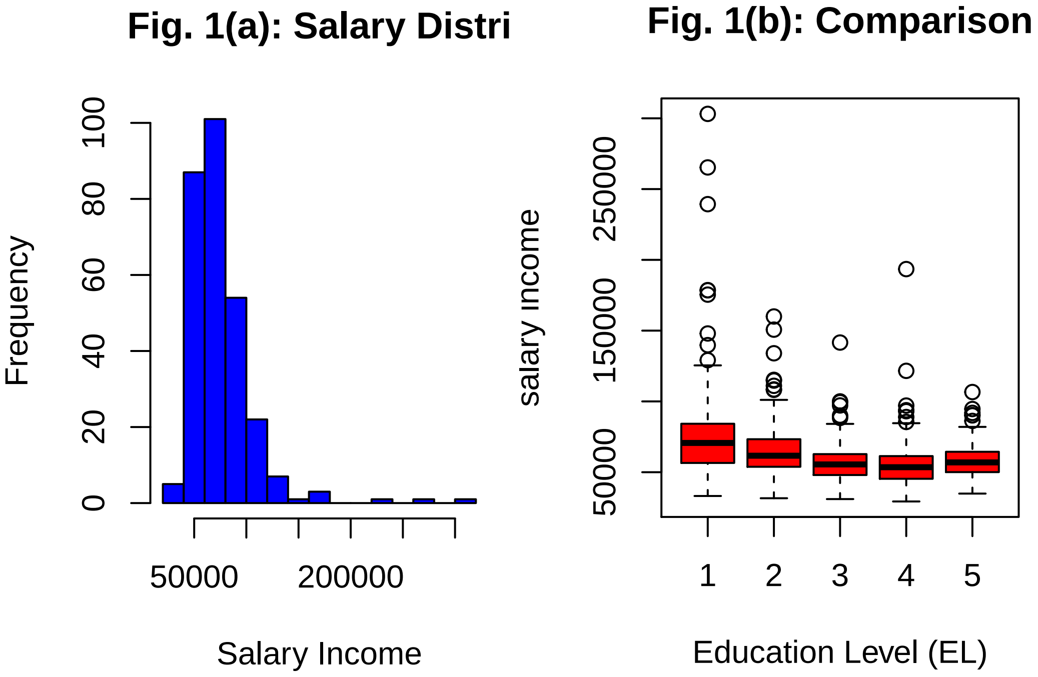
Next we need to examine the salary distribution for each EL. Note that irrespective of EL’s, the employees’ salary shows skewed distributions. As a representative example, in Fig. 1(a), the salary income shows a right-skewed distribution. The mean income is $75,308, the median income is $70,640, but the maximum can be even $303,100! Thus it is more important to look at the median and Q1, Q3 values using boxplot, and compare the salary income for each EL.

We further reconstruct the vector arrays for different degrees as a (289 × 5) × 2 data frame. The procedures are tedious, as follows:

```
degree <- rep(1,length(resrhPhD))
salaryData <- data.frame(resrhPhD,degree)
names(salaryData) <- c("salary","degree")

degree <- rep(2,length(otherPhD))
salary <- data.frame(otherPhD,degree)
names(salary) <- c("salary","degree")
salaryData <- rbind(salaryData,salary)

degree <- rep(3,length(master))
salary <- data.frame(master,degree)
names(salary) <- c("salary","degree")
salaryData <- rbind(salaryData,salary)

degree <- rep(4,length(bachelor))
salary <- data.frame(bachelor,degree)
names(salary) <- c("salary","degree")
salaryData <- rbind(salaryData,salary)

degree <- rep(5,length(twoYear))
salary <- data.frame(twoYear,degree)
names(salary) <- c("salary","degree")
salaryData <- rbind(salaryData,salary)
salaryData <- salaryData[is.na(salaryData$salary)==FALSE,]
head(salaryData,10)
```

```
##    salary degree
## 1   91824      1
## 2   51520      1
## 3   42735      1
## 4   51806      1
## 5   86219      1
## 6  100393      1
## 7   49339      1
## 8   42618      1
## 9   79488      1
## 10  77783      1
```

Now the "salaryData" data frame has the second column labeled as the EL: "1" means Research PhD, "2" other PhD, "3" master, "4" bachelor and "5" two-Year. The first column is the averaged salary for different positions from the raw data.



In Fig. 1(b), on average, we can observe that the employees who received PhD from research universities have higher salary than others, including master and bachelor degrees. Roughly speaking, a higher degree a higher averaged income (the median income is higher). We also can note, at higher EL, the distribution is more skewed. For EL=1, the mean income is $75,308, but the median income is $70,640. The maximum can be even $303,100. On the other hand, for EL=4 (bachelor degree), the mean income is $55,849, the median income is $53,520, and the maximum income is $193,500, showing a less skewed distribution compared to the "Research PhD" case.

```
summary(resrhPhD); mean(resrhPhD[!is.na(resrhPhD)])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   33200   56530   70650   75310   84230  303100       6
## [1] 75308.33
```

```
summary(bachelor); mean(bachelor[!is.na(bachelor)])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   29330   45340   53520   55850   61300  193500      47
```

```
## [1] 55849.07
```

From these analysis, pursing a higher degree for an employee is likely to get better pay and could be worthy to invest. (The problem is, however, that pursing a PhD may waste several years with risk, and during the time you can also make money.)

Another interesting observation found here is that an employee with the two-year college education even slightly makes more money than people who have the 4-year college (which earns a bachelor degree). Thus it is negotiable to attend a two-year or four-year college since you may pay more tuition for 4-year college.

Simple Regression Model

Next question we are interested is how much salary income an employee can earn more if he/she is wondering to stay in/go back schools to chase a higher degree? Whether we can construct a regression to predict the distribution behavior for next year or even few years later?

To quantitatively answer these questions, we implement the linear regression Machine Learning algorithms to train the data. First let us randomize the data frame “salaryData” and generate “randSalary”:

```
randSalary <- salaryData[sample(nrow(salaryData), dim(salaryData)[1]), ]
```

Then we can partition 70% of the original data as the training set to train, and the remaining 30% of the original data as the testing set to test our regression.

```
library(ggplot2)
library(lattice)
library(caret)
inTrain <- createDataPartition(y=randSalary$salary,p=0.7, list=FALSE)
training <- randSalary[inTrain,]
testing <- randSalary[-inTrain,]
```

Here we consider our problem in a relatively simple way, since we involve a predictor in the training only: $y_{\text{income}} = \alpha + \beta x_{\text{EL}}$, where y_{income} is the averaged salary income for a given job, and x_{EL} is the EL, $x_{\text{EL}} = 1, 2, 3, 4, 5$ representing “Research PhD”, “other PHD”, “Master”, “Bachelor” and “Two-Year”, separately. We use method=‘lm’, linear regression method, to train the dataset “training”.

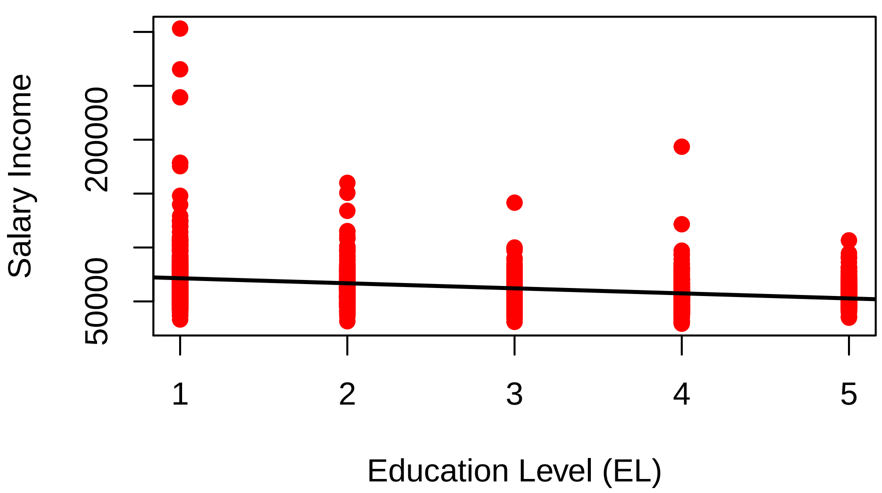
```
set.seed(123456)
modFit <- train(salary ~., data = training, method='lm')
modFit$finalModel
```

```
## 
## Call:
## lm(formula = .outcome ~ ., data = dat)
## 
## Coefficients:
## (Intercept)       degree  
##       75852        -4567
```

The above fitted outcome shows the intercept is $\alpha \sim 75000$ and the slop $\beta \sim -4500$ with big error bar. This can be roughly explained that, the expected salary of the employee, who earned a bachelor degree and a master degree, is possibly $4,500 more than that of who only earned a bachelor degree. And then if he/she continues to purse a master degree, one can earn $ β in pay, and for further PhD degree from some other institutions (meaning not from research universities) one can earn even $ 2β more.

We summarize the fitting results in Fig. 2. The red dots are the preliminary data and the black line is the regression line, described by the above equation. The negative slope indicates a message: higher education is possibly with higher salary. This is qualitatively consistent with our previous conclusion, and can be used to give a guide for people who are considering their education levels.

Fig. 2



Discussion and Future Plan

Although here I gave a simple model to study the relation between the salary income vs degree, there are a lot of flaws. One of them is that in the regression, the fitting result shows that even at $x_{\mathrm{EL}} = 0$, the intercept is still $75,000. Also, there should be no difference between "Two-Year" and "Bacc..". These show that the model only considering EL to explain the salary income is too simple. In the testing set, the accuracy of the predicted income using the model compared to the observed values (provided by the test data) is actually poor. Therefore, the model only captures the qualitative behavior and fails to predict the relation quantitatively.

The regression with the simple assumption, which salary income is only relevant to EL, is insufficient, and we still need to explore more external features to study the relation between the EL and the salary income. Other possible reasons include the school reputation, the employees' locations, further considering job categories and so on. Furthermore, the regression can be even year-dependent, such as the year-by-year variation of the salary vs EL. Then it becomes a dynamical problem.

Overall, the relation between the income and education background is a complicated problem, and such investigation needs to collect more data from various sources, and introduce more possible features. I believe that many people are interested in this topic, and would like to know further. This is my project proposal in the Data Incubator Fellowship Program. I will be excited to join your team, participate your program, and wish to study such interesting problems or others more deeply with other excellent data scientists.