CIS 585 Project

Kodai Watanabe

05/05/2022

## Abstract

In this project, I try to find an answer for a research question, How does years of experience of developers affect their compensation. Figuring out the answer is beneficial for HR people and job seekers. It is expected that HR people can use the finding to provide a better estimate of compensation and the job seekers can know how much they should expect as salaries based on their experience.

It is argued that more experience lead to higher compensation in general. In this research, I examine the relationship in the context of developers. To implement the research, I get data from a survey conducted by Stack Overflow in 2021 and did a regression analysis. I find that for each one percent increase in the professional experience as a developer, there is an increase of 0.181% in the compensation amount. In the regression analysis, the length of mere coding experience does not significantly affect the compensation from a standpoint of statistics.

However, this relationship between the professional experience and the compensation is applied to only developers who make less than a compensation amount of 12 in the logarithmic scale. This limitation arises because of non-linearity in the data.

Based on the findings, I would recommend HR people to focus on the professional experience of coding rather than mere coding experience when estimating the compensation. More professional experience causes the increase of the compensation, so failing to pay enough money for certain professional experience lead to not be able to hire a good candidate.

For job seekers, I would encourage to gain internships to gain professional experience.

## Introduction

This research project tries to answer a question, how does years of experience of developers affect their salaries in the U.S.? Years of experiences are generally considered one of factors that determine the amount of salary. This is because, the longer you do a job, the more productive you become (Torpey, 2015). It is mentioned by an article that it is typical to get paid more if you have more experience (Salary.com, n.d.). In job postings of developers, some years of experience, such as 5 years, is often seen as a requirement. This research project will test how significant the factor is for the determination of salary of developers in the U.S.

This research project is important for two reasons. One is that the findings can give people in HR an insight about how impactful years of experience are on a figure of salary. They could provide more accurate estimate of salary based on the findings. Second is that the results of the project could help job seekers and students to determine whether their years of experience are valid against salaries shown by employers.

## Literature Review

You get paid more as you gain more experience even though this is observed in years of experiences up to 25 years; the base salary in the U.S. increases by $1500 for one-year increase in years of experience on average (Sauro, 2014). In a study focusing on the Indian IT industry, it is suggested that salary is positively related with work experience (Dash et al., 2017).

Other factors also affect the amount of salary. For instance, education is one of them. Workers who possess higher degree, professional certification, or license might get paid more than others even though it is about the same industry (Torpey, 2015).

## Theory

The years of experiences as developers increases the amount of their compensation, but the slope of the growth get flattened.

## Data

I get data from a survey conducted by Stack Overflow in 2021.

<https://insights.stackoverflow.com/survey>

## Methodology

From this section, I show all of what I did in this project to arrrive the conclusion I discussed earlier.

Install Packages

install.packages("rmarkdown")  
install.packages("knitr")  
install.packages("ggplot2")  
install.packages("plyr")  
install.packages("dplyr")  
install.packages("stringr")  
install.packages("caret")  
install.packages("car")  
install.packages("lmtest")  
install.packages("sandwich")  
install.packages("magrittr")  
install.packages("stargazer")  
install.packages("leaps")

Load packages

library("rmarkdown")  
library("knitr")  
library("ggplot2")  
library("plyr")  
library("dplyr")

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library("stringr")  
library("car")

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library("lmtest")

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library("sandwich")  
library("magrittr")  
library("stargazer")

##   
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library("caret")

## Loading required package: lattice

library("leaps")

## Read data

I download a dataset from Stack Overflow 2021 survey.

dataset <- read.csv('Datasets/survey\_results\_public.csv')

See what the data looks like. First, I check data types of columns. Some data types need to be transformed. For instance, Response ID should have character data type.

str(dataset)

## 'data.frame': 83439 obs. of 48 variables:  
## $ ResponseId : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ MainBranch : chr "I am a developer by profession" "I am a student who is learning to code" "I am not primarily a developer, but I write code sometimes as part of my work" "I am a developer by profession" ...  
## $ Employment : chr "Independent contractor, freelancer, or self-employed" "Student, full-time" "Student, full-time" "Employed full-time" ...  
## $ Country : chr "Slovakia" "Netherlands" "Russian Federation" "Austria" ...  
## $ US\_State : chr NA NA NA NA ...  
## $ UK\_Country : chr NA NA NA NA ...  
## $ EdLevel : chr "Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)" "Bachelorâ\200\231s degree (B.A., B.S., B.Eng., etc.)" "Bachelorâ\200\231s degree (B.A., B.S., B.Eng., etc.)" "Masterâ\200\231s degree (M.A., M.S., M.Eng., MBA, etc.)" ...  
## $ Age1stCode : chr "18 - 24 years" "11 - 17 years" "11 - 17 years" "11 - 17 years" ...  
## $ LearnCode : chr "Coding Bootcamp;Other online resources (ex: videos, blogs, etc)" "Other online resources (ex: videos, blogs, etc);School" "Other online resources (ex: videos, blogs, etc);Online Forum" NA ...  
## $ YearsCode : chr NA "7" NA NA ...  
## $ YearsCodePro : chr NA NA NA NA ...  
## $ DevType : chr "Developer, mobile" NA NA "Developer, front-end" ...  
## $ OrgSize : chr "20 to 99 employees" NA NA "100 to 499 employees" ...  
## $ Currency : chr "EUR European Euro" NA NA "EUR European Euro" ...  
## $ CompTotal : num 4800 NA NA NA NA NA NA NA NA 42000 ...  
## $ CompFreq : chr "Monthly" NA NA "Monthly" ...  
## $ LanguageHaveWorkedWith : chr "C++;HTML/CSS;JavaScript;Objective-C;PHP;Swift" "JavaScript;Python" "Assembly;C;Python;R;Rust" "JavaScript;TypeScript" ...  
## $ LanguageWantToWorkWith : chr "Swift" NA "Julia;Python;Rust" "JavaScript;TypeScript" ...  
## $ DatabaseHaveWorkedWith : chr "PostgreSQL;SQLite" "PostgreSQL" "SQLite" NA ...  
## $ DatabaseWantToWorkWith : chr "SQLite" NA "SQLite" NA ...  
## $ PlatformHaveWorkedWith : chr NA NA "Heroku" NA ...  
## $ PlatformWantToWorkWith : chr NA NA NA NA ...  
## $ WebframeHaveWorkedWith : chr "Laravel;Symfony" "Angular;Flask;Vue.js" "Flask" "Angular;jQuery" ...  
## $ WebframeWantToWorkWith : chr NA NA "Flask" "Angular;jQuery" ...  
## $ MiscTechHaveWorkedWith : chr NA "Cordova" "NumPy;Pandas;TensorFlow;Torch/PyTorch" NA ...  
## $ MiscTechWantToWorkWith : chr NA NA "Keras;NumPy;Pandas;TensorFlow;Torch/PyTorch" NA ...  
## $ ToolsTechHaveWorkedWith : chr NA "Docker;Git;Yarn" NA NA ...  
## $ ToolsTechWantToWorkWith : chr NA "Git" NA NA ...  
## $ NEWCollabToolsHaveWorkedWith: chr "PHPStorm;Xcode" "Android Studio;IntelliJ;Notepad++;PyCharm" "IPython/Jupyter;PyCharm;RStudio;Sublime Text;Visual Studio Code" NA ...  
## $ NEWCollabToolsWantToWorkWith: chr "Atom;Xcode" NA "IPython/Jupyter;RStudio;Sublime Text;Visual Studio Code" NA ...  
## $ OpSys : chr "MacOS" "Windows" "MacOS" "Windows" ...  
## $ NEWStuck : chr "Call a coworker or friend;Visit Stack Overflow;Go for a walk or other physical activity;Google it" "Visit Stack Overflow;Google it" "Visit Stack Overflow;Google it;Watch help / tutorial videos;Do other work and come back later" "Call a coworker or friend;Visit Stack Overflow;Go for a walk or other physical activity;Google it" ...  
## $ NEWSOSites : chr "Stack Overflow" "Stack Overflow" "Stack Overflow;Stack Exchange" "Stack Overflow" ...  
## $ SOVisitFreq : chr "Multiple times per day" "Daily or almost daily" "Multiple times per day" "Daily or almost daily" ...  
## $ SOAccount : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ SOPartFreq : chr "A few times per month or weekly" "Daily or almost daily" "Multiple times per day" "Daily or almost daily" ...  
## $ SOComm : chr "Yes, definitely" "Yes, definitely" "Yes, definitely" "Neutral" ...  
## $ NEWOtherComms : chr "No" "No" "Yes" "No" ...  
## $ Age : chr "25-34 years old" "18-24 years old" "18-24 years old" "35-44 years old" ...  
## $ Gender : chr "Man" "Man" "Man" "Man" ...  
## $ Trans : chr "No" "No" "No" "No" ...  
## $ Sexuality : chr "Straight / Heterosexual" "Straight / Heterosexual" "Prefer not to say" "Straight / Heterosexual" ...  
## $ Ethnicity : chr "White or of European descent" "White or of European descent" "Prefer not to say" "White or of European descent" ...  
## $ Accessibility : chr "None of the above" "None of the above" "None of the above" "I am deaf / hard of hearing" ...  
## $ MentalHealth : chr "None of the above" "None of the above" "None of the above" NA ...  
## $ SurveyLength : chr "Appropriate in length" "Appropriate in length" "Appropriate in length" "Appropriate in length" ...  
## $ SurveyEase : chr "Easy" "Easy" "Easy" "Neither easy nor difficult" ...  
## $ ConvertedCompYearly : int 62268 NA NA NA NA NA NA NA NA 51552 ...

Count null values for each of the columns

null\_counts <- sapply(dataset, (function(x) sum(is.na(x))))  
null\_counts

## ResponseId MainBranch   
## 0 0   
## Employment Country   
## 116 0   
## US\_State UK\_Country   
## 68519 79021   
## EdLevel Age1stCode   
## 313 196   
## LearnCode YearsCode   
## 476 1798   
## YearsCodePro DevType   
## 22223 16955   
## OrgSize Currency   
## 22713 22359   
## CompTotal CompFreq   
## 36256 31289   
## LanguageHaveWorkedWith LanguageWantToWorkWith   
## 1082 6618   
## DatabaseHaveWorkedWith DatabaseWantToWorkWith   
## 13893 25140   
## PlatformHaveWorkedWith PlatformWantToWorkWith   
## 31304 41820   
## WebframeHaveWorkedWith WebframeWantToWorkWith   
## 21732 31344   
## MiscTechHaveWorkedWith MiscTechWantToWorkWith   
## 36384 45418   
## ToolsTechHaveWorkedWith ToolsTechWantToWorkWith   
## 10902 17959   
## NEWCollabToolsHaveWorkedWith NEWCollabToolsWantToWorkWith   
## 2205 10417   
## OpSys NEWStuck   
## 145 387   
## NEWSOSites SOVisitFreq   
## 268 1026   
## SOAccount SOPartFreq   
## 914 15886   
## SOComm NEWOtherComms   
## 1120 611   
## Age Gender   
## 1032 1153   
## Trans Sexuality   
## 2761 10073   
## Ethnicity Accessibility   
## 3975 5836   
## MentalHealth SurveyLength   
## 6519 1728   
## SurveyEase ConvertedCompYearly   
## 1491 36595

Get respondents who are developers by profession and live in the U.S.because they are what I am interested in.

dataset <-subset(dataset, MainBranch=='I am a developer by profession' & Country=='United States of America')

## Data Cleaning

From here, I start cleaning the data.

First, I drop definitely unnecessary columns. I drop columns related to this survey. They should not be correlated with the compensations of the respondents.

dataset <- subset(dataset, select=-c(SurveyLength,SurveyEase,UK\_Country))

Change the data type of ID column. To prevent from applying a calculation on the column, I convert it into character data type.

dataset$ResponseId <- as.character(dataset$ResponseId)

Looking at YearsCode column, there are two string entries, less than 1 year and more than 50 years. They occupy only 1.2% of all respondents. Thus, they can be considered outliers. So, I drop rows containing the entries. Then, I change the column’s data type into numeric.

unique(dataset$YearsCode)

## [1] "8" "20" "30"   
## [4] "25" "18" "10"   
## [7] "12" "7" "27"   
## [10] "38" "41" NA   
## [13] "37" "24" "22"   
## [16] "21" "6" "14"   
## [19] "15" "4" "40"   
## [22] "9" "5" "39"   
## [25] "3" "45" "13"   
## [28] "33" "43" "35"   
## [31] "16" "23" "42"   
## [34] "46" "1" "17"   
## [37] "32" "19" "11"   
## [40] "28" "50" "26"   
## [43] "47" "34" "49"   
## [46] "2" "More than 50 years" "36"   
## [49] "31" "48" "44"   
## [52] "29" "Less than 1 year"

dataset <- dataset[!dataset$YearsCode %in% c("More than 50 years","Less than 1 year"),]  
dataset$YearsCode <- as.numeric(dataset$YearsCode)

Same logic is applied to YearsCodePro column.

unique(dataset$YearsCodePro)

## [1] "Less than 1 year" "15" "30" "1"   
## [5] "5" "6" "16" "34"   
## [9] "24" "9" "22" NA   
## [13] "13" "31" "12" "10"   
## [17] "4" "20" "25" "3"   
## [21] "35" "2" "19" "21"   
## [25] "8" "17" "11" "33"   
## [29] "7" "37" "43" "26"   
## [33] "23" "14" "42" "28"   
## [37] "18" "40" "32" "36"   
## [41] "38" "29" "27" "44"   
## [45] "45" "39" "41" "47"   
## [49] "50" "49" "46" "48"

dataset <- dataset[!dataset$YearsCodePro == "Less than 1 year",]  
dataset$YearsCodePro <- as.numeric(dataset$YearsCodePro)

Let’s look at OrgSize column. One of the unique values is a very long string. Let’s rename it. I call it Other.

unique(dataset$OrgSize)

## [1] "1,000 to 4,999 employees"   
## [2] "2 to 9 employees"   
## [3] "100 to 499 employees"   
## [4] "20 to 99 employees"   
## [5] "Just me - I am a freelancer, sole proprietor, etc."  
## [6] "10,000 or more employees"   
## [7] NA   
## [8] "500 to 999 employees"   
## [9] "I donâ\200\231t know"   
## [10] "5,000 to 9,999 employees"   
## [11] "10 to 19 employees"

dataset$OrgSize <- mapvalues(dataset$OrgSize,from=c("Just me - I am a freelancer, sole proprietor, etc."),  
 to=c("Other"))

About EdLevel, most of values of this variable are very long descriptions. Let’s rename them.

unique(dataset$EdLevel)

## [1] "Some college/university study without earning a degree"   
## [2] "Masterâ\200\231s degree (M.A., M.S., M.Eng., MBA, etc.)"   
## [3] "Bachelorâ\200\231s degree (B.A., B.S., B.Eng., etc.)"   
## [4] "Other doctoral degree (Ph.D., Ed.D., etc.)"   
## [5] "Associate degree (A.A., A.S., etc.)"   
## [6] NA   
## [7] "Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)"  
## [8] "Primary/elementary school"   
## [9] "Professional degree (JD, MD, etc.)"   
## [10] "Something else"

from <- c("Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)","Other doctoral degree (Ph.D., Ed.D., etc.)" ,  
 "Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)", "Associate degree (A.A., A.S., etc.)",  
 "Secondary school (e.g. American high school, German Realschule or Gymnasium, etc.)",  
 "Some college/university study without earning a degree" ,  
 "Primary/elementary school",  
 "Professional degree (JD, MD, etc.)",  
 "Something else" )  
to <- c("Bachelor","Doctorate","Master","Associate","Secondary School","College study wihtout degree","Elementary school",  
 "Professional degree","Something else")  
dataset$EdLevel <- mapvalues(dataset$EdLevel,from,to)

Regarding Employement, this variable also contains a very long string. Let’s rename it.

unique(dataset$Employment)

## [1] "Employed full-time"   
## [2] "Independent contractor, freelancer, or self-employed"  
## [3] NA   
## [4] "Employed part-time"   
## [5] "Retired"   
## [6] "I prefer not to say"

dataset$Employment <- mapvalues(dataset$Employment,from=c("Independent contractor, freelancer, or self-employed"),to=c("Other"))

There are columns regarding tools developers want to work with and have worked with. I drop columns that show tools they want to work with. It is hard to imagine that tools they want to use are related to their current compensation. On the other hand, columns about tools they have worked with might be useful. They could describe how experienced the developers are. I deal with this type of columns in the later section.

wanto\_columns <- grepl(pattern = "WantToWorkWith", x=names(dataset))  
dataset <- dataset[,!wanto\_columns]

There are columns about Stack Overflow.The usage of the site has nothing to do with the compensation amount, so let’s drop the columns.

dataset<- subset(dataset,select=-c(NEWSOSites,SOVisitFreq,SOAccount,SOPartFreq,SOComm))  
dataset<- subset(dataset,select=-c(NEWOtherComms))  
dataset<- subset(dataset, select=-c(NEWStuck))

Drop a column which tell what kinds of OS they use. It is irrelevant to the research question.

dataset <- subset(dataset, select=-c(OpSys))

Let’s go back to the worked with columns. Those columns contain tools’ names they have used. I convert the list of names into the number of tool. s First, let’s define a function to calculate the number of tools.

calculate\_elements <- function(elements)  
{  
 if(any(is.na(elements))){  
 return (0)  
 }  
 else{  
 return (length(elements))  
 }  
   
}

Then, apply the function to the worked with columns.

columns <- names(dataset)[grep(pattern = "HaveWorkedWith",x=names(dataset))]  
  
  
for(col\_name in columns)  
{  
 dataset[,col\_name] <- rapply(strsplit(dataset[,col\_name],";"),calculate\_elements)  
  
}

Drop rows that contain NA values

dataset <- na.omit(dataset)

I do not believe that how people learn code is crucial in deciding the compensation amount. So, let’s drop a column about how they learned coding.

dataset <- subset(dataset, select=-c(LearnCode))

Types of developers could affect the amount of compensation. For instance, Data Scientists could earn more than front-end developers. In this survey, one respondent can choose more than one type of developers. Thus, for each developer type available in this survey, let’s create a binary column which tells a person belongs to the type.

First, get all dev types available in this survey

all\_types =c()  
for (types in dataset$DevType){  
 for(type in strsplit(types,";")){  
 for(content in type){  
 if(!(as.character(content) %in% all\_types)){  
 all\_types <-append(all\_types,as.character(content))  
 }  
 }  
 }  
}

Then, create the columns in the original data set.

dev\_types <- setNames(data.frame(matrix(0,ncol = length(all\_types), nrow = nrow(dataset))), all\_types)  
dataset<- cbind(dataset,dev\_types)  
  
for(id in dataset$ResponseId){  
 job\_desc <- dataset[dataset$ResponseId==id,"DevType"]  
 description <- strsplit(job\_desc,";")  
 for(outer\_list in description)  
 {  
 for(inner\_list in outer\_list)  
 {  
 if (inner\_list == "Other (please specify):")  
 {  
 dataset[dataset$ResponseId==id,"Other (please specify):"] <- 1  
 } else{  
 dataset[dataset$ResponseId==id,inner\_list] <- 1  
 }  
 }  
 }  
}

This is what the columns look like.

head(dataset[c('Developer, full-stack','Engineer, data','Data scientist or machine learning specialist')])

## Developer, full-stack Engineer, data  
## 38 0 0  
## 76 1 0  
## 88 0 1  
## 119 1 0  
## 129 1 0  
## 144 0 0  
## Data scientist or machine learning specialist  
## 38 0  
## 76 0  
## 88 1  
## 119 0  
## 129 0  
## 144 0

Clean names of columns

colnames(dataset) <- gsub(" ","\_",colnames(dataset))  
colnames(dataset) <- gsub(",","",colnames(dataset))  
colnames(dataset) <- gsub("-","\_",colnames(dataset))  
colnames(dataset) <- gsub("/","",colnames(dataset))

Rename some of column names.

dataset$Other\_devtypes <- dataset$`Other\_(please\_specify):`  
dataset$Senior\_excecutives <- dataset$`Senior\_Executive\_(C\_Suite\_VP\_etc.)`  
  
dataset <- subset(dataset, select=-c(`Other\_(please\_specify):`,`Senior\_Executive\_(C\_Suite\_VP\_etc.)`))

Drop the original dev types column

dataset <- subset(dataset, select=-c(DevType))

There is a sexuality column. I do not think that a sexual preference significantly affects the amount of salary. Thus, I will drop the column.

dataset <- subset(dataset,select =-c(Sexuality))

## Looking at each variable

### Gender

Some respondents choose more than one gender categories. I think that it will be too detailed if I include all of these information. Since majority of respondents is male and the second biggest proportion is female. I will keep only male and female respondents.

dataset %>%group\_by(Gender) %>%summarise(cnt = n())%>%mutate(percentage=(cnt/sum(cnt))\*100)%>%head()

## # A tibble: 6 x 3  
## Gender cnt percentage  
## <chr> <int> <dbl>  
## 1 Man 6857 89.7   
## 2 Man;Non-binary, genderqueer, or gender non-conforming 47 0.615   
## 3 Man;Or, in your own words: 19 0.249   
## 4 Man;Woman 3 0.0393  
## 5 Man;Woman;Non-binary, genderqueer, or gender non-conforming 2 0.0262  
## 6 Man;Woman;Non-binary, genderqueer, or gender non-conforming;~ 1 0.0131

dataset<- dataset[(dataset$Gender=="Man") | (dataset$Gender=="Woman"),]

### Ethnicity

This Ethnicity categorical column is also very imbalanced. Most of values are white. Proportions of other unique values spread out. The problem is that the distinction is too detailed. Moreover, some of respondents have more than one Ethnicity even though there is a option called multiracial. Since I don’t know how to deal with multiple entries of ethnicity for one respondent and most of respondents have one ethnicity type, I will pick only rows that contain one value for this variable.

unique(dataset$Ethnicity)[1:10]

## [1] "White or of European descent"   
## [2] "South Asian"   
## [3] "White or of European descent;Hispanic or Latino/a/x"   
## [4] "White or of European descent;Or, in your own words:"   
## [5] "East Asian"   
## [6] "White or of European descent;Hispanic or Latino/a/x;Indigenous (such as Native American, Pacific Islander, or Indigenous Australian)"  
## [7] "White or of European descent;Middle Eastern"   
## [8] "Hispanic or Latino/a/x"   
## [9] "Multiracial"   
## [10] "I don't know"

dataset %>%group\_by(Ethnicity) %>%summarise(cnt = n())%>%mutate(percentage=(cnt/sum(cnt))\*100)%>%arrange(desc(percentage))%>%head(20)

## # A tibble: 20 x 3  
## Ethnicity cnt percentage  
## <chr> <int> <dbl>  
## 1 White or of European descent 5669 77.5   
## 2 Hispanic or Latino/a/x 221 3.02   
## 3 South Asian 217 2.97   
## 4 East Asian 172 2.35   
## 5 Prefer not to say 123 1.68   
## 6 White or of European descent;Hispanic or Latino/a/x 118 1.61   
## 7 Black or of African descent 112 1.53   
## 8 Southeast Asian 78 1.07   
## 9 White or of European descent;Indigenous (such as Native Ame~ 67 0.916  
## 10 Or, in your own words: 61 0.834  
## 11 White or of European descent;Middle Eastern 42 0.574  
## 12 Middle Eastern 40 0.547  
## 13 Multiracial 32 0.438  
## 14 White or of European descent;Or, in your own words: 27 0.369  
## 15 White or of European descent;East Asian 23 0.315  
## 16 White or of European descent;Hispanic or Latino/a/x;Biracial 22 0.301  
## 17 I don't know 18 0.246  
## 18 White or of European descent;East Asian;Biracial 16 0.219  
## 19 Biracial 15 0.205  
## 20 White or of European descent;Biracial;Black or of African d~ 11 0.150

dataset <- dataset[(rapply(strsplit(dataset$Ethnicity,";"),length)==1),]

Now the variable looks like this

unique(dataset$Ethnicity)

## [1] "White or of European descent"   
## [2] "South Asian"   
## [3] "East Asian"   
## [4] "Hispanic or Latino/a/x"   
## [5] "Multiracial"   
## [6] "I don't know"   
## [7] "Prefer not to say"   
## [8] "Or, in your own words:"   
## [9] "Middle Eastern"   
## [10] "Southeast Asian"   
## [11] "Biracial"   
## [12] "Black or of African descent"   
## [13] "Indigenous (such as Native American, Pacific Islander, or Indigenous Australian)"

Since the entry of indigenous is too long, let’s make it shorter.

dataset$Ethnicity <- mapvalues(dataset$Ethnicity, from=c("Indigenous (such as Native American, Pacific Islander, or Indigenous Australian)"), to=c("Indigenous"))

### Accessibility

This variable is too detailed as well. Moreover, since respondents in this dataset have jobs, their disabilities are not so severe that they cannot work as developers if they have. However, the disabilities might affect efficiency of their work. If I make dummy variable for each of disabilities, it will be too detailed. So, let’s make this column into a binary column that indicates whether people have a disability or not.

unique(dataset$Accessibility)[1:10]

## [1] "None of the above"   
## [2] "I am unable to / find it difficult to walk or stand without assistance"   
## [3] "Prefer not to say"   
## [4] "I am blind / have difficulty seeing"   
## [5] "Or, in your own words:"   
## [6] "I am deaf / hard of hearing"   
## [7] "I am deaf / hard of hearing;Or, in your own words:"   
## [8] "I am blind / have difficulty seeing;Or, in your own words:"   
## [9] "I am blind / have difficulty seeing;I am unable to / find it difficult to type"  
## [10] "I am unable to / find it difficult to type"

dataset %>%group\_by(Accessibility) %>%summarise(cnt = n())%>%mutate(percentage=(cnt/sum(cnt))\*100)

## # A tibble: 16 x 3  
## Accessibility cnt percentage  
## <chr> <int> <dbl>  
## 1 I am blind / have difficulty seeing 43 0.636   
## 2 I am blind / have difficulty seeing;I am unable to / find i~ 1 0.0148  
## 3 I am blind / have difficulty seeing;I am unable to / find i~ 1 0.0148  
## 4 I am blind / have difficulty seeing;Or, in your own words: 1 0.0148  
## 5 I am deaf / hard of hearing 53 0.784   
## 6 I am deaf / hard of hearing;I am blind / have difficulty se~ 1 0.0148  
## 7 I am deaf / hard of hearing;I am blind / have difficulty se~ 1 0.0148  
## 8 I am deaf / hard of hearing;I am blind / have difficulty se~ 1 0.0148  
## 9 I am deaf / hard of hearing;Or, in your own words: 3 0.0444  
## 10 I am unable to / find it difficult to type 17 0.251   
## 11 I am unable to / find it difficult to type;I am unable to /~ 2 0.0296  
## 12 I am unable to / find it difficult to type;Or, in your own ~ 4 0.0591  
## 13 I am unable to / find it difficult to walk or stand without~ 13 0.192   
## 14 None of the above 6520 96.4   
## 15 Or, in your own words: 54 0.798   
## 16 Prefer not to say 48 0.710

Convert the column into a binary column

dataset[dataset$Accessibility=="None of the above","Accessibility"] <- 0  
dataset[dataset$Accessibility!=0,"Accessibility"] <- 1

### Mental Health

Apply same logic as what I used for the accessibility variable

unique(dataset$MentalHealth)

dataset %>%group\_by(MentalHealth) %>%summarise(cnt = n())%>%mutate(percentage=(cnt/sum(cnt))\*100)%>%arrange(desc(percentage))%>%head(5)

## # A tibble: 5 x 3  
## MentalHealth cnt percentage  
## <chr> <int> <dbl>  
## 1 None of the above 4414 65.3   
## 2 I have a concentration and/or memory disorder (e.g. ADHD) 517 7.64  
## 3 I have an anxiety disorder 330 4.88  
## 4 I have a mood or emotional disorder (e.g. depression, bipola~ 261 3.86  
## 5 I have a mood or emotional disorder (e.g. depression, bipola~ 249 3.68

Create a binary column

dataset[dataset$MentalHealth=="None of the above","MentalHealth"] <- 0  
dataset[dataset$MentalHealth!=0,"MentalHealth"] <- 1

### Currency

Let’s look at currency. A very few respondents get paid in non-us dollars. Since I selected only developers in the United States, this does not sound intuitive. Given that the number is very few and the situation is strange, I guess they are wrong entries. Let’s drop the rows that contain non-us currency

unique(dataset$Currency)

## [1] "USD\tUnited States dollar" "UGX\tUgandan shilling"   
## [3] "EUR European Euro" "CAD\tCanadian dollar"   
## [5] "MKD\tMacedonian denar"

dataset %>%group\_by(Currency) %>%summarise(cnt = n())%>%mutate(percentage=(cnt/sum(cnt))\*100)%>%arrange(desc(percentage))

## # A tibble: 5 x 3  
## Currency cnt percentage  
## <chr> <int> <dbl>  
## 1 "USD\tUnited States dollar" 6757 99.9   
## 2 "CAD\tCanadian dollar" 3 0.0444  
## 3 "EUR European Euro" 1 0.0148  
## 4 "MKD\tMacedonian denar" 1 0.0148  
## 5 "UGX\tUgandan shilling" 1 0.0148

dataset<- dataset[dataset$Currency == "USD\tUnited States dollar",]

### Trans

Let’s get rows containing Yes or No

unique(dataset$Trans)

## [1] "No" "Yes" "Prefer not to say"   
## [4] "Or, in your own words:"

dataset <- dataset[(dataset$Trans=="No"|dataset$Trans=="Yes"),]

I use ConvertedCompYearly as a dependent variable. Let’s again drop some variables that are not necessary for our project.

dataset <- subset(dataset,select=-c(Currency, CompTotal, CompFreq))  
dataset <- subset(dataset,select=-c(MainBranch,Country))  
dataset <- subset(dataset,select=-c(ResponseId))

Okay. Let’s end our preliminary data cleaning here. We will move onto some data analysis. First, let’s look at the dependent variable.

## Data Analysis

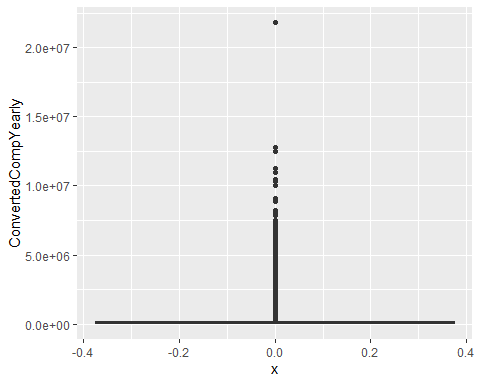
It seems that the variable is skewed a lot. Median and mean are totally different. We need to consider dropping some observations.

summary(dataset$ConvertedCompYearly)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 95000 130000 261462 173000 21822250

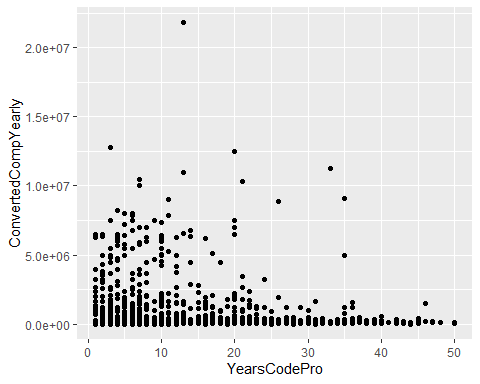
Let’s look at the box plot. Since the distribution is very skewed, it is impossible to see the whole distribution.

ggplot(dataset,aes(x=0,y=ConvertedCompYearly))+  
 geom\_boxplot()



Let’s see a relationship between the dependent variable and a variable of interest. Due to very high compensation amount, we can’t see anything about the most of values of the dependent variable. Let’s divide the variable into two groups, high and low.

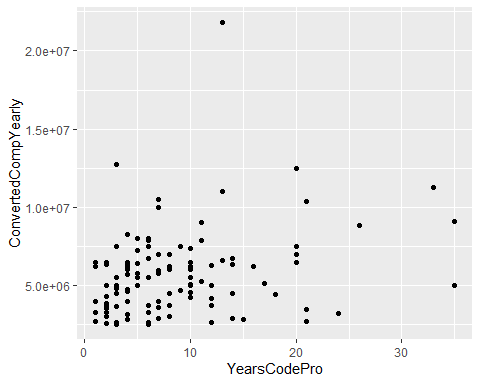
ggplot(dataset)+  
 geom\_point(aes(x=YearsCodePro,y=ConvertedCompYearly))



high\_comp <-dataset[dataset$ConvertedCompYearly>=2.5e+06,]  
low\_comp <-dataset[dataset$ConvertedCompYearly<2.5e+06,]

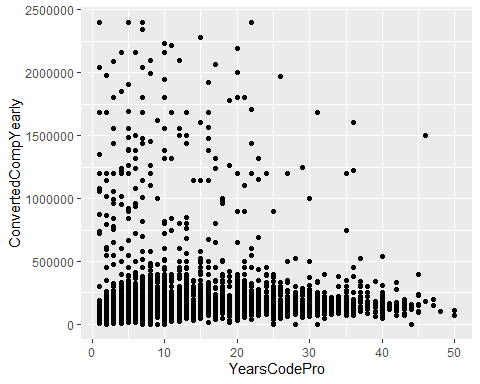
Let’s visualize them. This is high compenstaion. There might be a linear pattern in this.

#let's see plots and correlatino coefficient   
ggplot(high\_comp)+  
 geom\_point(aes(x=YearsCodePro,y=ConvertedCompYearly))



This is low compensation. It looks like that there are still some outliers. Majority of values are under 500000. Since I would like to find a pattern that many observations follow, in this research, I focus on the majority values.

ggplot(low\_comp)+  
 geom\_point(aes(x=YearsCodePro,y=ConvertedCompYearly))

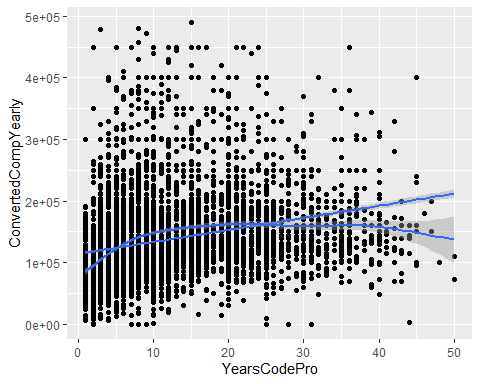


Get majority of the dependent variable and plot them. When looking at this, there is a non linear pattern. The compensation amount goes up by a certain point and move horizontally(or maybe decrease)

majority\_part <-dataset[dataset$ConvertedCompYearly<500000,]  
ggplot(majority\_part,aes(x=YearsCodePro,y=ConvertedCompYearly))+  
 geom\_point()+  
 geom\_smooth(method = "lm")+  
 geom\_smooth()

## `geom\_smooth()` using formula 'y ~ x'

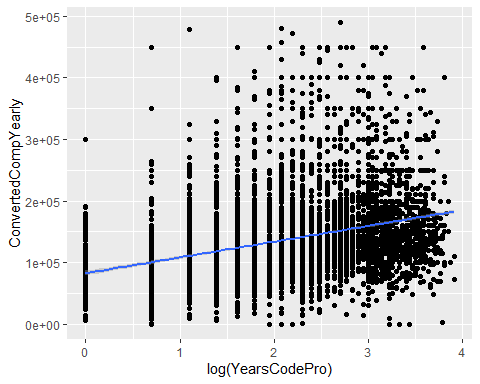
## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



Let’s apply log transformation on the independent variable. This is more linear.

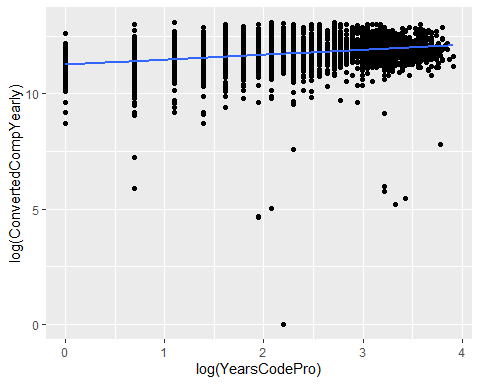
ggplot(majority\_part,aes(x=log(YearsCodePro),y=ConvertedCompYearly))+  
 geom\_point()+  
 geom\_smooth(method="lm")

## `geom\_smooth()` using formula 'y ~ x'

 Apply it on both of dependent and independent. This is great.

ggplot(majority\_part,aes(x=log(YearsCodePro),y=log(ConvertedCompYearly)))+  
 geom\_point()+  
 geom\_smooth(method="lm")

## `geom\_smooth()` using formula 'y ~ x'

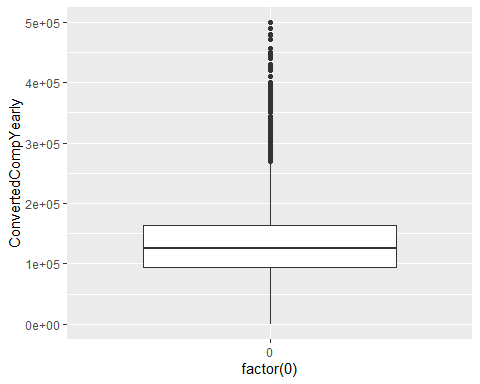


Given these plots, I think that it would be better to focus on the majority of data point because the original dependent variable is very skewed and there is likely a patter in the majority of data points.

Look at the distribution of the majority part. There are some data points that can be considered outliers(black points). I think that it is okay to actually consider them outliers.

ggplot(dataset,aes(x=factor(0),y=ConvertedCompYearly))+  
 geom\_boxplot()+  
 ylim(0,500000)

## Warning: Removed 296 rows containing non-finite values (stat\_boxplot).



Points outside of the whiskers are considered as outliers in the boxplot. So, Let’s calculate the upper and lower bound of the whiskers.

upper\_limit <-(173000-95000)\*1.5 + 173000  
lower\_limit <-(173000-95000)\*1.5- 95000

Okay. Then, let’s get rows that are not outliers.

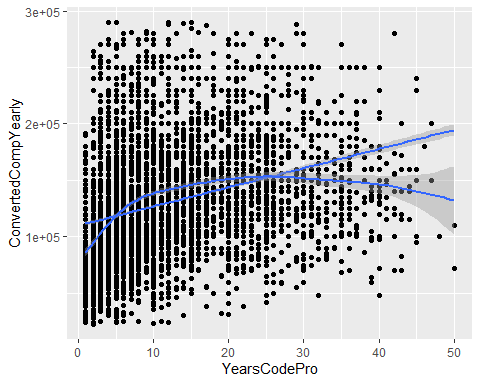
dataset <- dataset[dataset$ConvertedCompYearly<=upper\_limit & dataset$ConvertedCompYearly>=lower\_limit ,]

Then, let’s see the scatter plots again

ggplot(dataset,aes(x=YearsCodePro,y=ConvertedCompYearly))+  
 geom\_point()+  
 geom\_smooth(method="lm")+  
 geom\_smooth()

## `geom\_smooth()` using formula 'y ~ x'

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

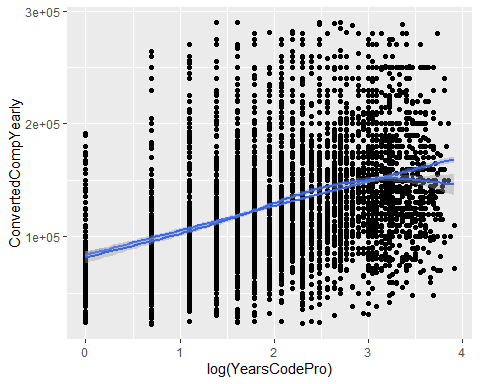


Log transformation on the dependent variable

ggplot(dataset,aes(x=log(YearsCodePro),y=ConvertedCompYearly))+  
 geom\_point()+  
 geom\_smooth(method="lm")+  
 geom\_smooth()

## `geom\_smooth()` using formula 'y ~ x'

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

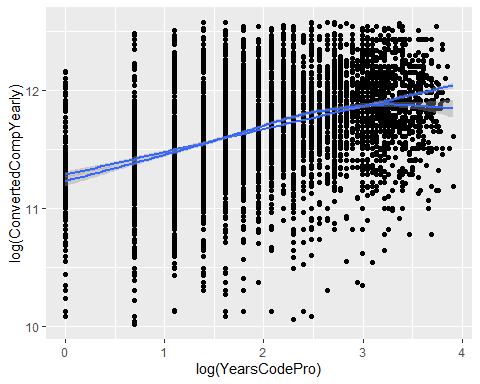


Log transformation on both of the variables. Even after log-log transformation is applied, from a certain point, the pattern becomes non linear. The compensation becomes constant or decreased from a certain point of years of professional coding experience.

ggplot(dataset,aes(x=log(YearsCodePro),y=log(ConvertedCompYearly)))+  
 geom\_point()+  
 geom\_smooth(method = "lm")+  
 geom\_smooth()

## `geom\_smooth()` using formula 'y ~ x'

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

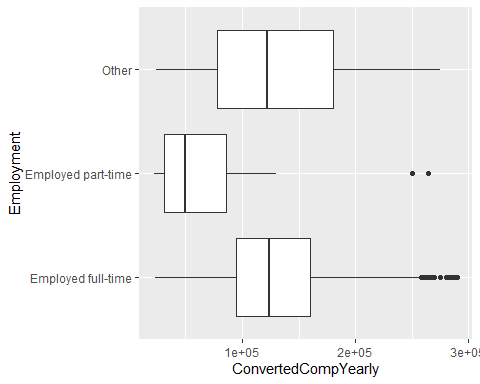


Alright, let’s see a relationship between the dependent and the each of the independent variables.

By Employement

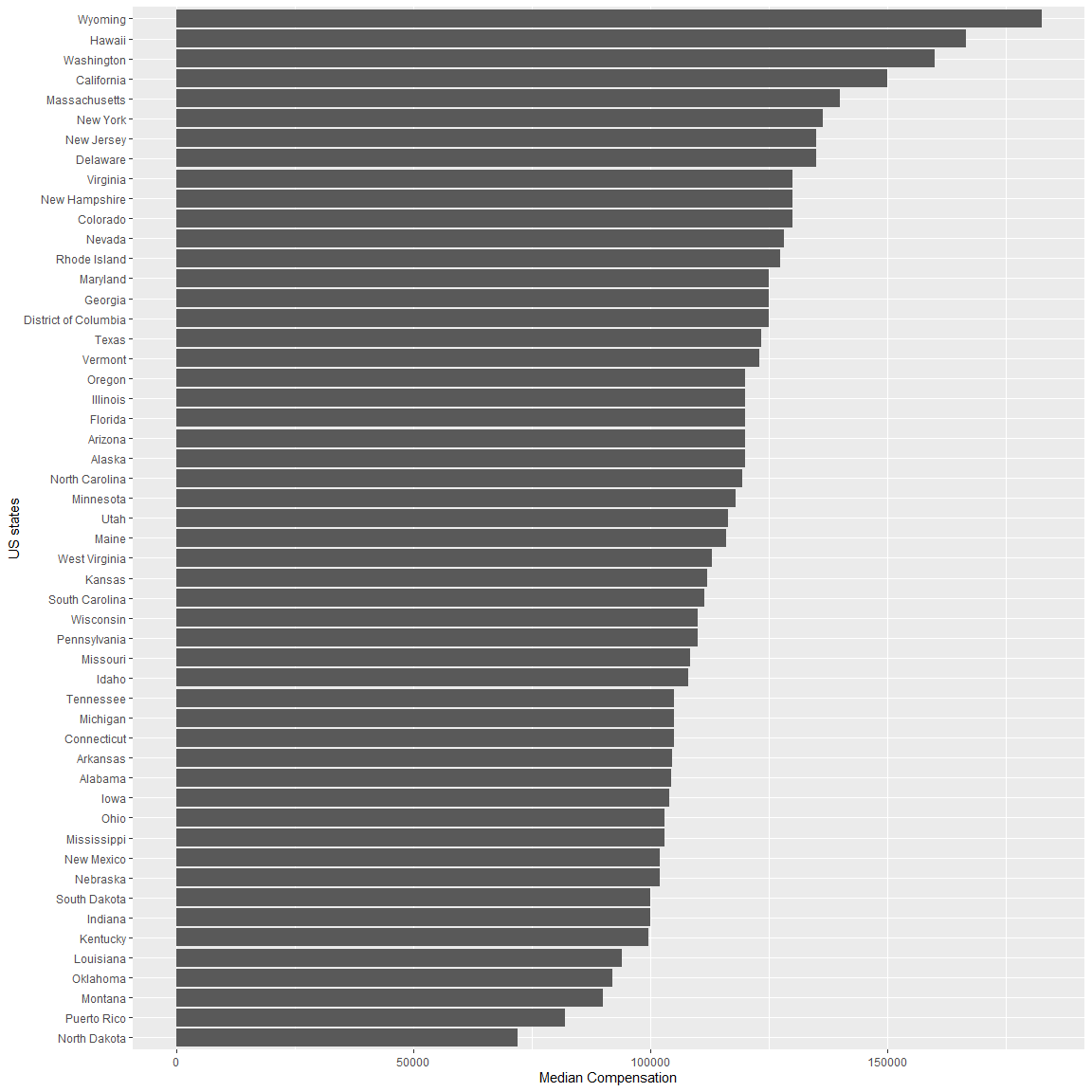
ggplot(data = dataset, mapping = aes(x =Employment, y =ConvertedCompYearly))+  
 geom\_boxplot(horizontal=TRUE)+  
 coord\_flip()

## Warning: Ignoring unknown parameters: horizontal



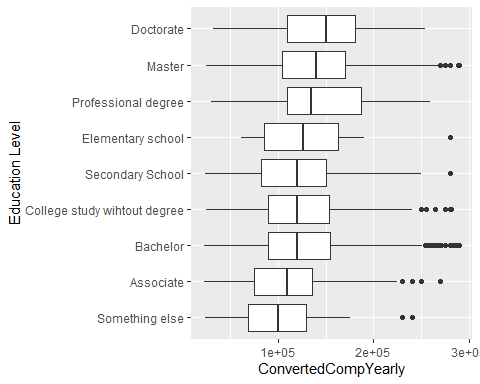
By States

group\_by\_state <- group\_by(dataset,US\_State) %>% summarise(median=median(ConvertedCompYearly))  
ggplot(data=group\_by\_state, mapping=aes(x=reorder(US\_State,median),y=median))+  
 geom\_bar(stat="identity")+  
 xlab("US states")+  
 ylab("Median Compensation")+  
 coord\_flip()

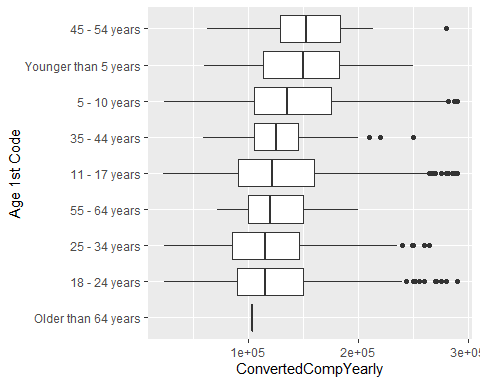


By Education Level

ggplot(data = dataset, mapping = aes(x = reorder(EdLevel,ConvertedCompYearly, FUN=median), y =ConvertedCompYearly))+  
 geom\_boxplot()+  
 xlab("Education Level")+  
 coord\_flip()

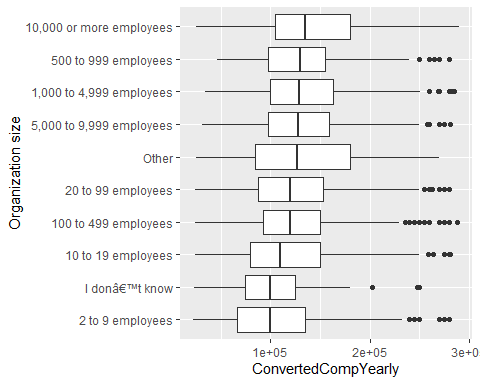
 Age1stCode This variables is about time when respondents write the first code.

ggplot(data = dataset, mapping = aes(x = reorder(Age1stCode,ConvertedCompYearly,FUN=median), y =ConvertedCompYearly))+  
 xlab("Age 1st Code")+  
 geom\_boxplot()+coord\_flip()



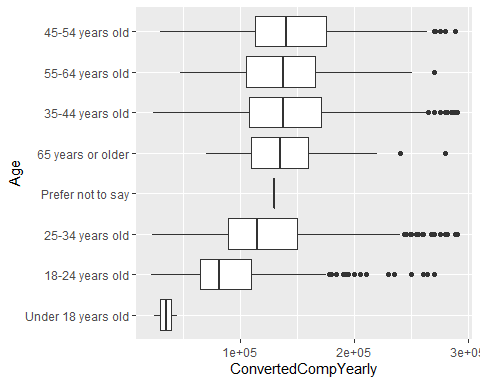
Organization Size

ggplot(data = dataset, mapping = aes(x = reorder(OrgSize,ConvertedCompYearly,FUN=median), y =ConvertedCompYearly))+  
 xlab("Organization size")+  
 geom\_boxplot()+coord\_flip()



Age

ggplot(data = dataset, mapping = aes(x = reorder(Age,ConvertedCompYearly,FUN=median), y =ConvertedCompYearly))+  
 xlab("Age")+  
 geom\_boxplot()+coord\_flip()



Looks like that there are some respondents who are under 18 years old. I am skeptic that there are professional developers who are under 18 years old. In this project, I consider them errors. Moreover, There are only two respondents who fall in the category. So, it is safe to drop them.

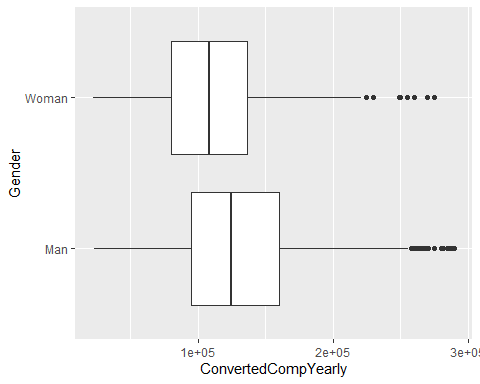
#4 persons are under 18   
sum(dataset$Age=="Under 18 years old")

## [1] 2

#drop them   
dataset <- dataset[dataset$Age!="Under 18 years old",]

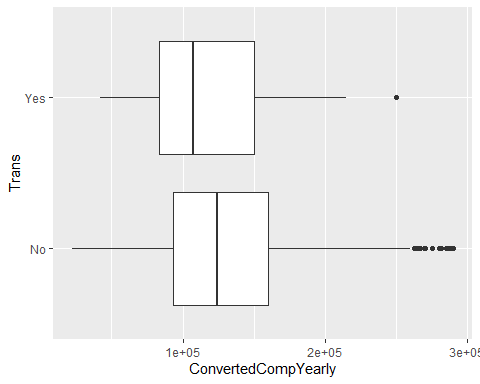
Gender

ggplot(data = dataset,aes(x = Gender, y=ConvertedCompYearly))+  
 geom\_boxplot()+coord\_flip()



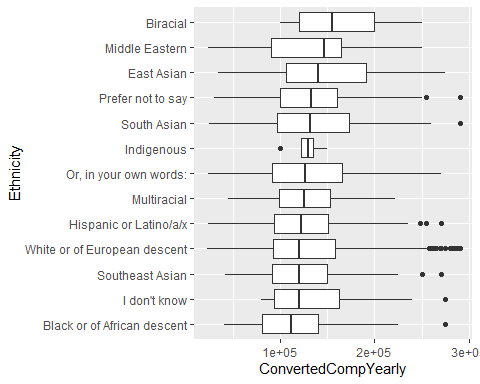
Trans

ggplot(data = dataset, mapping = aes(x =Trans, y =ConvertedCompYearly))+  
 geom\_boxplot()+coord\_flip()



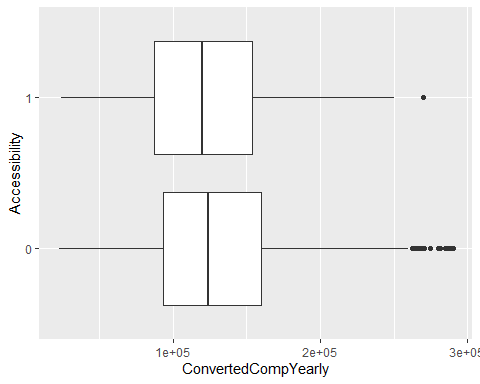
Ethinicity

ggplot(data = dataset, mapping = aes(x =reorder(Ethnicity,ConvertedCompYearly,FUN=median), y =ConvertedCompYearly))+  
 xlab("Ethnicity")+  
 geom\_boxplot()+coord\_flip()



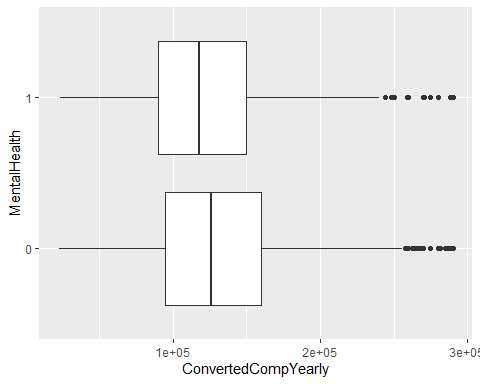
Accessibility

ggplot(data = dataset, mapping = aes(x = Accessibility, y =ConvertedCompYearly))+  
 geom\_boxplot()+coord\_flip()



Mentalhealth

ggplot(data = dataset,aes(x=MentalHealth,y=ConvertedCompYearly))+  
 geom\_boxplot()+coord\_flip()



## Modeling

Let’s start modeling

Since there are too many independent variables, let’s select them using the backward selection. I apply the feature selection on numerical columns only because R considers one of dummy variables as one variable and might drop it.

numerical\_data <-select\_if(dataset, is.numeric)  
regfit.bwd <- regsubsets(ConvertedCompYearly~. ,nvmax=45,data=numerical\_data,method="backward")  
regfit\_bwd\_summary <- summary(regfit.bwd)

Get the combination of variables which produce the highest adjusted R2

which.max(regfit\_bwd\_summary$adjr2)

## [1] 28

#let's use this numerical columns for our regression model  
numerical\_selected <- names(coef(regfit.bwd,28))[-1]  
names(coef(regfit.bwd,28))[-1]

## [1] "YearsCode"   
## [2] "YearsCodePro"   
## [3] "LanguageHaveWorkedWith"   
## [4] "PlatformHaveWorkedWith"   
## [5] "WebframeHaveWorkedWith"   
## [6] "ToolsTechHaveWorkedWith"   
## [7] "NEWCollabToolsHaveWorkedWith"   
## [8] "Developer\_back\_end"   
## [9] "Developer\_front\_end"   
## [10] "Developer\_full\_stack"   
## [11] "Engineer\_data"   
## [12] "Data\_scientist\_or\_machine\_learning\_specialist"  
## [13] "Developer\_desktop\_or\_enterprise\_applications"   
## [14] "Academic\_researcher"   
## [15] "Database\_administrator"   
## [16] "System\_administrator"   
## [17] "Developer\_embedded\_applications\_or\_devices"   
## [18] "DevOps\_specialist"   
## [19] "Engineering\_manager"   
## [20] "Engineer\_site\_reliability"   
## [21] "Developer\_mobile"   
## [22] "Developer\_QA\_or\_test"   
## [23] "Designer"   
## [24] "Developer\_game\_or\_graphics"   
## [25] "Data\_or\_business\_analyst"   
## [26] "Student"   
## [27] "Other\_devtypes"   
## [28] "Senior\_excecutives"

Prepare a dataset for modeling.

categorical <-names(select\_if(dataset,is.character))  
variables <- c(numerical\_selected,categorical)  
regression\_df <- subset(dataset,select=variables)  
regression\_df$ConvertedCompYearly <- dataset$ConvertedCompYearly

## First Model

Adjusted R-squared is 0.3931

Let’s look at YearsCode and YearsCodePro because they are this research’s primary interest. YearsCodePro is statistically significant at 5% significance level. By one year increase in professional experience, the yearly compensation raises by $2042 on average, all else equal. On the other hand, YearsCode is significant at 10 % level and the coefficient estimate is much smaller than YearsCodePro.

When looking at other columns, different variables affect the amount of compensation. States where developers live, types of developers, organization size, age, gender, and mental health affect the amount of compensation.

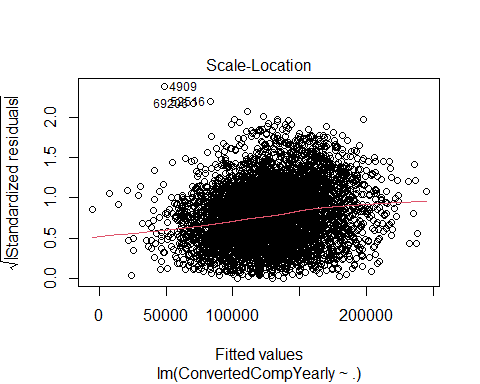
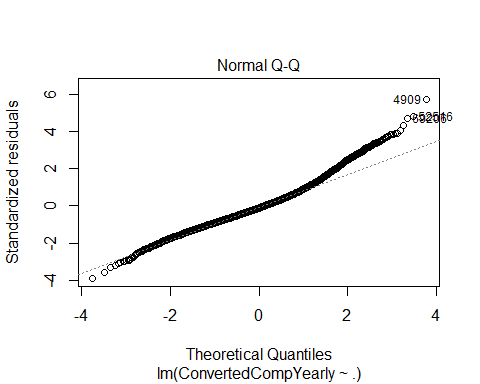
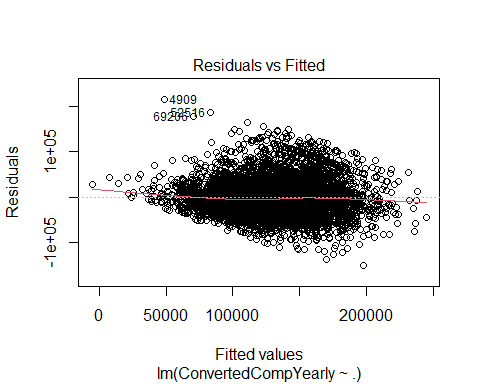
Check plots regarding this regression model next.

reg1 <- lm(ConvertedCompYearly~.,data=regression\_df)  
stargazer(reg1,type = "text")

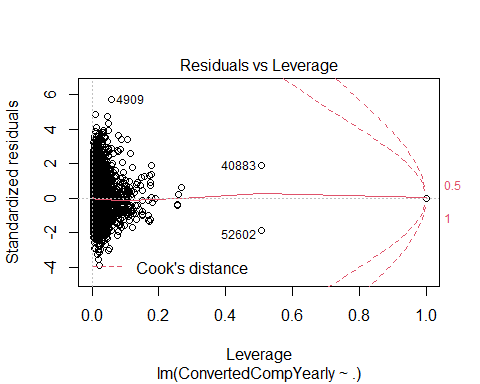
##   
## =========================================================================  
## Dependent variable:   
## ---------------------------  
## ConvertedCompYearly   
## -------------------------------------------------------------------------  
## YearsCode 40.167   
## (131.756)   
##   
## YearsCodePro 2,150.231\*\*\*   
## (156.524)   
##   
## LanguageHaveWorkedWith 325.558   
## (233.627)   
##   
## PlatformHaveWorkedWith 3,270.881\*\*\*   
## (576.368)   
##   
## WebframeHaveWorkedWith -1,639.621\*\*\*   
## (366.500)   
##   
## ToolsTechHaveWorkedWith 5,388.417\*\*\*   
## (433.577)   
##   
## NEWCollabToolsHaveWorkedWith -978.117\*\*\*   
## (335.396)   
##   
## Developer\_back\_end 6,687.142\*\*\*   
## (1,112.557)   
##   
## Developer\_front\_end -5,819.854\*\*\*   
## (1,292.507)   
##   
## Developer\_full\_stack -4,990.585\*\*\*   
## (1,179.184)   
##   
## Engineer\_data 4,110.041\*\*   
## (2,024.965)   
##   
## Data\_scientist\_or\_machine\_learning\_specialist 1,428.644   
## (2,430.254)   
##   
## Developer\_desktop\_or\_enterprise\_applications -3,038.745\*\*   
## (1,380.630)   
##   
## Academic\_researcher -28,521.950\*\*\*   
## (3,652.069)   
##   
## Database\_administrator -1,429.808   
## (2,112.289)   
##   
## System\_administrator -9,672.124\*\*\*   
## (2,122.615)   
##   
## Developer\_embedded\_applications\_or\_devices -2,896.370   
## (2,049.625)   
##   
## DevOps\_specialist 682.468   
## (1,707.041)   
##   
## Engineering\_manager 15,998.570\*\*\*   
## (2,010.508)   
##   
## Engineer\_site\_reliability 7,872.096\*\*\*   
## (2,556.696)   
##   
## Developer\_mobile 4,923.425\*\*\*   
## (1,739.732)   
##   
## Developer\_QA\_or\_test -8,414.873\*\*\*   
## (2,178.903)   
##   
## Designer -5,290.717\*\*   
## (2,144.085)   
##   
## Developer\_game\_or\_graphics 2,409.487   
## (3,330.008)   
##   
## Data\_or\_business\_analyst -13,001.570\*\*\*   
## (2,432.587)   
##   
## Student -4,613.981   
## (4,253.008)   
##   
## Other\_devtypes 6,765.919\*\*   
## (2,758.753)   
##   
## Senior\_excecutives 19,876.000\*\*\*   
## (3,324.876)   
##   
## EmploymentEmployed part-time -26,902.580\*\*\*   
## (6,345.974)   
##   
## EmploymentOther 1,802.021   
## (3,466.884)   
##   
## US\_StateAlaska 4,102.875   
## (15,822.620)   
##   
## US\_StateArizona 15,726.410\*\*   
## (6,685.961)   
##   
## US\_StateArkansas 12,302.760   
## (8,578.300)   
##   
## US\_StateCalifornia 42,609.750\*\*\*   
## (5,877.998)   
##   
## US\_StateColorado 21,160.270\*\*\*   
## (6,168.698)   
##   
## US\_StateConnecticut 18,343.320\*\*   
## (7,966.194)   
##   
## US\_StateDelaware 22,082.840   
## (15,000.580)   
##   
## US\_StateDistrict of Columbia 32,328.550\*\*\*   
## (8,603.861)   
##   
## US\_StateFlorida 15,733.680\*\*   
## (6,203.831)   
##   
## US\_StateGeorgia 18,260.810\*\*\*   
## (6,474.131)   
##   
## US\_StateHawaii 50,744.470\*\*\*   
## (13,622.090)   
##   
## US\_StateIdaho 2,803.428   
## (7,720.738)   
##   
## US\_StateIllinois 21,157.030\*\*\*   
## (6,225.086)   
##   
## US\_StateIndiana 2,843.308   
## (6,979.908)   
##   
## US\_StateIowa 6,049.586   
## (7,600.246)   
##   
## US\_StateKansas 12,680.640   
## (7,865.972)   
##   
## US\_StateKentucky -3,621.442   
## (7,476.619)   
##   
## US\_StateLouisiana -2,679.836   
## (8,953.227)   
##   
## US\_StateMaine 7,741.338   
## (9,676.391)   
##   
## US\_StateMaryland 20,175.130\*\*\*   
## (6,545.769)   
##   
## US\_StateMassachusetts 25,856.340\*\*\*   
## (6,232.922)   
##   
## US\_StateMichigan 5,148.937   
## (6,450.037)   
##   
## US\_StateMinnesota 10,133.280   
## (6,411.610)   
##   
## US\_StateMississippi -4,103.628   
## (13,656.030)   
##   
## US\_StateMissouri 6,701.754   
## (6,732.959)   
##   
## US\_StateMontana -12,700.400   
## (11,912.930)   
##   
## US\_StateNebraska 2,511.202   
## (8,008.334)   
##   
## US\_StateNevada 30,146.820\*\*\*   
## (10,123.340)   
##   
## US\_StateNew Hampshire 25,102.930\*\*\*   
## (8,715.388)   
##   
## US\_StateNew Jersey 23,227.800\*\*\*   
## (6,776.380)   
##   
## US\_StateNew Mexico -57.619   
## (8,474.258)   
##   
## US\_StateNew York 31,267.240\*\*\*   
## (6,054.081)   
##   
## US\_StateNorth Carolina 17,432.080\*\*\*   
## (6,380.227)   
##   
## US\_StateNorth Dakota -16,836.770   
## (13,620.010)   
##   
## US\_StateOhio 4,760.142   
## (6,301.494)   
##   
## US\_StateOklahoma -3,414.303   
## (8,788.472)   
##   
## US\_StateOregon 14,593.100\*\*   
## (6,478.614)   
##   
## US\_StatePennsylvania 10,553.780\*   
## (6,295.793)   
##   
## US\_StatePuerto Rico 22,979.270   
## (17,074.870)   
##   
## US\_StateRhode Island 16,071.910   
## (10,828.460)   
##   
## US\_StateSouth Carolina 16,007.510\*\*   
## (7,800.964)   
##   
## US\_StateSouth Dakota 7,517.268   
## (14,223.240)   
##   
## US\_StateTennessee 5,429.031   
## (7,025.800)   
##   
## US\_StateTexas 18,904.550\*\*\*   
## (5,988.478)   
##   
## US\_StateUtah 18,354.760\*\*\*   
## (6,285.863)   
##   
## US\_StateVermont 13,917.900   
## (9,045.944)   
##   
## US\_StateVirginia 23,174.560\*\*\*   
## (6,363.058)   
##   
## US\_StateWashington 47,935.700\*\*\*   
## (6,046.131)   
##   
## US\_StateWest Virginia 8,702.220   
## (16,952.040)   
##   
## US\_StateWisconsin 10,670.760   
## (6,607.939)   
##   
## US\_StateWyoming 78,837.830\*\*\*   
## (28,253.380)   
##   
## EdLevelBachelor 12,875.610\*\*\*   
## (2,465.577)   
##   
## EdLevelCollege study wihtout degree 5,552.049\*\*   
## (2,770.546)   
##   
## EdLevelDoctorate 30,584.390\*\*\*   
## (3,998.450)   
##   
## EdLevelElementary school 24,075.960\*\*   
## (10,789.860)   
##   
## EdLevelMaster 18,327.620\*\*\*   
## (2,702.151)   
##   
## EdLevelProfessional degree 20,427.500\*\*   
## (8,745.241)   
##   
## EdLevelSecondary School 4,370.733   
## (4,549.031)   
##   
## EdLevelSomething else 3,700.475   
## (9,098.288)   
##   
## Age1stCode18 - 24 years -2,220.771   
## (1,379.998)   
##   
## Age1stCode25 - 34 years -1,542.471   
## (2,467.645)   
##   
## Age1stCode35 - 44 years -4,022.956   
## (4,722.348)   
##   
## Age1stCode45 - 54 years 2,020.969   
## (7,039.341)   
##   
## Age1stCode5 - 10 years 3,289.163\*\*   
## (1,531.791)   
##   
## Age1stCode55 - 64 years -8,428.220   
## (9,819.649)   
##   
## Age1stCodeOlder than 64 years -16,263.230   
## (39,628.150)   
##   
## Age1stCodeYounger than 5 years 141.809   
## (6,867.576)   
##   
## OrgSize10 to 19 employees -17,139.040\*\*\*   
## (2,567.322)   
##   
## OrgSize10,000 or more employees 8,973.284\*\*\*   
## (1,841.497)   
##   
## OrgSize100 to 499 employees -7,989.291\*\*\*   
## (1,829.905)   
##   
## OrgSize2 to 9 employees -27,663.010\*\*\*   
## (2,637.663)   
##   
## OrgSize20 to 99 employees -10,326.570\*\*\*   
## (1,881.328)   
##   
## OrgSize5,000 to 9,999 employees -1,523.524   
## (2,665.597)   
##   
## OrgSize500 to 999 employees -3,618.135   
## (2,419.706)   
##   
## OrgSizeI donâ€™t know -16,858.460\*\*\*   
## (4,181.238)   
##   
## OrgSizeOther -7,545.690\*   
## (4,491.603)   
##   
## Age25-34 years old 22,223.130\*\*\*   
## (1,939.821)   
##   
## Age35-44 years old 22,295.040\*\*\*   
## (2,393.172)   
##   
## Age45-54 years old 9,946.907\*\*\*   
## (3,243.685)   
##   
## Age55-64 years old -13,420.900\*\*\*   
## (4,408.690)   
##   
## Age65 years or older -22,793.900\*\*\*   
## (7,270.733)   
##   
## AgePrefer not to say 16,295.830   
## (40,107.990)   
##   
## GenderWoman -7,279.288\*\*\*   
## (2,209.882)   
##   
## TransYes -433.617   
## (5,977.835)   
##   
## EthnicityBlack or of African descent -16,632.760   
## (11,979.360)   
##   
## EthnicityEast Asian -990.402   
## (11,777.800)   
##   
## EthnicityHispanic or Latino/a/x -14,140.850   
## (11,679.740)   
##   
## EthnicityI don't know -10,829.470   
## (15,221.480)   
##   
## EthnicityIndigenous -8,534.268   
## (22,608.960)   
##   
## EthnicityMiddle Eastern -12,882.760   
## (13,136.240)   
##   
## EthnicityMultiracial -18,497.720   
## (13,535.600)   
##   
## EthnicityOr, in your own words: -12,581.260   
## (12,626.580)   
##   
## EthnicityPrefer not to say -10,992.810   
## (11,987.920)   
##   
## EthnicitySouth Asian -12,390.480   
## (11,707.320)   
##   
## EthnicitySoutheast Asian -19,452.650   
## (12,254.680)   
##   
## EthnicityWhite or of European descent -11,850.710   
## (11,348.640)   
##   
## Accessibility1 -3,747.792   
## (2,869.033)   
##   
## MentalHealth1 -4,253.948\*\*\*   
## (1,095.746)   
##   
## Constant 63,961.630\*\*\*   
## (13,173.300)   
##   
## -------------------------------------------------------------------------  
## Observations 6,102   
## R2 0.406   
## Adjusted R2 0.393   
## Residual Std. Error 38,939.820 (df = 5973)   
## F Statistic 31.873\*\*\* (df = 128; 5973)   
## =========================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Make several plots regarding the model. Although there are no outliers since no square root of standardized residuals is more than 3, it looks like that there is heteroskadasticity. Let’s see what causes it.

plot(reg1)

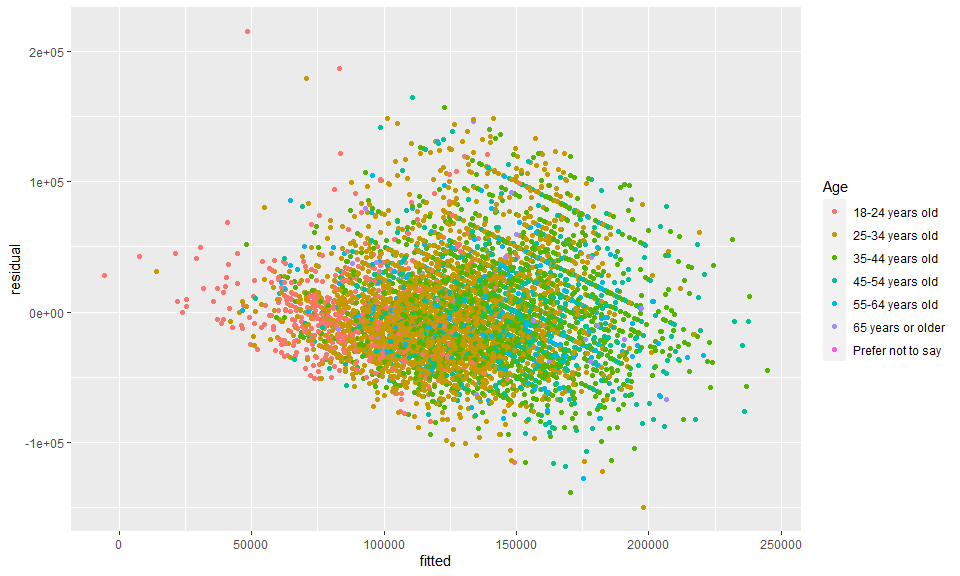


## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced  
  
## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced

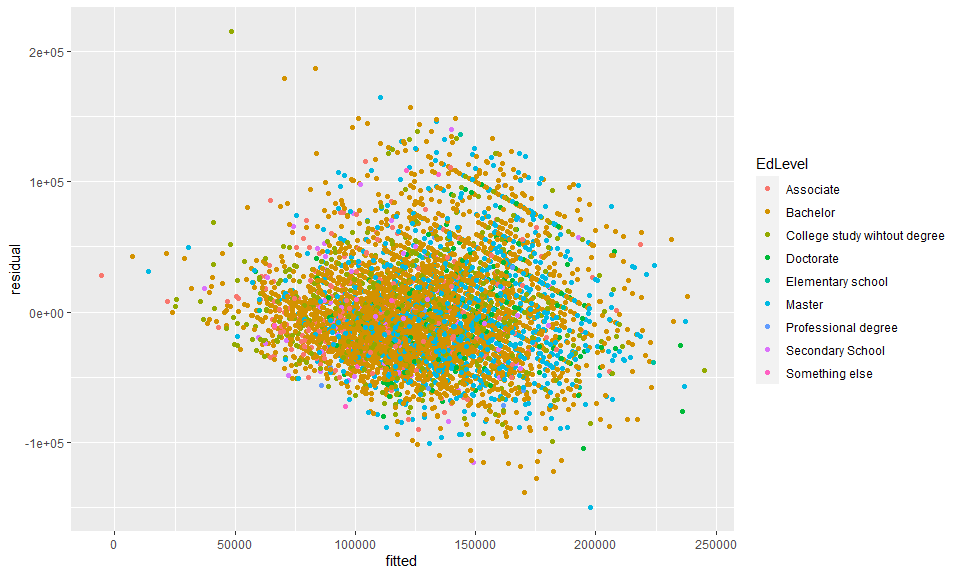


Check what causes the residual to vary aross fitted values Age

hetero\_check <- data.frame(dataset)  
hetero\_check$residual <- reg1$residuals  
hetero\_check$fitted <- reg1$fitted.values  
ggplot(hetero\_check,aes(x=fitted,y=residual,color=Age))+  
 geom\_point()

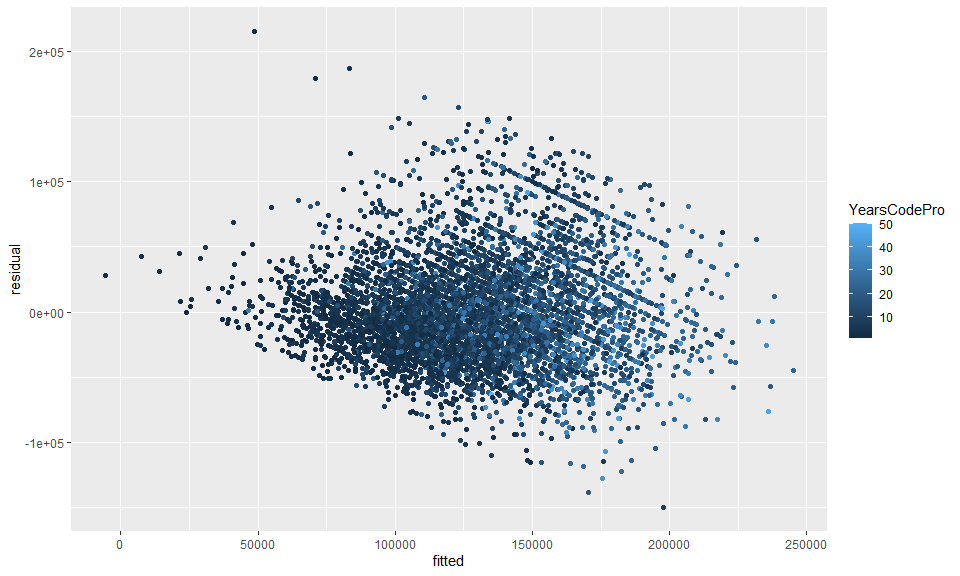
 Education level

ggplot(hetero\_check,aes(x=fitted,y=residual,color=EdLevel))+  
 geom\_point()



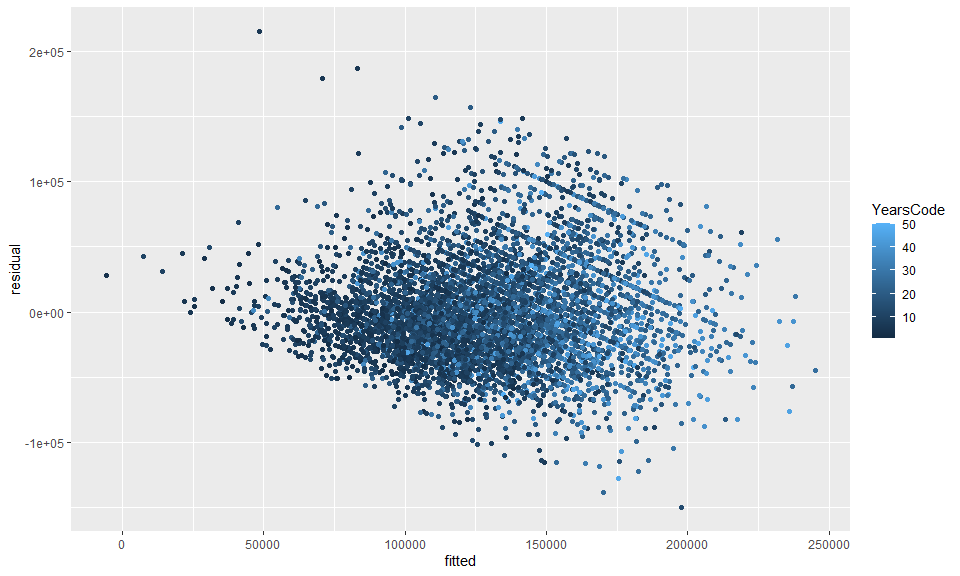
Years of experience as professional

ggplot(hetero\_check,aes(x=fitted,y=residual,color=YearsCodePro))+  
 geom\_point()



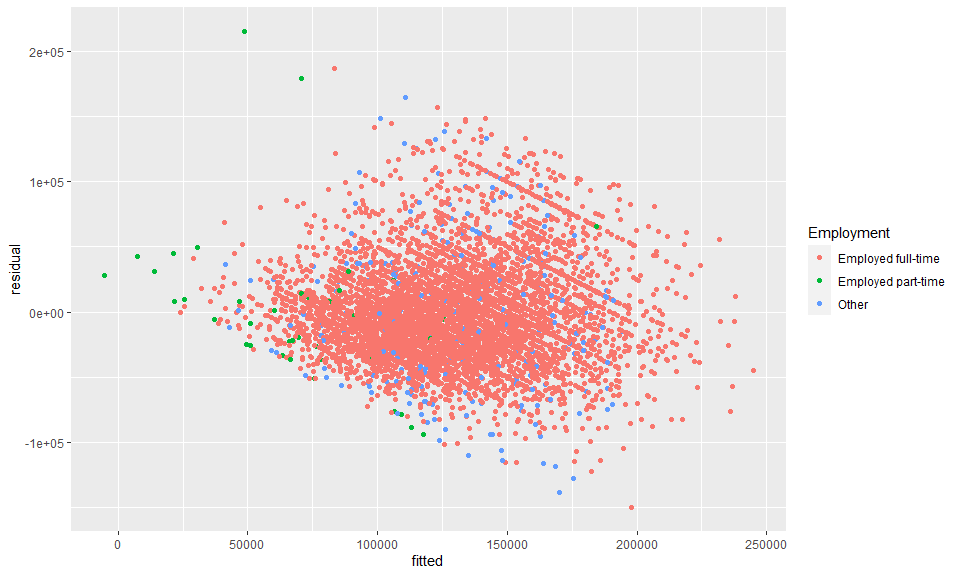
Years of coding

ggplot(hetero\_check,aes(x=fitted,y=residual,color=YearsCode))+  
 geom\_point()



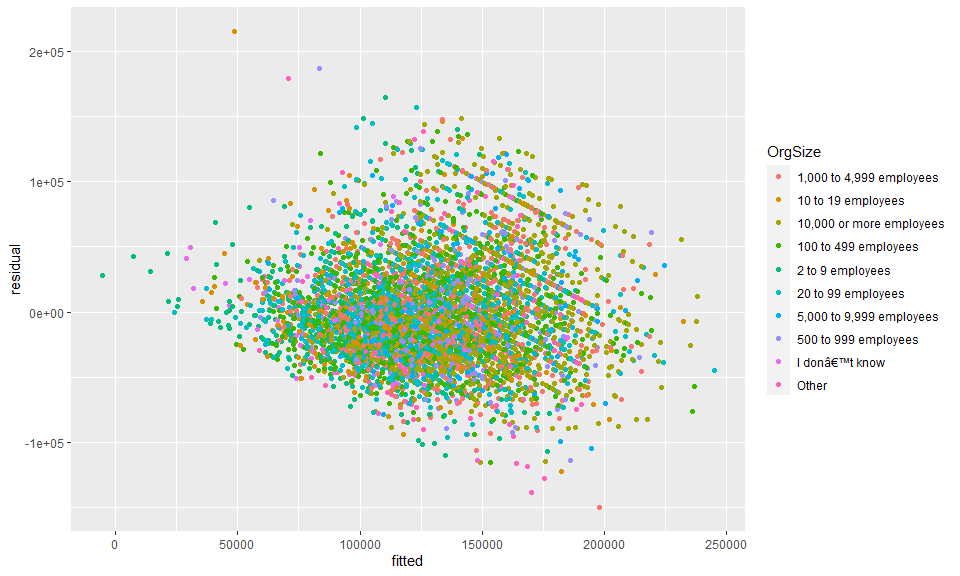
Employment Type

ggplot(hetero\_check,aes(x=fitted,y=residual,color=Employment))+  
 geom\_point()



Organization size

ggplot(hetero\_check,aes(x=fitted,y=residual,color=OrgSize))+  
 geom\_point()

 Looks like professional code experience and age cause this situation. Let’s solve this issue later. I also want to check if there is multicollinearity. I use VIF score.

It does not look like there is such an issue because no score exceeds 10.

vif(reg1)

## GVIF Df GVIF^(1/(2\*Df))  
## YearsCode 7.949072 1 2.819410  
## YearsCodePro 8.511147 1 2.917387  
## LanguageHaveWorkedWith 1.675441 1 1.294388  
## PlatformHaveWorkedWith 1.417078 1 1.190411  
## WebframeHaveWorkedWith 1.817131 1 1.348010  
## ToolsTechHaveWorkedWith 1.453637 1 1.205669  
## NEWCollabToolsHaveWorkedWith 1.402067 1 1.184089  
## Developer\_back\_end 1.238084 1 1.112692  
## Developer\_front\_end 1.339341 1 1.157299  
## Developer\_full\_stack 1.358579 1 1.165581  
## Engineer\_data 1.170813 1 1.082041  
## Data\_scientist\_or\_machine\_learning\_specialist 1.264368 1 1.124441  
## Developer\_desktop\_or\_enterprise\_applications 1.188785 1 1.090314  
## Academic\_researcher 1.152793 1 1.073682  
## Database\_administrator 1.551585 1 1.245626  
## System\_administrator 1.430665 1 1.196104  
## Developer\_embedded\_applications\_or\_devices 1.121636 1 1.059073  
## DevOps\_specialist 1.298773 1 1.139637  
## Engineering\_manager 1.095204 1 1.046520  
## Engineer\_site\_reliability 1.241349 1 1.114158  
## Developer\_mobile 1.152953 1 1.073756  
## Developer\_QA\_or\_test 1.175755 1 1.084322  
## Designer 1.224803 1 1.106708  
## Developer\_game\_or\_graphics 1.063047 1 1.031042  
## Data\_or\_business\_analyst 1.218195 1 1.103719  
## Student 1.080923 1 1.039674  
## Other\_devtypes 1.041114 1 1.020350  
## Senior\_excecutives 1.122088 1 1.059286  
## Employment 2.174830 2 1.214385  
## US\_State 2.404642 51 1.008639  
## EdLevel 1.597710 8 1.029719  
## Age1stCode 2.115971 8 1.047959  
## OrgSize 2.880770 9 1.060543  
## Age 6.327843 6 1.166195  
## Gender 1.161735 1 1.077838  
## Trans 1.122306 1 1.059390  
## Ethnicity 1.508874 12 1.017288  
## Accessibility 1.060265 1 1.029691  
## MentalHealth 1.097147 1 1.047448

## Model after dropping unsignificant variabels from the frist

Adjusted R-squared is 0.3922

Let’s drop Trans ad Ethnicity because they are insignificant and their boxplots do not show significant difference.

regression\_df <-subset(regression\_df,select=-c(Age1stCode,Trans,Ethnicity,Accessibility))

Run a new model with new variables. There is not a significant change.

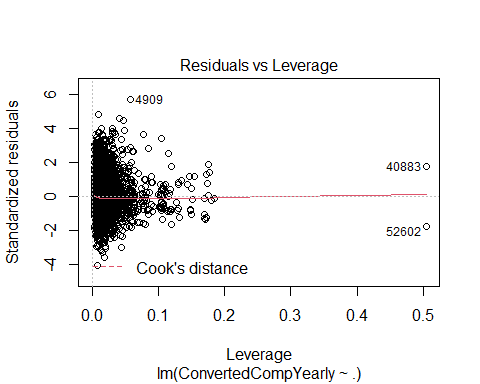
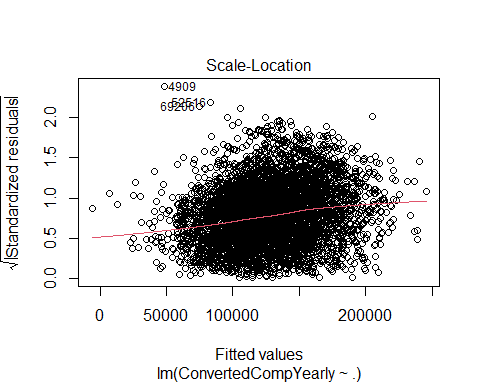
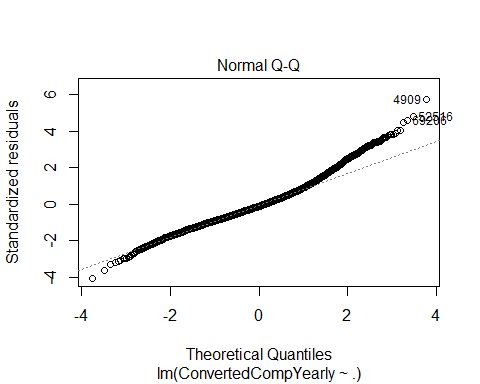
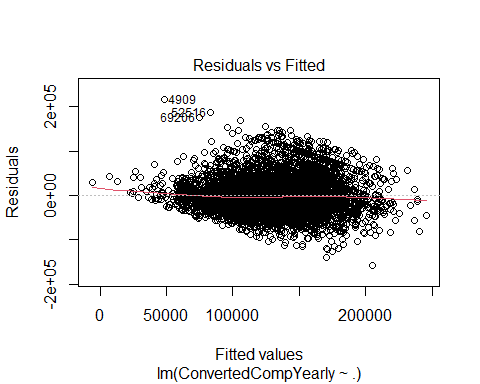
model\_2nd <- lm(ConvertedCompYearly~.,data=regression\_df)  
stargazer(model\_2nd,type = "text")

##   
## =========================================================================  
## Dependent variable:   
## ---------------------------  
## ConvertedCompYearly   
## -------------------------------------------------------------------------  
## YearsCode 202.585\*   
## (117.944)   
##   
## YearsCodePro 2,063.710\*\*\*   
## (153.044)   
##   
## LanguageHaveWorkedWith 376.546   
## (232.368)   
##   
## PlatformHaveWorkedWith 3,321.768\*\*\*   
## (575.688)   
##   
## WebframeHaveWorkedWith -1,710.899\*\*\*   
## (365.240)   
##   
## ToolsTechHaveWorkedWith 5,364.555\*\*\*   
## (433.466)   
##   
## NEWCollabToolsHaveWorkedWith -996.242\*\*\*   
## (334.337)   
##   
## Developer\_back\_end 6,669.416\*\*\*   
## (1,111.218)   
##   
## Developer\_front\_end -5,749.048\*\*\*   
## (1,290.980)   
##   
## Developer\_full\_stack -4,995.193\*\*\*   
## (1,178.295)   
##   
## Engineer\_data 4,043.559\*\*   
## (2,024.544)   
##   
## Data\_scientist\_or\_machine\_learning\_specialist 1,764.037   
## (2,423.063)   
##   
## Developer\_desktop\_or\_enterprise\_applications -2,902.288\*\*   
## (1,379.873)   
##   
## Academic\_researcher -28,642.860\*\*\*   
## (3,648.390)   
##   
## Database\_administrator -1,364.840   
## (2,109.502)   
##   
## System\_administrator -9,754.703\*\*\*   
## (2,119.706)   
##   
## Developer\_embedded\_applications\_or\_devices -2,838.669   
## (2,047.473)   
##   
## DevOps\_specialist 864.481   
## (1,704.949)   
##   
## Engineering\_manager 16,019.430\*\*\*   
## (2,008.818)   
##   
## Engineer\_site\_reliability 8,247.314\*\*\*   
## (2,553.218)   
##   
## Developer\_mobile 4,668.195\*\*\*   
## (1,737.762)   
##   
## Developer\_QA\_or\_test -8,790.407\*\*\*   
## (2,175.700)   
##   
## Designer -5,291.777\*\*   
## (2,142.575)   
##   
## Developer\_game\_or\_graphics 2,348.927   
## (3,323.203)   
##   
## Data\_or\_business\_analyst -13,115.440\*\*\*   
## (2,431.008)   
##   
## Student -4,378.042   
## (4,252.094)   
##   
## Other\_devtypes 6,656.693\*\*   
## (2,751.041)   
##   
## Senior\_excecutives 20,231.400\*\*\*   
## (3,320.536)   
##   
## EmploymentEmployed part-time -27,377.980\*\*\*   
## (6,339.263)   
##   
## EmploymentOther 1,695.286   
## (3,461.825)   
##   
## US\_StateAlaska 4,623.414   
## (15,822.800)   
##   
## US\_StateArizona 16,484.570\*\*   
## (6,673.226)   
##   
## US\_StateArkansas 12,627.290   
## (8,564.259)   
##   
## US\_StateCalifornia 43,912.830\*\*\*   
## (5,848.304)   
##   
## US\_StateColorado 21,762.270\*\*\*   
## (6,151.098)   
##   
## US\_StateConnecticut 18,471.640\*\*   
## (7,954.305)   
##   
## US\_StateDelaware 23,065.760   
## (14,931.520)   
##   
## US\_StateDistrict of Columbia 32,506.340\*\*\*   
## (8,586.570)   
##   
## US\_StateFlorida 16,161.360\*\*\*   
## (6,176.967)   
##   
## US\_StateGeorgia 18,567.660\*\*\*   
## (6,461.270)   
##   
## US\_StateHawaii 52,189.970\*\*\*   
## (13,605.360)   
##   
## US\_StateIdaho 4,099.130   
## (7,703.686)   
##   
## US\_StateIllinois 21,990.810\*\*\*   
## (6,205.467)   
##   
## US\_StateIndiana 3,089.395   
## (6,967.016)   
##   
## US\_StateIowa 6,847.216   
## (7,585.119)   
##   
## US\_StateKansas 13,312.050\*   
## (7,848.331)   
##   
## US\_StateKentucky -2,407.253   
## (7,454.179)   
##   
## US\_StateLouisiana -2,386.033   
## (8,941.523)   
##   
## US\_StateMaine 8,637.307   
## (9,667.242)   
##   
## US\_StateMaryland 20,774.020\*\*\*   
## (6,526.525)   
##   
## US\_StateMassachusetts 27,029.340\*\*\*   
## (6,212.215)   
##   
## US\_StateMichigan 5,723.158   
## (6,432.482)   
##   
## US\_StateMinnesota 10,895.660\*   
## (6,391.426)   
##   
## US\_StateMississippi -4,355.945   
## (13,629.680)   
##   
## US\_StateMissouri 7,221.497   
## (6,717.374)   
##   
## US\_StateMontana -12,323.370   
## (11,903.190)   
##   
## US\_StateNebraska 3,083.022   
## (7,988.301)   
##   
## US\_StateNevada 32,334.320\*\*\*   
## (10,077.100)   
##   
## US\_StateNew Hampshire 25,753.800\*\*\*   
## (8,705.995)   
##   
## US\_StateNew Jersey 24,223.230\*\*\*   
## (6,758.064)   
##   
## US\_StateNew Mexico -290.265   
## (8,453.159)   
##   
## US\_StateNew York 32,315.150\*\*\*   
## (6,039.198)   
##   
## US\_StateNorth Carolina 17,647.670\*\*\*   
## (6,366.834)   
##   
## US\_StateNorth Dakota -15,430.980   
## (13,611.440)   
##   
## US\_StateOhio 5,667.954   
## (6,281.309)   
##   
## US\_StateOklahoma -2,663.811   
## (8,778.577)   
##   
## US\_StateOregon 15,508.380\*\*   
## (6,459.903)   
##   
## US\_StatePennsylvania 11,492.560\*   
## (6,275.204)   
##   
## US\_StatePuerto Rico 20,931.410   
## (16,966.200)   
##   
## US\_StateRhode Island 16,526.000   
## (10,819.490)   
##   
## US\_StateSouth Carolina 16,271.920\*\*   
## (7,784.961)   
##   
## US\_StateSouth Dakota 6,714.634   
## (14,217.110)   
##   
## US\_StateTennessee 5,841.728   
## (7,013.059)   
##   
## US\_StateTexas 19,772.310\*\*\*   
## (5,963.646)   
##   
## US\_StateUtah 19,269.990\*\*\*   
## (6,263.618)   
##   
## US\_StateVermont 15,323.950\*   
## (9,033.811)   
##   
## US\_StateVirginia 23,465.450\*\*\*   
## (6,341.055)   
##   
## US\_StateWashington 48,816.080\*\*\*   
## (6,022.951)   
##   
## US\_StateWest Virginia 9,154.894   
## (16,949.740)   
##   
## US\_StateWisconsin 11,385.200\*   
## (6,591.315)   
##   
## US\_StateWyoming 77,131.230\*\*\*   
## (28,243.310)   
##   
## EdLevelBachelor 12,801.990\*\*\*   
## (2,460.576)   
##   
## EdLevelCollege study wihtout degree 5,708.651\*\*   
## (2,767.610)   
##   
## EdLevelDoctorate 30,422.640\*\*\*   
## (3,995.298)   
##   
## EdLevelElementary school 23,945.850\*\*   
## (10,780.030)   
##   
## EdLevelMaster 18,102.410\*\*\*   
## (2,683.601)   
##   
## EdLevelProfessional degree 20,781.180\*\*   
## (8,739.630)   
##   
## EdLevelSecondary School 4,476.344   
## (4,546.327)   
##   
## EdLevelSomething else 3,198.149   
## (9,091.406)   
##   
## OrgSize10 to 19 employees -17,276.800\*\*\*   
## (2,563.674)   
##   
## OrgSize10,000 or more employees 9,104.863\*\*\*   
## (1,838.101)   
##   
## OrgSize100 to 499 employees -8,018.253\*\*\*   
## (1,827.193)   
##   
## OrgSize2 to 9 employees -27,728.980\*\*\*   
## (2,635.484)   
##   
## OrgSize20 to 99 employees -10,432.010\*\*\*   
## (1,878.715)   
##   
## OrgSize5,000 to 9,999 employees -1,407.613   
## (2,662.091)   
##   
## OrgSize500 to 999 employees -3,479.826   
## (2,417.255)   
##   
## OrgSizeI donâ€™t know -17,204.290\*\*\*   
## (4,175.455)   
##   
## OrgSizeOther -7,564.249\*   
## (4,487.099)   
##   
## Age25-34 years old 21,424.220\*\*\*   
## (1,910.548)   
##   
## Age35-44 years old 21,108.860\*\*\*   
## (2,319.362)   
##   
## Age45-54 years old 8,502.398\*\*\*   
## (3,130.870)   
##   
## Age55-64 years old -17,236.320\*\*\*   
## (4,181.506)   
##   
## Age65 years or older -27,630.610\*\*\*   
## (6,974.641)   
##   
## AgePrefer not to say 9,403.968   
## (39,236.560)   
##   
## GenderWoman -7,056.808\*\*\*   
## (2,110.389)   
##   
## MentalHealth1 -4,342.843\*\*\*   
## (1,082.155)   
##   
## Constant 50,397.830\*\*\*   
## (6,653.605)   
##   
## -------------------------------------------------------------------------  
## Observations 6,102   
## R2 0.403   
## Adjusted R2 0.392   
## Residual Std. Error 38,967.460 (df = 5995)   
## F Statistic 38.146\*\*\* (df = 106; 5995)   
## =========================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Look at the plots

plot(model\_2nd)

## Warning: not plotting observations with leverage one:  
## 338



Given the standardized residuals plot, there is no outlier. But, what about high leverage values? They tend to distort the linear model. So, if any, let’s drop them.

hv<-as.data.frame(hatvalues(model\_2nd))  
mn<-mean(hatvalues(model\_2nd))  
hv$warn <- ifelse(hv[, 'hatvalues(model\_2nd)']>3\*mn, 'x3',  
 ifelse(hv[, 'hatvalues(model\_2nd)']>2\*mn, 'x3', '-' ))  
hv\_high <- subset(hv, warn%in%c("x2", "x3"))  
regression\_df <-regression\_df[!rownames(regression\_df) %in% rownames(hv\_high),]

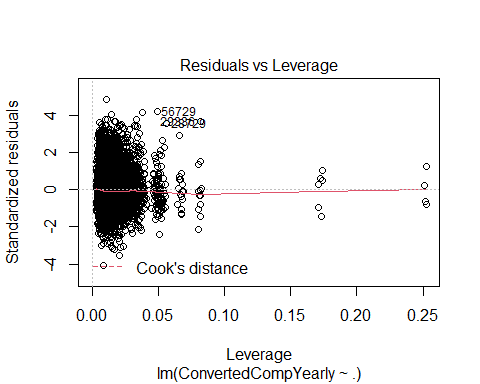
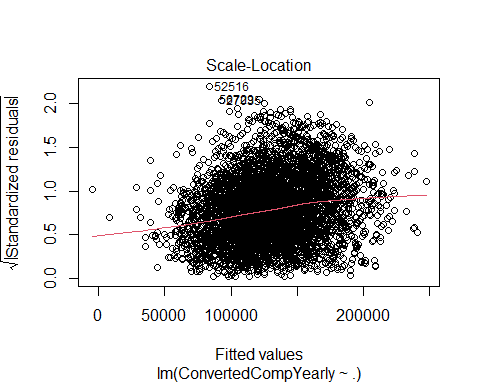
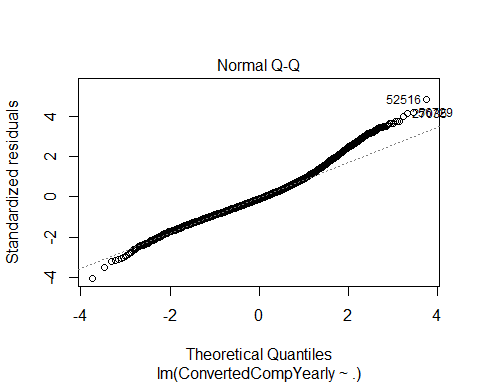
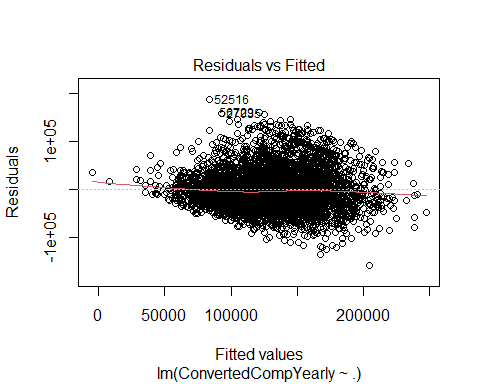
## Model without high leverage observations

Run the model. There is not a significant change.

model\_3rd <-lm(ConvertedCompYearly~.,data=regression\_df)  
stargazer(model\_3rd,type = "text")

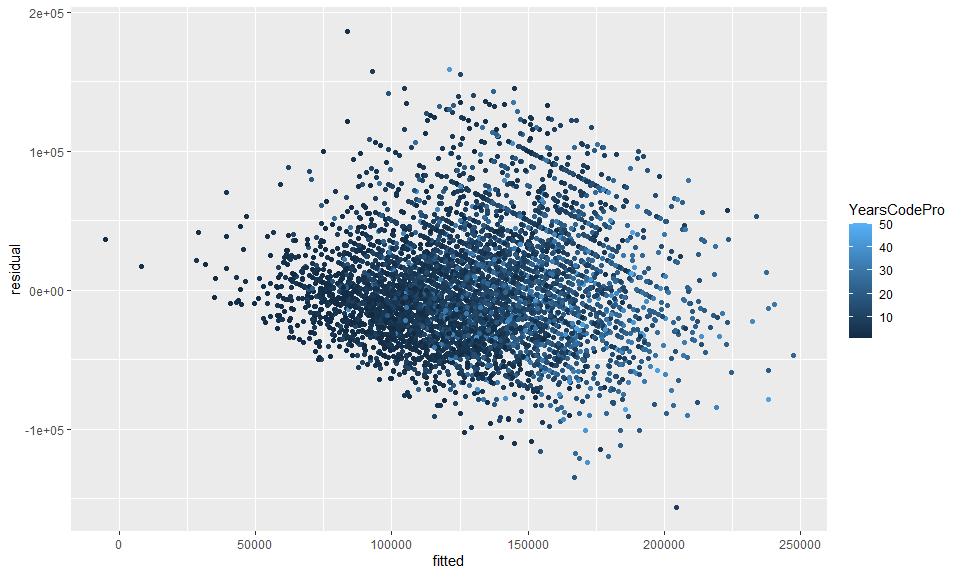
##   
## =========================================================================  
## Dependent variable:   
## ---------------------------  
## ConvertedCompYearly   
## -------------------------------------------------------------------------  
## YearsCode 225.378\*   
## (123.151)   
##   
## YearsCodePro 2,042.247\*\*\*   
## (160.892)   
##   
## LanguageHaveWorkedWith 519.284\*\*   
## (241.602)   
##   
## PlatformHaveWorkedWith 3,101.682\*\*\*   
## (598.446)   
##   
## WebframeHaveWorkedWith -1,748.212\*\*\*   
## (379.925)   
##   
## ToolsTechHaveWorkedWith 5,377.708\*\*\*   
## (449.016)   
##   
## NEWCollabToolsHaveWorkedWith -1,060.316\*\*\*   
## (345.946)   
##   
## Developer\_back\_end 6,653.657\*\*\*   
## (1,145.094)   
##   
## Developer\_front\_end -6,639.009\*\*\*   
## (1,336.717)   
##   
## Developer\_full\_stack -4,834.465\*\*\*   
## (1,217.074)   
##   
## Engineer\_data 5,024.345\*\*   
## (2,126.521)   
##   
## Data\_scientist\_or\_machine\_learning\_specialist 2,001.700   
## (2,542.997)   
##   
## Developer\_desktop\_or\_enterprise\_applications -2,838.104\*\*   
## (1,429.064)   
##   
## Academic\_researcher -28,730.250\*\*\*   
## (4,096.819)   
##   
## Database\_administrator -1,630.761   
## (2,207.394)   
##   
## System\_administrator -9,256.174\*\*\*   
## (2,218.330)   
##   
## Developer\_embedded\_applications\_or\_devices -3,367.609   
## (2,114.562)   
##   
## DevOps\_specialist 1,039.971   
## (1,767.014)   
##   
## Engineering\_manager 16,151.840\*\*\*   
## (2,050.458)   
##   
## Engineer\_site\_reliability 8,576.259\*\*\*   
## (2,654.827)   
##   
## Developer\_mobile 5,395.279\*\*\*   
## (1,805.749)   
##   
## Developer\_QA\_or\_test -9,114.336\*\*\*   
## (2,270.007)   
##   
## Designer -5,295.869\*\*   
## (2,264.825)   
##   
## Developer\_game\_or\_graphics 1,015.553   
## (3,524.882)   
##   
## Data\_or\_business\_analyst -14,161.450\*\*\*   
## (2,572.674)   
##   
## Student -6,376.209   
## (4,957.273)   
##   
## Other\_devtypes 6,711.470\*\*   
## (2,934.965)   
##   
## Senior\_excecutives 19,488.080\*\*\*   
## (3,628.841)   
##   
## EmploymentEmployed part-time -70,523.740\*\*\*   
## (15,923.230)   
##   
## EmploymentOther 751.650   
## (3,782.678)   
##   
## US\_StateArizona 16,925.760\*\*   
## (7,106.380)   
##   
## US\_StateArkansas 5,126.832   
## (10,176.860)   
##   
## US\_StateCalifornia 44,637.120\*\*\*   
## (6,334.799)   
##   
## US\_StateColorado 22,210.040\*\*\*   
## (6,611.995)   
##   
## US\_StateConnecticut 18,455.790\*\*   
## (8,758.597)   
##   
## US\_StateDistrict of Columbia 37,776.800\*\*\*   
## (10,488.120)   
##   
## US\_StateFlorida 15,817.990\*\*   
## (6,655.354)   
##   
## US\_StateGeorgia 18,139.790\*\*\*   
## (6,913.149)   
##   
## US\_StateIdaho 5,030.517   
## (8,124.399)   
##   
## US\_StateIllinois 21,893.210\*\*\*   
## (6,666.573)   
##   
## US\_StateIndiana 3,962.588   
## (7,408.233)   
##   
## US\_StateIowa 6,515.235   
## (7,977.275)   
##   
## US\_StateKansas 6,242.807   
## (8,470.992)   
##   
## US\_StateKentucky 555.916   
## (7,946.034)   
##   
## US\_StateLouisiana 11,470.190   
## (20,353.640)   
##   
## US\_StateMaryland 20,492.400\*\*\*   
## (6,973.675)   
##   
## US\_StateMassachusetts 27,662.350\*\*\*   
## (6,670.414)   
##   
## US\_StateMichigan 5,629.303   
## (6,879.544)   
##   
## US\_StateMinnesota 10,662.360   
## (6,833.027)   
##   
## US\_StateMissouri 7,273.840   
## (7,166.673)   
##   
## US\_StateNebraska 1,944.280   
## (8,544.816)   
##   
## US\_StateNew Hampshire 25,285.800\*\*   
## (11,501.680)   
##   
## US\_StateNew Jersey 25,317.770\*\*\*   
## (7,196.059)   
##   
## US\_StateNew Mexico -8,404.419   
## (10,323.350)   
##   
## US\_StateNew York 33,229.540\*\*\*   
## (6,516.602)   
##   
## US\_StateNorth Carolina 17,816.870\*\*\*   
## (6,827.141)   
##   
## US\_StateOhio 5,535.779   
## (6,735.526)   
##   
## US\_StateOklahoma 6,963.100   
## (12,404.630)   
##   
## US\_StateOregon 15,085.600\*\*   
## (6,926.554)   
##   
## US\_StatePennsylvania 12,393.890\*   
## (6,732.603)   
##   
## US\_StateSouth Carolina 16,647.240\*\*   
## (8,389.325)   
##   
## US\_StateTennessee 6,023.034   
## (7,411.239)   
##   
## US\_StateTexas 19,661.280\*\*\*   
## (6,442.552)   
##   
## US\_StateUtah 20,517.760\*\*\*   
## (6,727.999)   
##   
## US\_StateVirginia 23,841.910\*\*\*   
## (6,794.924)   
##   
## US\_StateWashington 49,478.260\*\*\*   
## (6,492.516)   
##   
## US\_StateWisconsin 11,528.410   
## (7,025.712)   
##   
## EdLevelBachelor 12,413.570\*\*\*   
## (2,563.297)   
##   
## EdLevelCollege study wihtout degree 5,696.264\*\*   
## (2,879.226)   
##   
## EdLevelDoctorate 27,157.560\*\*\*   
## (4,278.003)   
##   
## EdLevelMaster 17,585.410\*\*\*   
## (2,792.610)   
##   
## EdLevelSecondary School 3,739.891   
## (4,833.287)   
##   
## OrgSize10 to 19 employees -17,776.990\*\*\*   
## (2,660.210)   
##   
## OrgSize10,000 or more employees 9,693.600\*\*\*   
## (1,872.508)   
##   
## OrgSize100 to 499 employees -7,729.973\*\*\*   
## (1,870.552)   
##   
## OrgSize2 to 9 employees -28,752.030\*\*\*   
## (2,750.499)   
##   
## OrgSize20 to 99 employees -9,944.330\*\*\*   
## (1,919.658)   
##   
## OrgSize5,000 to 9,999 employees -1,083.006   
## (2,727.578)   
##   
## OrgSize500 to 999 employees -2,902.230   
## (2,473.328)   
##   
## OrgSizeI donâ€™t know -16,932.100\*\*\*   
## (4,463.069)   
##   
## OrgSizeOther -9,293.968\*   
## (4,895.339)   
##   
## Age25-34 years old 21,413.820\*\*\*   
## (1,963.442)   
##   
## Age35-44 years old 21,649.010\*\*\*   
## (2,396.375)   
##   
## Age45-54 years old 8,636.552\*\*\*   
## (3,250.560)   
##   
## Age55-64 years old -17,549.480\*\*\*   
## (4,380.047)   
##   
## Age65 years or older -38,063.030\*\*\*   
## (8,814.678)   
##   
## GenderWoman -5,721.419\*\*\*   
## (2,195.147)   
##   
## MentalHealth1 -4,273.429\*\*\*   
## (1,117.381)   
##   
## Constant 49,784.320\*\*\*   
## (7,141.058)   
##   
## -------------------------------------------------------------------------  
## Observations 5,660   
## R2 0.401   
## Adjusted R2 0.391   
## Residual Std. Error 38,727.810 (df = 5571)   
## F Statistic 42.356\*\*\* (df = 88; 5571)   
## =========================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

plot(model\_3rd)

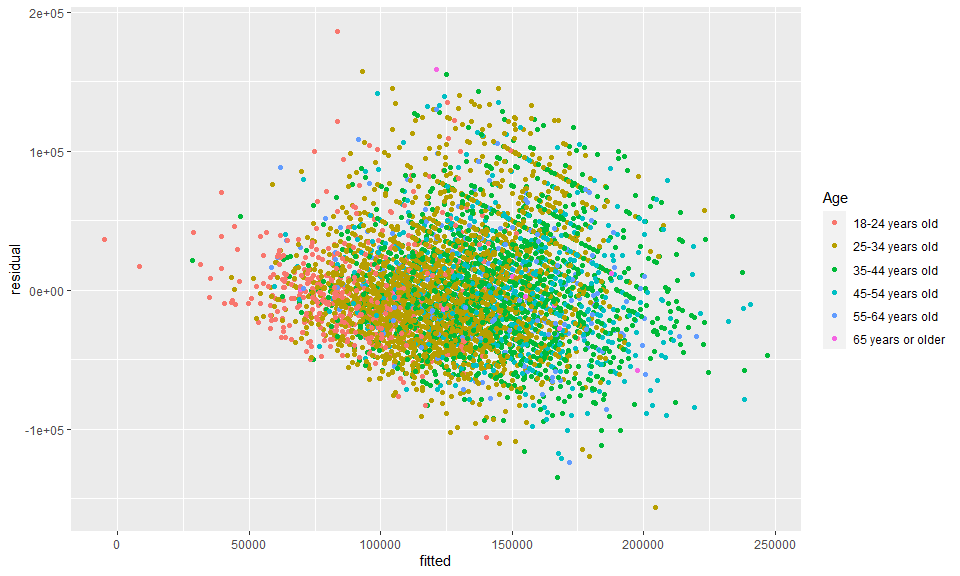


Check residual and fitted values plot again.

hetero\_check\_reg3 <- data.frame(regression\_df)  
hetero\_check\_reg3$residual <- model\_3rd$residuals  
hetero\_check\_reg3$fitted <- model\_3rd$fitted.values  
ggplot(hetero\_check\_reg3,aes(x=fitted,y=residual,color=YearsCodePro))+  
 geom\_point()

 Age

hetero\_check\_reg3$fitted <- model\_3rd$fitted.values  
ggplot(hetero\_check\_reg3,aes(x=fitted,y=residual,color=Age))+  
 geom\_point()



## Model with log transformation on both independent and dependent variables.

Let’s apply log transformation on YearsCodePro, YearsCode, and the dependent variable.This is because, in the scatter plots shown in the previous section, non-linear relationships were observed.

Apply log transformation

no\_leverage\_df\_log\_log <- data.frame(regression\_df)  
no\_leverage\_df\_log\_log$log\_compensation <- log(no\_leverage\_df\_log\_log$ConvertedCompYearly)  
no\_leverage\_df\_log\_log$log\_yearscode <- log(no\_leverage\_df\_log\_log$YearsCode)  
no\_leverage\_df\_log\_log$log\_yearscodepro <- log(no\_leverage\_df\_log\_log$YearsCodePro)  
  
  
no\_leverage\_df\_log\_log<-subset(no\_leverage\_df\_log\_log,select=-c(ConvertedCompYearly, YearsCode, YearsCodePro))

Run the model.

Adjusted R2 is 0.432.

Since I applied the log transformation, the interpretations on the log transformed coefficient estimates are changed. They are interpreted in terms of percent change. For instance, by one percent increase in professional coding experience, the yearly compensation increases by 0.181%.

model\_log\_log <- lm(log\_compensation~., data=no\_leverage\_df\_log\_log)  
stargazer(model\_log\_log,type = "text")

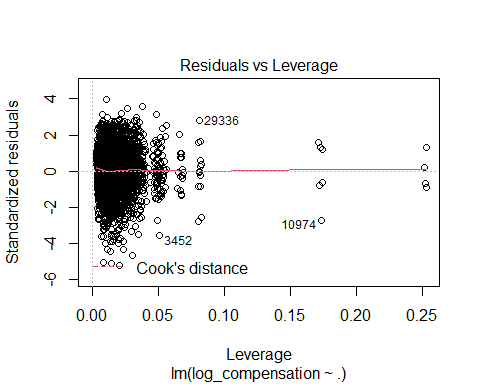
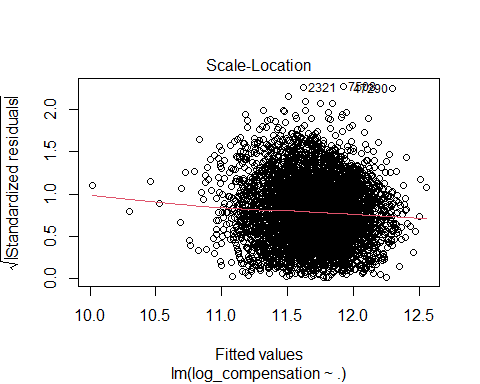
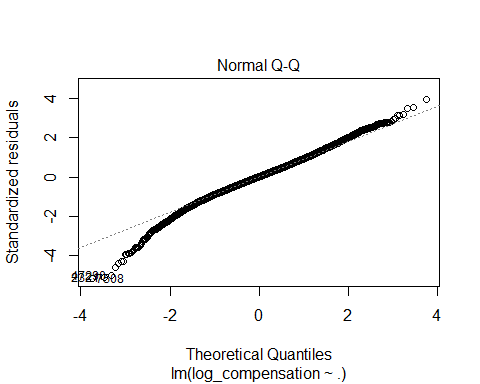
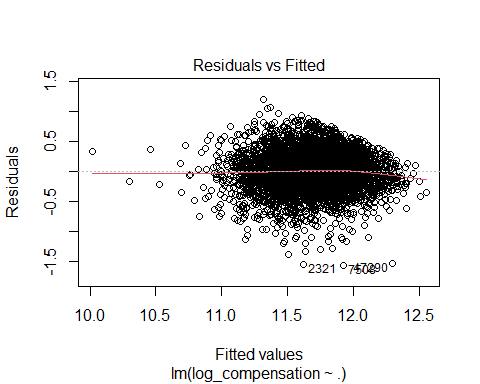
##   
## =========================================================================  
## Dependent variable:   
## ---------------------------  
## log\_compensation   
## -------------------------------------------------------------------------  
## LanguageHaveWorkedWith 0.002   
## (0.002)   
##   
## PlatformHaveWorkedWith 0.023\*\*\*   
## (0.005)   
##   
## WebframeHaveWorkedWith -0.011\*\*\*   
## (0.003)   
##   
## ToolsTechHaveWorkedWith 0.044\*\*\*   
## (0.004)   
##   
## NEWCollabToolsHaveWorkedWith -0.007\*\*\*   
## (0.003)   
##   
## Developer\_back\_end 0.043\*\*\*   
## (0.009)   
##   
## Developer\_front\_end -0.048\*\*\*   
## (0.010)   
##   
## Developer\_full\_stack -0.031\*\*\*   
## (0.010)   
##   
## Engineer\_data 0.049\*\*\*   
## (0.017)   
##   
## Data\_scientist\_or\_machine\_learning\_specialist 0.002   
## (0.020)   
##   
## Developer\_desktop\_or\_enterprise\_applications -0.023\*\*   
## (0.011)   
##   
## Academic\_researcher -0.261\*\*\*   
## (0.032)   
##   
## Database\_administrator -0.024   
## (0.017)   
##   
## System\_administrator -0.073\*\*\*   
## (0.017)   
##   
## Developer\_embedded\_applications\_or\_devices -0.015   
## (0.017)   
##   
## DevOps\_specialist 0.009   
## (0.014)   
##   
## Engineering\_manager 0.108\*\*\*   
## (0.016)   
##   
## Engineer\_site\_reliability 0.060\*\*\*   
## (0.021)   
##   
## Developer\_mobile 0.038\*\*\*   
## (0.014)   
##   
## Developer\_QA\_or\_test -0.072\*\*\*   
## (0.018)   
##   
## Designer -0.054\*\*\*   
## (0.018)   
##   
## Developer\_game\_or\_graphics -0.001   
## (0.028)   
##   
## Data\_or\_business\_analyst -0.102\*\*\*   
## (0.020)   
##   
## Student -0.079\*\*   
## (0.039)   
##   
## Other\_devtypes 0.053\*\*   
## (0.023)   
##   
## Senior\_excecutives 0.134\*\*\*   
## (0.028)   
##   
## EmploymentEmployed part-time -1.009\*\*\*   
## (0.125)   
##   
## EmploymentOther -0.042   
## (0.030)   
##   
## US\_StateArizona 0.155\*\*\*   
## (0.056)   
##   
## US\_StateArkansas 0.008   
## (0.080)   
##   
## US\_StateCalifornia 0.345\*\*\*   
## (0.050)   
##   
## US\_StateColorado 0.199\*\*\*   
## (0.052)   
##   
## US\_StateConnecticut 0.144\*\*   
## (0.069)   
##   
## US\_StateDistrict of Columbia 0.293\*\*\*   
## (0.082)   
##   
## US\_StateFlorida 0.128\*\*   
## (0.052)   
##   
## US\_StateGeorgia 0.171\*\*\*   
## (0.054)   
##   
## US\_StateIdaho 0.040   
## (0.064)   
##   
## US\_StateIllinois 0.177\*\*\*   
## (0.052)   
##   
## US\_StateIndiana 0.019   
## (0.058)   
##   
## US\_StateIowa 0.044   
## (0.062)   
##   
## US\_StateKansas 0.040   
## (0.066)   
##   
## US\_StateKentucky 0.011   
## (0.062)   
##   
## US\_StateLouisiana 0.103   
## (0.159)   
##   
## US\_StateMaryland 0.193\*\*\*   
## (0.055)   
##   
## US\_StateMassachusetts 0.234\*\*\*   
## (0.052)   
##   
## US\_StateMichigan 0.053   
## (0.054)   
##   
## US\_StateMinnesota 0.105\*\*   
## (0.053)   
##   
## US\_StateMissouri 0.051   
## (0.056)   
##   
## US\_StateNebraska 0.002   
## (0.067)   
##   
## US\_StateNew Hampshire 0.233\*\*\*   
## (0.090)   
##   
## US\_StateNew Jersey 0.207\*\*\*   
## (0.056)   
##   
## US\_StateNew Mexico -0.090   
## (0.081)   
##   
## US\_StateNew York 0.256\*\*\*   
## (0.051)   
##   
## US\_StateNorth Carolina 0.144\*\*\*   
## (0.053)   
##   
## US\_StateOhio 0.044   
## (0.053)   
##   
## US\_StateOklahoma -0.008   
## (0.097)   
##   
## US\_StateOregon 0.128\*\*   
## (0.054)   
##   
## US\_StatePennsylvania 0.090\*   
## (0.053)   
##   
## US\_StateSouth Carolina 0.140\*\*   
## (0.066)   
##   
## US\_StateTennessee 0.058   
## (0.058)   
##   
## US\_StateTexas 0.172\*\*\*   
## (0.050)   
##   
## US\_StateUtah 0.171\*\*\*   
## (0.053)   
##   
## US\_StateVirginia 0.198\*\*\*   
## (0.053)   
##   
## US\_StateWashington 0.369\*\*\*   
## (0.051)   
##   
## US\_StateWisconsin 0.082   
## (0.055)   
##   
## EdLevelBachelor 0.119\*\*\*   
## (0.020)   
##   
## EdLevelCollege study wihtout degree 0.063\*\*\*   
## (0.023)   
##   
## EdLevelDoctorate 0.230\*\*\*   
## (0.033)   
##   
## EdLevelMaster 0.155\*\*\*   
## (0.022)   
##   
## EdLevelSecondary School 0.033   
## (0.038)   
##   
## OrgSize10 to 19 employees -0.146\*\*\*   
## (0.021)   
##   
## OrgSize10,000 or more employees 0.085\*\*\*   
## (0.015)   
##   
## OrgSize100 to 499 employees -0.052\*\*\*   
## (0.015)   
##   
## OrgSize2 to 9 employees -0.269\*\*\*   
## (0.022)   
##   
## OrgSize20 to 99 employees -0.070\*\*\*   
## (0.015)   
##   
## OrgSize5,000 to 9,999 employees 0.002   
## (0.021)   
##   
## OrgSize500 to 999 employees -0.015   
## (0.019)   
##   
## OrgSizeI donâ€™t know -0.148\*\*\*   
## (0.035)   
##   
## OrgSizeOther -0.087\*\*   
## (0.038)   
##   
## Age25-34 years old 0.125\*\*\*   
## (0.017)   
##   
## Age35-44 years old 0.121\*\*\*   
## (0.020)   
##   
## Age45-54 years old 0.079\*\*\*   
## (0.025)   
##   
## Age55-64 years old -0.025   
## (0.030)   
##   
## Age65 years or older -0.123\*   
## (0.065)   
##   
## GenderWoman -0.042\*\*   
## (0.017)   
##   
## MentalHealth1 -0.032\*\*\*   
## (0.009)   
##   
## log\_yearscode 0.017   
## (0.013)   
##   
## log\_yearscodepro 0.181\*\*\*   
## (0.011)   
##   
## Constant 10.866\*\*\*   
## (0.059)   
##   
## -------------------------------------------------------------------------  
## Observations 5,660   
## R2 0.440   
## Adjusted R2 0.431   
## Residual Std. Error 0.303 (df = 5571)   
## F Statistic 49.809\*\*\* (df = 88; 5571)   
## =========================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Make plots regarding the model

The red line in the residuals and fitted values plot becomes non linear and the residuals tend to be negative from a certain fitted value, 12. I think that this is because the growth of salary amount get smaller and smaller as the salary amount increases. So, from a certain point, there is no linear relationship. This phenomenon is also seen in the scatter plot between professional experience and the amount of compensation.

Moreover, from this plot, I can see that the variance of errors is not still constant. I might want to utilize the robust regression model to make our p-values accurate. Let’s apply that.

plot(model\_log\_log)



The result of this robust regression model should be considered the best accurate model of all. This is because, previous models are likely to suffer from heteroskadasticity and it causes calculation of p-values to be inaccurate. By using the robust model, the problem should be solved. Thus, I use a result from this robust model to draw an answer for my research question.

Log transformed years of coding experience is not statistically significant. On the other hand log transformed years of professional experience is still significant at 1 % level. This tells that professional expericen matters more than the mere coding experience. The coefficient estimate is same as the previous model because we just added white standard errors to the model. By one percent increase in professional coding experience, the yearly compensation increases by 0.181%.

model\_log\_log\_robust <- coeftest(model\_log\_log, vcovHC)  
stargazer(model\_log\_log\_robust,type = "text")

##   
## =========================================================================  
## Dependent variable:   
## ---------------------------  
##   
## -------------------------------------------------------------------------  
## LanguageHaveWorkedWith 0.002   
## (0.002)   
##   
## PlatformHaveWorkedWith 0.023\*\*\*   
## (0.005)   
##   
## WebframeHaveWorkedWith -0.011\*\*\*   
## (0.003)   
##   
## ToolsTechHaveWorkedWith 0.044\*\*\*   
## (0.004)   
##   
## NEWCollabToolsHaveWorkedWith -0.007\*\*\*   
## (0.003)   
##   
## Developer\_back\_end 0.043\*\*\*   
## (0.009)   
##   
## Developer\_front\_end -0.048\*\*\*   
## (0.011)   
##   
## Developer\_full\_stack -0.031\*\*\*   
## (0.010)   
##   
## Engineer\_data 0.049\*\*\*   
## (0.016)   
##   
## Data\_scientist\_or\_machine\_learning\_specialist 0.002   
## (0.023)   
##   
## Developer\_desktop\_or\_enterprise\_applications -0.023\*\*   
## (0.011)   
##   
## Academic\_researcher -0.261\*\*\*   
## (0.047)   
##   
## Database\_administrator -0.024   
## (0.018)   
##   
## System\_administrator -0.073\*\*\*   
## (0.019)   
##   
## Developer\_embedded\_applications\_or\_devices -0.015   
## (0.017)   
##   
## DevOps\_specialist 0.009   
## (0.013)   
##   
## Engineering\_manager 0.108\*\*\*   
## (0.015)   
##   
## Engineer\_site\_reliability 0.060\*\*\*   
## (0.020)   
##   
## Developer\_mobile 0.038\*\*   
## (0.016)   
##   
## Developer\_QA\_or\_test -0.072\*\*\*   
## (0.019)   
##   
## Designer -0.054\*\*\*   
## (0.021)   
##   
## Developer\_game\_or\_graphics -0.001   
## (0.027)   
##   
## Data\_or\_business\_analyst -0.102\*\*\*   
## (0.021)   
##   
## Student -0.079   
## (0.049)   
##   
## Other\_devtypes 0.053\*\*   
## (0.024)   
##   
## Senior\_excecutives 0.134\*\*\*   
## (0.036)   
##   
## EmploymentEmployed part-time -1.009\*\*\*   
## (0.208)   
##   
## EmploymentOther -0.042   
## (0.045)   
##   
## US\_StateArizona 0.155\*\*\*   
## (0.054)   
##   
## US\_StateArkansas 0.008   
## (0.090)   
##   
## US\_StateCalifornia 0.345\*\*\*   
## (0.051)   
##   
## US\_StateColorado 0.199\*\*\*   
## (0.052)   
##   
## US\_StateConnecticut 0.144\*   
## (0.075)   
##   
## US\_StateDistrict of Columbia 0.293\*\*\*   
## (0.107)   
##   
## US\_StateFlorida 0.128\*\*   
## (0.053)   
##   
## US\_StateGeorgia 0.171\*\*\*   
## (0.053)   
##   
## US\_StateIdaho 0.040   
## (0.062)   
##   
## US\_StateIllinois 0.177\*\*\*   
## (0.053)   
##   
## US\_StateIndiana 0.019   
## (0.057)   
##   
## US\_StateIowa 0.044   
## (0.062)   
##   
## US\_StateKansas 0.040   
## (0.068)   
##   
## US\_StateKentucky 0.011   
## (0.060)   
##   
## US\_StateLouisiana 0.103   
## (0.162)   
##   
## US\_StateMaryland 0.193\*\*\*   
## (0.053)   
##   
## US\_StateMassachusetts 0.234\*\*\*   
## (0.053)   
##   
## US\_StateMichigan 0.053   
## (0.054)   
##   
## US\_StateMinnesota 0.105\*\*   
## (0.053)   
##   
## US\_StateMissouri 0.051   
## (0.057)   
##   
## US\_StateNebraska 0.002   
## (0.067)   
##   
## US\_StateNew Hampshire 0.233\*\*\*   
## (0.086)   
##   
## US\_StateNew Jersey 0.207\*\*\*   
## (0.056)   
##   
## US\_StateNew Mexico -0.090   
## (0.076)   
##   
## US\_StateNew York 0.256\*\*\*   
## (0.052)   
##   
## US\_StateNorth Carolina 0.144\*\*\*   
## (0.054)   
##   
## US\_StateOhio 0.044   
## (0.053)   
##   
## US\_StateOklahoma -0.008   
## (0.140)   
##   
## US\_StateOregon 0.128\*\*   
## (0.055)   
##   
## US\_StatePennsylvania 0.090\*   
## (0.053)   
##   
## US\_StateSouth Carolina 0.140\*\*   
## (0.061)   
##   
## US\_StateTennessee 0.058   
## (0.060)   
##   
## US\_StateTexas 0.172\*\*\*   
## (0.051)   
##   
## US\_StateUtah 0.171\*\*\*   
## (0.052)   
##   
## US\_StateVirginia 0.198\*\*\*   
## (0.054)   
##   
## US\_StateWashington 0.369\*\*\*   
## (0.052)   
##   
## US\_StateWisconsin 0.082   
## (0.056)   
##   
## EdLevelBachelor 0.119\*\*\*   
## (0.021)   
##   
## EdLevelCollege study wihtout degree 0.063\*\*\*   
## (0.023)   
##   
## EdLevelDoctorate 0.230\*\*\*   
## (0.034)   
##   
## EdLevelMaster 0.155\*\*\*   
## (0.022)   
##   
## EdLevelSecondary School 0.033   
## (0.042)   
##   
## OrgSize10 to 19 employees -0.146\*\*\*   
## (0.021)   
##   
## OrgSize10,000 or more employees 0.085\*\*\*   
## (0.014)   
##   
## OrgSize100 to 499 employees -0.052\*\*\*   
## (0.014)   
##   
## OrgSize2 to 9 employees -0.269\*\*\*   
## (0.026)   
##   
## OrgSize20 to 99 employees -0.070\*\*\*   
## (0.014)   
##   
## OrgSize5,000 to 9,999 employees 0.002   
## (0.019)   
##   
## OrgSize500 to 999 employees -0.015   
## (0.018)   
##   
## OrgSizeI donâ€™t know -0.148\*\*\*   
## (0.036)   
##   
## OrgSizeOther -0.087   
## (0.056)   
##   
## Age25-34 years old 0.125\*\*\*   
## (0.018)   
##   
## Age35-44 years old 0.121\*\*\*   
## (0.022)   
##   
## Age45-54 years old 0.079\*\*\*   
## (0.026)   
##   
## Age55-64 years old -0.025   
## (0.031)   
##   
## Age65 years or older -0.123\*\*   
## (0.063)   
##   
## GenderWoman -0.042\*\*   
## (0.018)   
##   
## MentalHealth1 -0.032\*\*\*   
## (0.009)   
##   
## log\_yearscode 0.017   
## (0.014)   
##   
## log\_yearscodepro 0.181\*\*\*   
## (0.011)   
##   
## Constant 10.866\*\*\*   
## (0.061)   
##   
## =========================================================================  
## =========================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

model\_log\_log\_robust

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value  
## (Intercept) 10.8656760 0.0609055 178.4022  
## LanguageHaveWorkedWith 0.0017734 0.0019336 0.9171  
## PlatformHaveWorkedWith 0.0230755 0.0048909 4.7181  
## WebframeHaveWorkedWith -0.0112603 0.0030249 -3.7226  
## ToolsTechHaveWorkedWith 0.0437707 0.0035273 12.4090  
## NEWCollabToolsHaveWorkedWith -0.0073474 0.0028420 -2.5853  
## Developer\_back\_end 0.0434086 0.0089123 4.8706  
## Developer\_front\_end -0.0484868 0.0105303 -4.6045  
## Developer\_full\_stack -0.0308636 0.0095961 -3.2163  
## Engineer\_data 0.0492790 0.0161701 3.0475  
## Data\_scientist\_or\_machine\_learning\_specialist 0.0024374 0.0228634 0.1066  
## Developer\_desktop\_or\_enterprise\_applications -0.0227835 0.0113523 -2.0069  
## Academic\_researcher -0.2612929 0.0470996 -5.5477  
## Database\_administrator -0.0243680 0.0181320 -1.3439  
## System\_administrator -0.0726058 0.0192600 -3.7698  
## Developer\_embedded\_applications\_or\_devices -0.0145268 0.0171105 -0.8490  
## DevOps\_specialist 0.0087910 0.0133141 0.6603  
## Engineering\_manager 0.1076838 0.0151163 7.1237  
## Engineer\_site\_reliability 0.0602625 0.0201407 2.9921  
## Developer\_mobile 0.0379337 0.0157885 2.4026  
## Developer\_QA\_or\_test -0.0722382 0.0187573 -3.8512  
## Designer -0.0539044 0.0208687 -2.5830  
## Developer\_game\_or\_graphics -0.0013896 0.0269196 -0.0516  
## Data\_or\_business\_analyst -0.1024059 0.0212501 -4.8191  
## Student -0.0794093 0.0492367 -1.6128  
## Other\_devtypes 0.0534862 0.0238384 2.2437  
## Senior\_excecutives 0.1338612 0.0364672 3.6707  
## EmploymentEmployed part-time -1.0090550 0.2079179 -4.8531  
## EmploymentOther -0.0417331 0.0446511 -0.9346  
## US\_StateArizona 0.1548118 0.0539074 2.8718  
## US\_StateArkansas 0.0081161 0.0895286 0.0907  
## US\_StateCalifornia 0.3453939 0.0505834 6.8282  
## US\_StateColorado 0.1987253 0.0515680 3.8537  
## US\_StateConnecticut 0.1435095 0.0748597 1.9170  
## US\_StateDistrict of Columbia 0.2932277 0.1070609 2.7389  
## US\_StateFlorida 0.1282766 0.0532107 2.4107  
## US\_StateGeorgia 0.1708225 0.0533603 3.2013  
## US\_StateIdaho 0.0402676 0.0621906 0.6475  
## US\_StateIllinois 0.1766763 0.0532428 3.3183  
## US\_StateIndiana 0.0188408 0.0571973 0.3294  
## US\_StateIowa 0.0443056 0.0615900 0.7194  
## US\_StateKansas 0.0403944 0.0677203 0.5965  
## US\_StateKentucky 0.0107904 0.0596081 0.1810  
## US\_StateLouisiana 0.1031955 0.1617944 0.6378  
## US\_StateMaryland 0.1927827 0.0531505 3.6271  
## US\_StateMassachusetts 0.2340442 0.0528138 4.4315  
## US\_StateMichigan 0.0530816 0.0544225 0.9754  
## US\_StateMinnesota 0.1049087 0.0530170 1.9788  
## US\_StateMissouri 0.0514515 0.0568551 0.9050  
## US\_StateNebraska 0.0016424 0.0672747 0.0244  
## US\_StateNew Hampshire 0.2325495 0.0857971 2.7105  
## US\_StateNew Jersey 0.2066031 0.0564799 3.6580  
## US\_StateNew Mexico -0.0904425 0.0762172 -1.1866  
## US\_StateNew York 0.2558938 0.0522377 4.8986  
## US\_StateNorth Carolina 0.1444698 0.0535924 2.6957  
## US\_StateOhio 0.0442474 0.0530934 0.8334  
## US\_StateOklahoma -0.0079997 0.1401430 -0.0571  
## US\_StateOregon 0.1280433 0.0549188 2.3315  
## US\_StatePennsylvania 0.0896670 0.0531187 1.6881  
## US\_StateSouth Carolina 0.1403195 0.0610340 2.2990  
## US\_StateTennessee 0.0582446 0.0595590 0.9779  
## US\_StateTexas 0.1716561 0.0509155 3.3714  
## US\_StateUtah 0.1714099 0.0523027 3.2773  
## US\_StateVirginia 0.1978435 0.0539036 3.6703  
## US\_StateWashington 0.3694381 0.0517896 7.1334  
## US\_StateWisconsin 0.0820275 0.0564649 1.4527  
## EdLevelBachelor 0.1193650 0.0207470 5.7534  
## EdLevelCollege study wihtout degree 0.0627821 0.0232860 2.6961  
## EdLevelDoctorate 0.2295843 0.0339706 6.7583  
## EdLevelMaster 0.1545685 0.0224814 6.8754  
## EdLevelSecondary School 0.0325892 0.0422416 0.7715  
## OrgSize10 to 19 employees -0.1459525 0.0213737 -6.8286  
## OrgSize10,000 or more employees 0.0848988 0.0143035 5.9355  
## OrgSize100 to 499 employees -0.0520023 0.0139057 -3.7396  
## OrgSize2 to 9 employees -0.2689518 0.0260976 -10.3056  
## OrgSize20 to 99 employees -0.0697179 0.0144749 -4.8165  
## OrgSize5,000 to 9,999 employees 0.0021539 0.0192936 0.1116  
## OrgSize500 to 999 employees -0.0154292 0.0176681 -0.8733  
## OrgSizeI donâ\200\231t know -0.1481179 0.0363381 -4.0761  
## OrgSizeOther -0.0871713 0.0563617 -1.5466  
## Age25-34 years old 0.1254822 0.0184905 6.7863  
## Age35-44 years old 0.1211827 0.0221777 5.4642  
## Age45-54 years old 0.0789821 0.0260621 3.0305  
## Age55-64 years old -0.0248485 0.0311227 -0.7984  
## Age65 years or older -0.1230316 0.0627468 -1.9608  
## GenderWoman -0.0420788 0.0177091 -2.3761  
## MentalHealth1 -0.0321157 0.0088842 -3.6149  
## log\_yearscode 0.0171113 0.0136437 1.2542  
## log\_yearscodepro 0.1810530 0.0112921 16.0336  
## Pr(>|t|)   
## (Intercept) < 2.2e-16 \*\*\*  
## LanguageHaveWorkedWith 0.3591194   
## PlatformHaveWorkedWith 2.439e-06 \*\*\*  
## WebframeHaveWorkedWith 0.0001991 \*\*\*  
## ToolsTechHaveWorkedWith < 2.2e-16 \*\*\*  
## NEWCollabToolsHaveWorkedWith 0.0097543 \*\*   
## Developer\_back\_end 1.143e-06 \*\*\*  
## Developer\_front\_end 4.226e-06 \*\*\*  
## Developer\_full\_stack 0.0013061 \*\*   
## Engineer\_data 0.0023180 \*\*   
## Data\_scientist\_or\_machine\_learning\_specialist 0.9151059   
## Developer\_desktop\_or\_enterprise\_applications 0.0448035 \*   
## Academic\_researcher 3.028e-08 \*\*\*  
## Database\_administrator 0.1790275   
## System\_administrator 0.0001651 \*\*\*  
## Developer\_embedded\_applications\_or\_devices 0.3959155   
## DevOps\_specialist 0.5091003   
## Engineering\_manager 1.184e-12 \*\*\*  
## Engineer\_site\_reliability 0.0027830 \*\*   
## Developer\_mobile 0.0163108 \*   
## Developer\_QA\_or\_test 0.0001189 \*\*\*  
## Designer 0.0098189 \*\*   
## Developer\_game\_or\_graphics 0.9588343   
## Data\_or\_business\_analyst 1.481e-06 \*\*\*  
## Student 0.1068432   
## Other\_devtypes 0.0248906 \*   
## Senior\_excecutives 0.0002441 \*\*\*  
## EmploymentEmployed part-time 1.248e-06 \*\*\*  
## EmploymentOther 0.3500096   
## US\_StateArizona 0.0040967 \*\*   
## US\_StateArkansas 0.9277710   
## US\_StateCalifornia 9.513e-12 \*\*\*  
## US\_StateColorado 0.0001177 \*\*\*  
## US\_StateConnecticut 0.0552829 .   
## US\_StateDistrict of Columbia 0.0061844 \*\*   
## US\_StateFlorida 0.0159529 \*   
## US\_StateGeorgia 0.0013758 \*\*   
## US\_StateIdaho 0.5173441   
## US\_StateIllinois 0.0009114 \*\*\*  
## US\_StateIndiana 0.7418653   
## US\_StateIowa 0.4719469   
## US\_StateKansas 0.5508730   
## US\_StateKentucky 0.8563572   
## US\_StateLouisiana 0.5236181   
## US\_StateMaryland 0.0002892 \*\*\*  
## US\_StateMassachusetts 9.538e-06 \*\*\*  
## US\_StateMichigan 0.3294233   
## US\_StateMinnesota 0.0478908 \*   
## US\_StateMissouri 0.3655262   
## US\_StateNebraska 0.9805238   
## US\_StateNew Hampshire 0.0067396 \*\*   
## US\_StateNew Jersey 0.0002565 \*\*\*  
## US\_StateNew Mexico 0.2354200   
## US\_StateNew York 9.923e-07 \*\*\*  
## US\_StateNorth Carolina 0.0070449 \*\*   
## US\_StateOhio 0.4046611   
## US\_StateOklahoma 0.9544815   
## US\_StateOregon 0.0197624 \*   
## US\_StatePennsylvania 0.0914576 .   
## US\_StateSouth Carolina 0.0215395 \*   
## US\_StateTennessee 0.3281504   
## US\_StateTexas 0.0007530 \*\*\*  
## US\_StateUtah 0.0010546 \*\*   
## US\_StateVirginia 0.0002445 \*\*\*  
## US\_StateWashington 1.104e-12 \*\*\*  
## US\_StateWisconsin 0.1463583   
## EdLevelBachelor 9.215e-09 \*\*\*  
## EdLevelCollege study wihtout degree 0.0070361 \*\*   
## EdLevelDoctorate 1.538e-11 \*\*\*  
## EdLevelMaster 6.859e-12 \*\*\*  
## EdLevelSecondary School 0.4404459   
## OrgSize10 to 19 employees 9.488e-12 \*\*\*  
## OrgSize10,000 or more employees 3.106e-09 \*\*\*  
## OrgSize100 to 499 employees 0.0001861 \*\*\*  
## OrgSize2 to 9 employees < 2.2e-16 \*\*\*  
## OrgSize20 to 99 employees 1.500e-06 \*\*\*  
## OrgSize5,000 to 9,999 employees 0.9111163   
## OrgSize500 to 999 employees 0.3825470   
## OrgSizeI donâ\200\231t know 4.643e-05 \*\*\*  
## OrgSizeOther 0.1220070   
## Age25-34 years old 1.270e-11 \*\*\*  
## Age35-44 years old 4.853e-08 \*\*\*  
## Age45-54 years old 0.0024524 \*\*   
## Age55-64 years old 0.4246696   
## Age65 years or older 0.0499563 \*   
## GenderWoman 0.0175296 \*   
## MentalHealth1 0.0003031 \*\*\*  
## log\_yearscode 0.2098374   
## log\_yearscodepro < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##browseURL('model\_log\_log\_robust.doc')

## Result

Through this project, I found that professional experience of coding positively influence the yearly amount of compensation. For each one percent increase in the professional experience of coding, there is an increase of 0.181% in the compensation amount. This is statistically significant at the 1% level. On the other hand, years of mere coding experience do not have statistically significant influence on the amount.

However, it is observed that there is a non linear relationship between years of professional experience and the compensation amount from a certain point of experience. Moreover, it was seen that the linear model overestimates the compensation when it estimated large compensation amounts, amounts over 12 in the log scale.

## Implications

For future researches, I would recommend to create a model that explains that non-linearity.

## Conclusion

My research question is how does years of experience of developers affect their salaries in the U.S.?

Through this project, I was able to find the significance of professional experience of coding on the yearly compensation amount of developers in the U.S. However, the interpretation about the effect on the compensation was only applied to developers who make less than 12 compensation in log scale. This was due to the non-linearities I discussed. Moreover, I also found out that years of coding experience does not significantly affect the compensation.

The theory I established before the project actually matched with what I observed in this project. There was a linearity in the compensation by a certain point, but from that point the amount stays same or decreases. Although it is hard to know what causes the change, one possible explanation is that when developers reach that point, they are somewhat old and start taking less pad jobs that allow them to have more personal time.

## References

Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL <https://CRAN.R-project.org/doc/Rnews/>

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2020). rmarkdown: Dynamic Documents for R. <https://github.com/rstudio/rmarkdown>

Dash, M., Bakshi, S., & Chugh, A. (2017, Nov). The Relationship Between Work Experience and Employee Compensation: A Case Study of The Indian IT Industry, Journal of Applied Management and Investment, vol 6(1), 5-10. Retrieved from <https://econpapers.repec.org/article/odsjournl/v_3a6_3ay_3a2017_3ai_3a1_3ap_3a5-10.htm>

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. <https://CRAN.R-project.org/package=dplyr>

Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. <https://CRAN.R-project.org/package=stringr>

Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. <https://CRAN.R-project.org/package=stargazer>

H. Wickham. (2016, n.d.). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag

Yihui Xie (2021, n.d.). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.37.

Max Kuhn (2022). caret: Classification and Regression Training. R package version 6.0-92. <https://CRAN.R-project.org/package=caret>

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>

Salary.com. (n.d.). Eight Factors That Can Affect Your Pay. Retrieved from <https://www.salary.com/articles/eight-factors-that-can-affect-your-pay/>

Sauro, J. (2014, Nov). User Experience Salaries & Calculator (2014). Measuring U. Retrieved from <https://measuringu.com/salary-survey2014/>

Stefan Milton Bache and Hadley Wickham (2020). magrittr: A Forward-Pipe Operator for R. R package version 2.0.1. <https://CRAN.R-project.org/package=magrittr>

Thomas Lumley based on Fortran code by Alan Miller (2020). leaps: Regression Subset Selection. R package version 3.1. <https://CRAN.R-project.org/package=leaps>

Torpey, E. (2015, May). Same occupation, different pay: How wages vary. U.S. BUREAU OF LABOR STATISTICS. Retrieved from <https://www.bls.gov/careeroutlook/2015/article/wage-differences.htm>

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. In Journal of Statistical Software (Vol. 40, Issue 1, pp. 1–29). <http://www.jstatsoft.org/v40/i01/>

Zeileis A, Köll S, Graham N (2020). “Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R.” *Journal of Statistical Software*, *95*(1), 1-36. doi: 10.18637/jss.v095.i01 (URL: <https://doi.org/10.18637/jss.v095.i01)citation>()