Impact of Age On Academic Achievement

Yujing Fang, Chang Lian, Zhengdao Song

July 1, 2016

rm(list=ls(all=TRUE)) # Load libraries

library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

# Load raw data

# Read all raw data  
data <- read.csv("result/reshaped\_data.csv", stringsAsFactors = F)

# Exploratory Analysis

# numbers of variables and numbers of observations  
dim(data)

## [1] 186 18

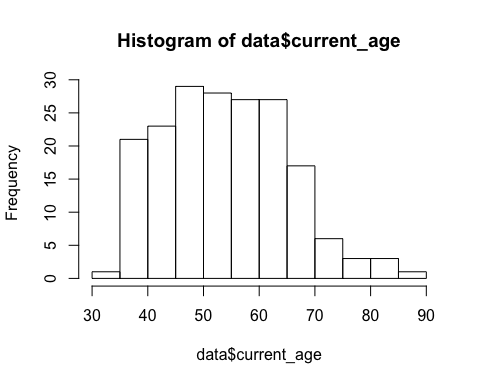
# structure of variables  
str(data)

## 'data.frame': 186 obs. of 18 variables:  
## $ first\_name : chr "james" "paul" "elaine" "anjon" ...  
## $ last\_name : chr "thomson" "ahlquist" "alarid" "audhya" ...  
## $ gender : chr "male" "male" "female" "male" ...  
## $ current\_age : int 58 63 54 42 60 65 56 44 52 64 ...  
## $ phd\_age : int 30 28 29 28 30 28 27 31 30 28 ...  
## $ current\_university : chr "university of wisconsin-madison" "university of wisconsin-madison" "university of wisconsin-madison" "university of wisconsin-madison" ...  
## $ current\_university\_rank : int 18 18 18 18 18 18 18 18 18 18 ...  
## $ current\_university\_tier : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ bs\_university : chr "university of illinois-urban champaign" "iowa state university" "university of california-berkeley" "brown university" ...  
## $ phd\_university : chr "university of pennsylvania" "university of wisconsin-madison" "university of california-berkeley" "university of california-san diego" ...  
## $ year\_get\_bs : int 1981 1976 1985 1997 1979 1974 1983 1995 1987 1975 ...  
## $ year\_get\_phd : int 1988 1981 1991 2002 1986 1979 1987 2003 1994 1980 ...  
## $ num\_articles : int 277 71 26 62 24 87 44 77 35 65 ...  
## $ num\_peered\_reviewed\_articles : int 14 2 0 1 0 4 0 0 0 1 ...  
## $ percent\_peered\_reviewed\_articles: num 0.0505 0.0282 0 0.0161 0 0.046 0 0 0 0.0154 ...  
## $ num\_nature : int 9 1 0 0 0 2 0 0 0 0 ...  
## $ num\_cell : int 3 0 0 1 0 1 0 0 0 0 ...  
## $ num\_science : int 2 1 0 0 0 1 0 0 0 1 ...

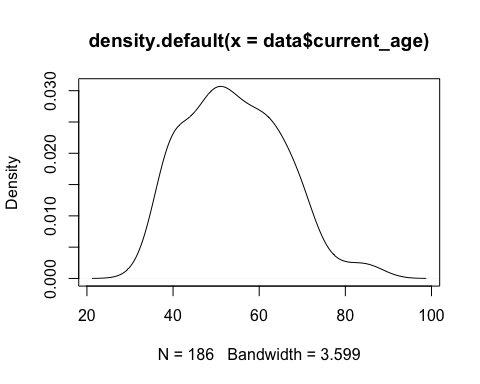
# Descriptive statistics  
summary(data)

## first\_name last\_name gender current\_age   
## Length:186 Length:186 Length:186 Min. :32.00   
## Class :character Class :character Class :character 1st Qu.:46.00   
## Mode :character Mode :character Mode :character Median :53.00   
## Mean :54.28   
## 3rd Qu.:63.00   
## Max. :88.00   
##   
## phd\_age current\_university current\_university\_rank  
## Min. :25.0 Length:186 Min. : 1.00   
## 1st Qu.:28.0 Class :character 1st Qu.: 4.00   
## Median :29.0 Mode :character Median :18.00   
## Mean :29.2 Mean :17.65   
## 3rd Qu.:30.0 3rd Qu.:19.00   
## Max. :37.0 Max. :55.00   
##   
## current\_university\_tier bs\_university phd\_university   
## Min. :1.000 Length:186 Length:186   
## 1st Qu.:1.000 Class :character Class :character   
## Median :2.000 Mode :character Mode :character   
## Mean :1.769   
## 3rd Qu.:2.000   
## Max. :3.000   
##   
## year\_get\_bs year\_get\_phd num\_articles   
## Min. :1951 Min. :1957 Min. : 0.00   
## 1st Qu.:1976 1st Qu.:1982 1st Qu.: 24.00   
## Median :1986 Median :1992 Median : 50.50   
## Mean :1985 Mean :1991 Mean : 86.26   
## 3rd Qu.:1993 3rd Qu.:2000 3rd Qu.: 98.75   
## Max. :2007 Max. :2012 Max. :698.00   
##   
## num\_peered\_reviewed\_articles percent\_peered\_reviewed\_articles  
## Min. : 0.000 Min. :0.00000   
## 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 1.000 Median :0.02840   
## Mean : 7.301 Mean :0.06084   
## 3rd Qu.: 6.000 3rd Qu.:0.08947   
## Max. :152.000 Max. :0.36170   
## NA's :2   
## num\_nature num\_cell num\_science   
## Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 0.000 Median : 0.000 Median : 0.000   
## Mean : 2.753 Mean : 2.441 Mean : 2.108   
## 3rd Qu.: 2.000 3rd Qu.: 1.000 3rd Qu.: 2.000   
## Max. :74.000 Max. :48.000 Max. :42.000   
##

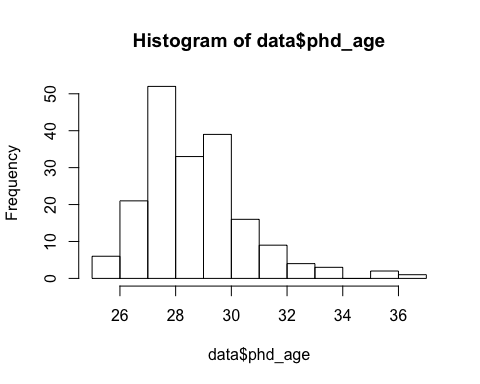
# Histogram  
hist(data$current\_age)



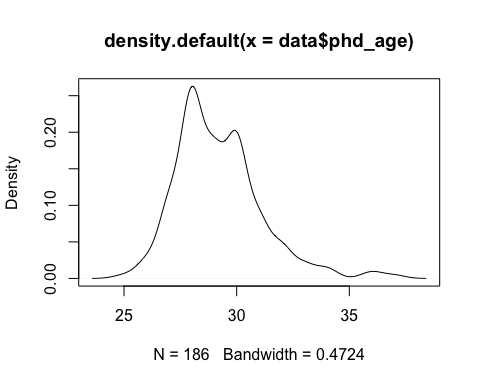
plot(density(data$current\_age)) #kernel density estimate



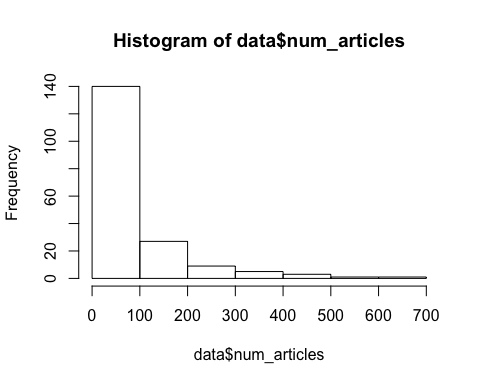
hist(data$phd\_age)



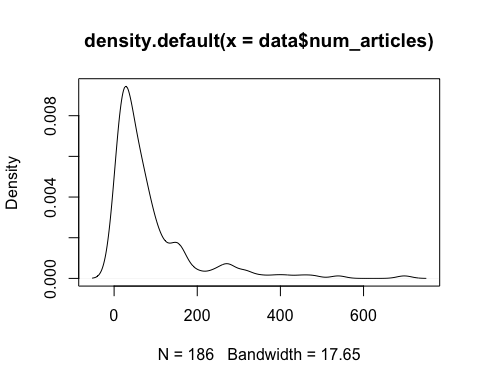
plot(density(data$phd\_age))



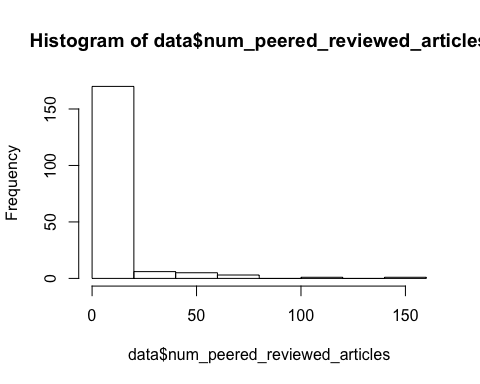
hist(data$num\_articles)



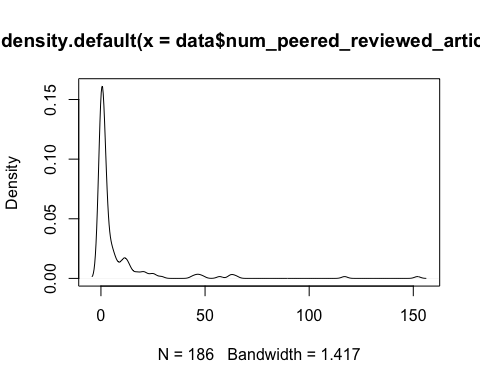
plot(density(data$num\_articles))



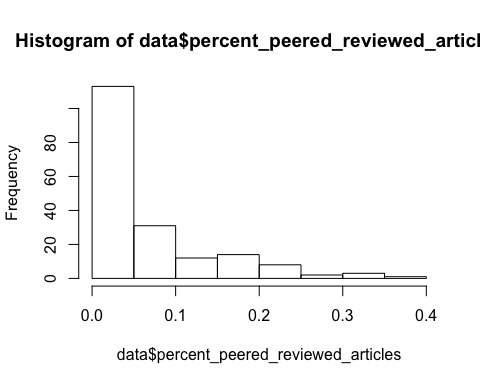
hist(data$num\_peered\_reviewed\_articles)



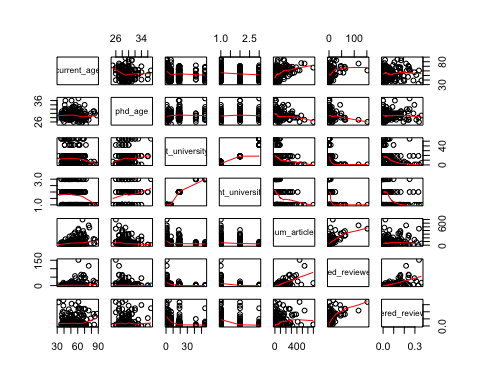
plot(density(data$num\_peered\_reviewed\_articles))



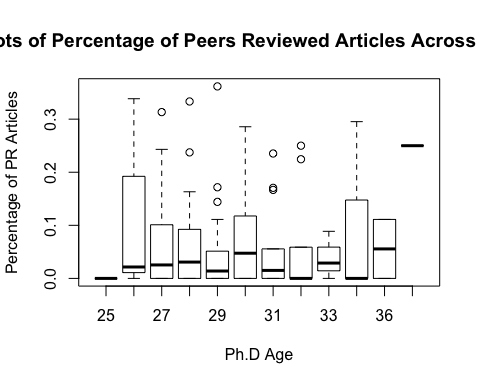
hist(data$percent\_peered\_reviewed\_articles)



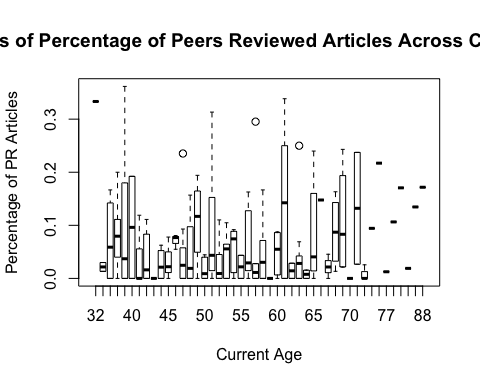
# Create scatterplots for numeric metircs  
data\_numeric <- select(data, current\_age, phd\_age, current\_university\_rank, current\_university\_tier,num\_articles, num\_peered\_reviewed\_articles, percent\_peered\_reviewed\_articles )  
  
pairs(data\_numeric,panel=panel.smooth)

  
 # Percentage of Peers Reviewed Articles ~ Ages Compared "centers" and "spreads" of Percentage of PR Articles across Ph.D Ages

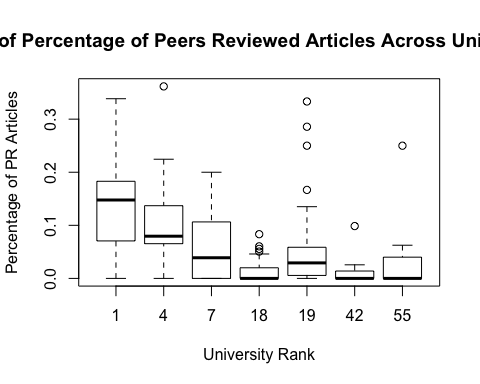
boxplot(percent\_peered\_reviewed\_articles ~ phd\_age, data=data, main="Boxplots of Percentage of Peers Reviewed Articles Across Ph.D Ages", xlab="Ph.D Age", ylab="Percentage of PR Articles")

  
 Compared "centers" and "spreads" of Percentage of PR Articles across Current Ages

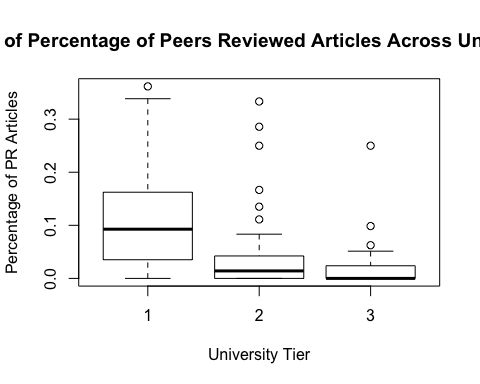
boxplot(percent\_peered\_reviewed\_articles ~ current\_age, data=data, main="Boxplots of Percentage of Peers Reviewed Articles Across Current Ages", xlab="Current Age", ylab="Percentage of PR Articles")

  
 # Percentage of Peers Reviewed Articles ~ Universities Compared "centers" and "spreads" of Percentage of PR Articles across University Ranks

boxplot(percent\_peered\_reviewed\_articles ~ current\_university\_rank, data=data, main="Boxplots of Percentage of Peers Reviewed Articles Across University Ranks", xlab="University Rank", ylab="Percentage of PR Articles")

  
 Compared "centers" and "spreads" of Percentage of PR Articles across University Tiers

boxplot(percent\_peered\_reviewed\_articles ~ current\_university\_tier, data=data, main="Boxplots of Percentage of Peers Reviewed Articles Across University Tiers", xlab="University Tier", ylab="Percentage of PR Articles")

  
 # Linear Regression for num\_articles num\_articles ~ current\_age, phd\_age, current\_university\_ranks

rg\_articles <- lm(num\_articles ~ current\_age + phd\_age + current\_university\_rank, data = data)  
summary(rg\_articles)

##   
## Call:  
## lm(formula = num\_articles ~ current\_age + phd\_age + current\_university\_rank,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -193.84 -51.05 -16.75 21.98 515.04   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 223.6371 111.5189 2.005 0.046405 \*   
## current\_age 3.0714 0.6041 5.084 9.12e-07 \*\*\*  
## phd\_age -9.4158 3.5648 -2.641 0.008976 \*\*   
## current\_university\_rank -1.6501 0.4296 -3.841 0.000169 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 91.79 on 182 degrees of freedom  
## Multiple R-squared: 0.2509, Adjusted R-squared: 0.2386   
## F-statistic: 20.32 on 3 and 182 DF, p-value: 2.1e-11

# Linear Regression for log(num\_articles)

log(num\_articles) ~ current\_age, phd\_age, current\_university\_ranks

# Create a replica of 'data' data frame for adding log(num\_article)  
data\_log\_num\_articles <- data  
data\_log\_num\_articles$log\_num\_articles <- log(data\_log\_num\_articles$num\_articles)  
  
# Drop -Inf  
data\_log\_num\_articles <- subset(data\_log\_num\_articles, log\_num\_articles!= "-Inf")  
  
# Run regression  
rg\_articles\_log <- lm(log\_num\_articles ~ current\_age + phd\_age + current\_university\_rank, data = data\_log\_num\_articles)  
summary(rg\_articles\_log)

##   
## Call:  
## lm(formula = log\_num\_articles ~ current\_age + phd\_age + current\_university\_rank,   
## data = data\_log\_num\_articles)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9126 -0.5456 0.0126 0.5767 2.1265   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.792069 1.183636 5.738 3.98e-08 \*\*\*  
## current\_age 0.033749 0.006472 5.215 5.02e-07 \*\*\*  
## phd\_age -0.152472 0.037970 -4.016 8.69e-05 \*\*\*  
## current\_university\_rank -0.015894 0.004704 -3.379 0.000893 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.973 on 180 degrees of freedom  
## Multiple R-squared: 0.2792, Adjusted R-squared: 0.2671   
## F-statistic: 23.24 on 3 and 180 DF, p-value: 9.251e-13

# Linear Regression for num\_peered\_reviewed\_articles

num\_peered\_reviewed\_articles ~ current\_age, phd\_age, current\_university\_ranks

rg\_peers\_articles <- lm(num\_peered\_reviewed\_articles ~ current\_age + phd\_age + current\_university\_rank, data = data)  
summary(rg\_peers\_articles)

##   
## Call:  
## lm(formula = num\_peered\_reviewed\_articles ~ current\_age + phd\_age +   
## current\_university\_rank, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.275 -6.978 -3.163 2.372 132.938   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.1492 19.8344 1.570 0.11804   
## current\_age 0.3505 0.1074 3.262 0.00132 \*\*   
## phd\_age -1.2750 0.6340 -2.011 0.04581 \*   
## current\_university\_rank -0.3196 0.0764 -4.184 4.46e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16.33 on 182 degrees of freedom  
## Multiple R-squared: 0.1851, Adjusted R-squared: 0.1716   
## F-statistic: 13.78 on 3 and 182 DF, p-value: 3.889e-08

# Linear Regression for log(num\_peered\_reviewed\_articles)

log(num\_peered\_reviewed\_articles) ~ current\_age, phd\_age, current\_university\_ranks

# Create a replica of 'data' data frame for adding log(num\_article)  
data\_log\_peers\_articles <- data  
data\_log\_peers\_articles$log\_num\_peered\_reviewed\_articles <- log(data\_log\_peers\_articles$num\_peered\_reviewed\_articles)  
  
# Drop -Inf  
data\_log\_peers\_articles <- subset(data\_log\_peers\_articles, log\_num\_peered\_reviewed\_articles!= "-Inf")  
  
# Run regression  
rg\_peers\_articles\_log <- lm(log\_num\_peered\_reviewed\_articles ~ current\_age + phd\_age + current\_university\_rank, data = data\_log\_peers\_articles)  
summary(rg\_peers\_articles\_log)

##   
## Call:  
## lm(formula = log\_num\_peered\_reviewed\_articles ~ current\_age +   
## phd\_age + current\_university\_rank, data = data\_log\_peers\_articles)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.42215 -0.83286 0.03809 0.71007 2.53055   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.130467 1.602115 1.954 0.053134 .   
## current\_age 0.028019 0.008244 3.398 0.000932 \*\*\*  
## phd\_age -0.088490 0.052382 -1.689 0.093866 .   
## current\_university\_rank -0.045526 0.006713 -6.782 5.41e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.051 on 115 degrees of freedom  
## Multiple R-squared: 0.3792, Adjusted R-squared: 0.363   
## F-statistic: 23.42 on 3 and 115 DF, p-value: 6.671e-12

# Linear Regression for percent\_peered\_reviewed\_articles

percent\_peered\_reviewed\_articles ~ current\_age, phd\_age, current\_university\_ranks

rg\_percent\_peers <- lm(percent\_peered\_reviewed\_articles ~ current\_age + phd\_age + current\_university\_rank, data = data)  
summary(rg\_percent\_peers)

##   
## Call:  
## lm(formula = percent\_peered\_reviewed\_articles ~ current\_age +   
## phd\_age + current\_university\_rank, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.09817 -0.05548 -0.01404 0.02354 0.27772   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.298e-02 9.145e-02 0.142 0.887   
## current\_age -8.261e-05 5.000e-04 -0.165 0.869   
## phd\_age 2.970e-03 2.934e-03 1.012 0.313   
## current\_university\_rank -1.995e-03 3.635e-04 -5.490 1.35e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.07518 on 180 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.1463, Adjusted R-squared: 0.1321   
## F-statistic: 10.28 on 3 and 180 DF, p-value: 2.777e-06

# Linear Regression for log(percent\_peered\_reviewed\_articles)

log(percent\_peered\_reviewed\_articles) ~ current\_age, phd\_age, current\_university\_ranks

# Create a replica of 'data' data frame for adding log(num\_article)  
data\_log\_percent\_peers <- data  
data\_log\_percent\_peers$log\_percent\_peered\_reviewed\_articles <- log(data\_log\_percent\_peers$percent\_peered\_reviewed\_articles)  
  
# Drop -Inf & NaN  
data\_log\_percent\_peers <- subset(data\_log\_percent\_peers, log\_percent\_peered\_reviewed\_articles!= "-Inf")  
data\_log\_percent\_peers <- subset(data\_log\_percent\_peers, log\_percent\_peered\_reviewed\_articles!= "NaN")  
  
# Run regression  
rg\_percent\_peers\_log <- lm(log\_percent\_peered\_reviewed\_articles ~ current\_age + phd\_age + current\_university\_rank, data = data\_log\_percent\_peers)  
summary(rg\_percent\_peers\_log)

##   
## Call:  
## lm(formula = log\_percent\_peered\_reviewed\_articles ~ current\_age +   
## phd\_age + current\_university\_rank, data = data\_log\_percent\_peers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.38652 -0.66105 0.07245 0.64665 1.81650   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.792155 1.308400 -3.663 0.000379 \*\*\*  
## current\_age -0.008371 0.006733 -1.243 0.216281   
## phd\_age 0.099016 0.042779 2.315 0.022410 \*   
## current\_university\_rank -0.028429 0.005482 -5.185 9.34e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8581 on 115 degrees of freedom  
## Multiple R-squared: 0.2098, Adjusted R-squared: 0.1892   
## F-statistic: 10.18 on 3 and 115 DF, p-value: 5.365e-06