Analysis on Factors That Affect the Salary of Undergraduates

Li Cao

Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong Province, 519000, China

**Corresponding author. Email:Vivian.wang@cas-harbour.org*

ABSTRACT

This paper uses linear regression to analyze the relationship between graduate students and their incomes. The dependent valuable indexes are the school from which they graduated, the student/teacher ratio, which city the school is located in, and their CET4/6 grades. By using R to modification the regression, the author detects the existence of multi-collinearity and promote the regression. In order to give graduates a direction in choosing a school and allow them to make good use of their time during school, this paper is proposed to give some suggestions.

Keywords: *Linear regression, graduate students, income*

1. INTRODUCTION

Through higher education, people can obtain more and much better employment opportunities in the vocational field and achieve correspondingly higher social status. The employment of college graduates has been widely concerned by the society, and the salary of college graduates is an important indicator to reflect the quality of China's higher education as well as the success of college students' own development. Therefore, the factors that exactly affect the salary of the college graduates need to be learned. By reading some papers and asking the relative teacher, the author found that the schools of the graduates, locations of 985, 211 (985: 50 universities which has the world advanced level. 211: focusing on building about 100 universities and a number of key disciplines) schools, college entrance examination scores, other objective conditions and subjective factors that are intertwined with each other, have different effects on the salary of college graduates. 80 samples are collected to analyze the impact of these factors on salary, and a multiple linear regression model is established.

By finishing this paper, the author expects to provide certain reference for the follow-up work of studying the employment status of college graduates, also giving suggestions for the high school students to know what kind of college is the best choice for their future life, meanwhile giving advice on how to make meaningful and valuable decisions with the use of college life to improve ourselves.

2. LITERATURE REVIEW

From some investigations, it is obvious to see that if the years of education increases one year, it can lead to the increase of personal income, in order to evaluate the efficiency of education investment and reflect the contribution made by the education to the economic development. As Chinese education popularization becomes larger, and the rapid development of marketization, a large number of studies have found that education in China has yielded distinctive rising trend annually since the 1990s. Junsen Zhang and Yao-hui Zhao's research shows that from 1988 to 1999 in the town, education yielded an increase from the original 4.7% to 11.5%. Later Li Shi and Ding Sai's research shows, the rate rose up from 2.43% in 1990 to 8.10% in 1999. In addition, Florence, Justin Chen Welding and Summul's research indicates that in 1991 it increased from 2.953% up to 8.534% in 2000[5]. The researches all indicate that China's urban education yields significant growth, reflecting the education in China's special economic system which has a positive effect to the laborer. The researches also show that the education plays an important parts in deciding the income of the graduate.

From USES-the survey data on the Beijing graduates, the statistical description and empirical studies of regression, the researchers verified the following point of human capital theory: the education, work and enterprise qualifications, work flow and other factors, as the important human capital investment are very significant for graduates salary[4]. It can be co-efficiently seen from the standard that education has the most positive influence on the starting salary of graduates at the beginning of their employment.

3. ECONOMETRIC MODELS AND ESTIMATION METHODS

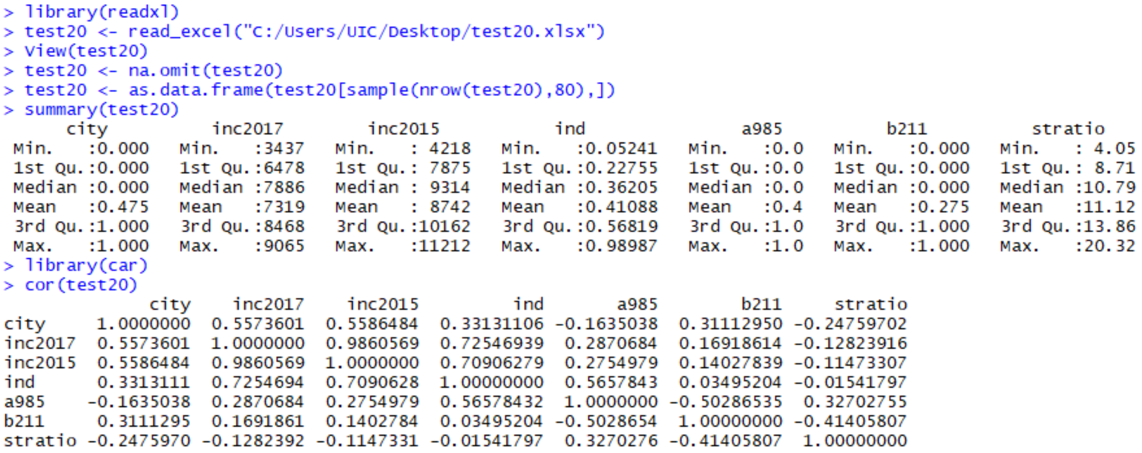
3.1. Statistics

The data mainly come from official websites of various schools, some online newspapers, journals and websites[1,2,3]. The author chooses several dependent variables to conduct the analysis. Whether the school belonging to 985 or 211 universities exists as a dummy variable. What is more, the income of every school, which can refer to the financial status of the university, is also selected as it can reflect the size and economic strength of the university. However, the author removes the income valuable when do the regression, because the income of the school was highly related to the investment in education at first. And after that the author did some further reading of the other news, the initial idea was changed as the relationship between income and investment on education was not as high as assumed. Therefore, the author removes it before conducting the regression.

The author chooses admission score total index as a dependent variable. The total index of college entrance examination score represents a student's own quality and exists as a proxy variable. The location of universities reflects the influence of regional location. The salary of college graduates in 2015 is a lagging variable to explore whether the past value of income itself will affect the current income level.

Table 1 Variables

Dependent variable	(inc2017) income of 2017 graduate students
Independent variables	inc2015(income of 2015 graduates students) ind (Admission score total index) a985(1=this collage is 985, 0=this collage is neither 985 nor 211) b211(1=this collage is 211, 0=this collage is neither 985 nor 211) Stratio (Student-faculty ratios) City(1= These cities are first-tier cities, 0= These cities are not first-tier cities)

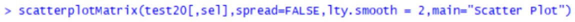


```
> library(readxl)
> test20 <- read_excel("C:/Users/UIC/Desktop/test20.xlsx")
> View(test20)
> test20 <- na.omit(test20)
> test20 <- as.data.frame(test20[sample(nrow(test20),80),])
> summary(test20)
      city          inc2017         inc2015           ind               a985          b211            stratio     
 Min.   :0.000   Min.   :3437   Min.   : 4218   Min.   :0.05241   Min.   :0.0   Min.   :0.000   Min.   : 4.05  
 1st Qu.:0.000   1st Qu.:6478   1st Qu.: 7875   1st Qu.:0.22755   1st Qu.:0.0   1st Qu.:0.000   1st Qu.: 8.71  
 Median :0.000   Median :7886   Median : 9314   Median :0.36205   Median :0.0   Median :0.000   Median :10.79  
 Mean   :0.475   Mean   :7319   Mean   : 8742   Mean   :0.41088   Mean   :0.4   Mean   :0.275   Mean   :11.12  
 3rd Qu.:1.000   3rd Qu.:8468   3rd Qu.:10162   3rd Qu.:0.56819   3rd Qu.:1.0   3rd Qu.:1.000   3rd Qu.:13.86  
 Max.   :1.000   Max.   :9065   Max.   :11212   Max.   :0.98987   Max.   :1.0   Max.   :1.000   Max.   :20.32  
> library(car)
> cor(test20)
              city    inc2017    inc2015        ind       a985        b211     stratio
city     1.0000000  0.5573601  0.5586484 0.33131106 -0.1635038  0.31112950 -0.24759702
inc2017  0.5573601  1.0000000  0.9860569 0.72546939  0.2870684  0.16918614 -0.12823916
inc2015  0.5586484  0.9860569  1.0000000 0.70906279  0.2754979  0.14027839 -0.11473307
ind      0.3313111  0.7254694  0.7090628 1.00000000  0.5657843  0.03495204 -0.01541797
a985    -0.1635038  0.2870684  0.2754979 0.56578432  1.0000000 -0.50286535  0.32702755
b211     0.3111295  0.1691861  0.1402784 0.03495204 -0.5028654  1.00000000 -0.41405807
stratio -0.2475970 -0.1282392 -0.1147331 -0.01541797  0.3270276 -0.41405807  1.00000000
```

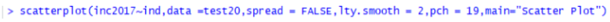
Figure 1 Correlation test

3.1.1. Plot

There is the scatter plot between independent variable & dependent variable.



```
> scatterplotMatrix(test20[,sel],spread=FALSE,lty.smooth = 2,main="Scatter Plot")
```



```
> scatterplot(inc2017~ind,data =test20,spread = FALSE,lty.smooth = 2,pch = 19,main="Scatter Plot")
```

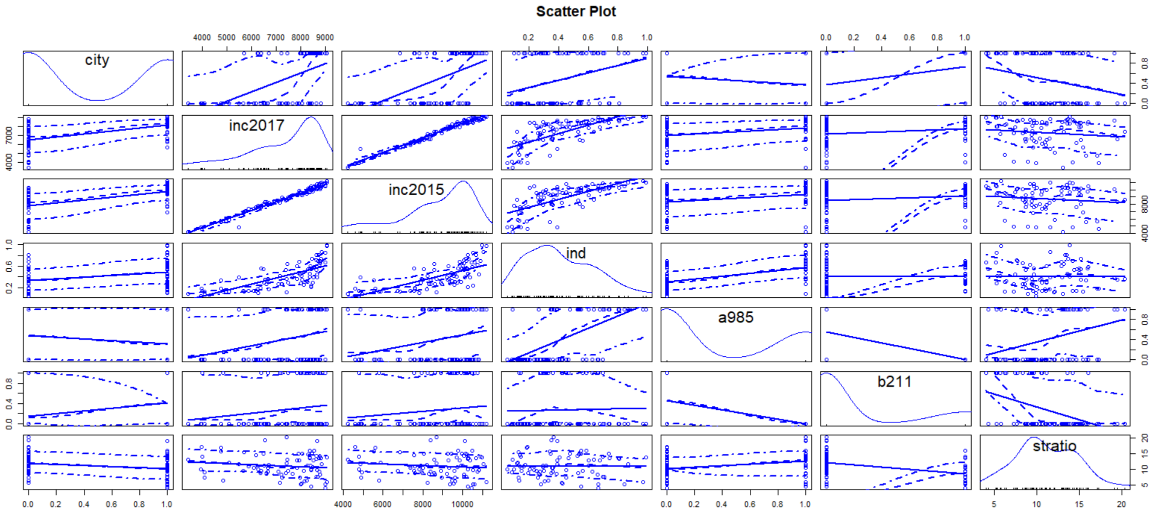


Figure 2 Distribution of dependent variable

*> scatterplot(inc2017~ind, data = test20, spread = FALSE, lty.smooth = 2, pch = 19, main ="Scatter Plot

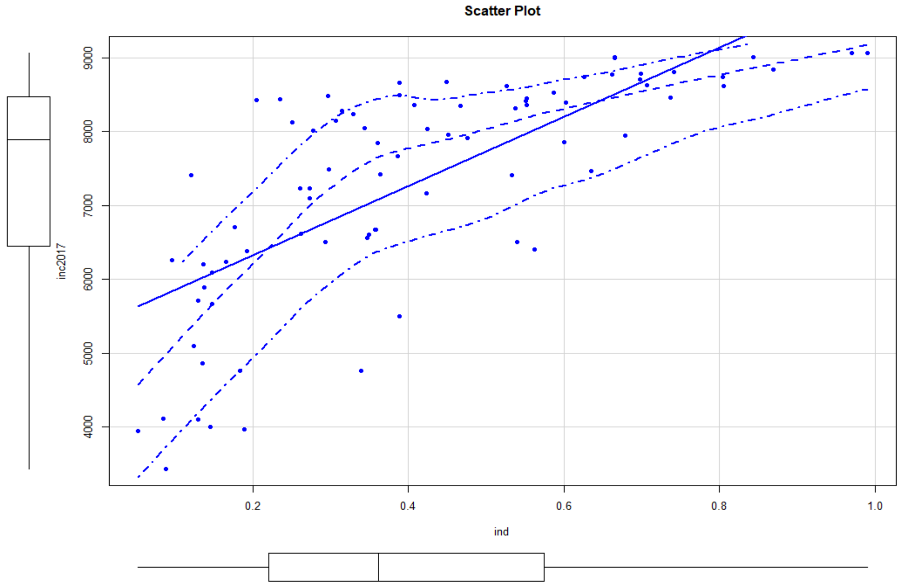
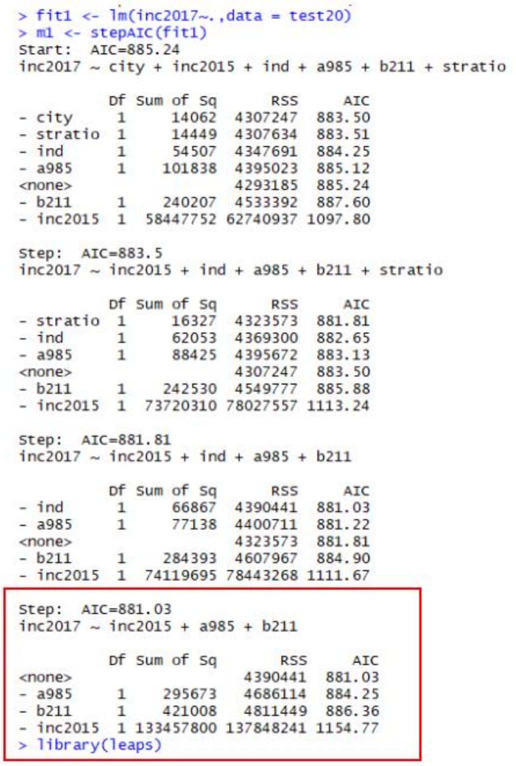


Figure 3 Scatter plot

Next, the author conducted the AIC to find the best fit regression term according to the dependent variable that are selected.



```
> fit1 <- lm(inc2017~.,data = test20)
> m1 <- stepAIC(fit1)
Start:  AIC=885.24
inc2017 ~ city + inc2015 + ind + a985 + b211 + stratio

          Df Sum of Sq      RSS     AIC
- city     1     14062  4307247  883.50
- stratio  1     14449  4307634  883.51
- ind      1     54507  4347691  884.25
- a985     1    101838  4395023  885.12
<none>                  4293185  885.24
- b211     1    240207  4533392  887.60
- inc2015  1  58447752 62740937 1097.80

Step:  AIC=883.5
inc2017 ~ inc2015 + ind + a985 + b211 + stratio

          Df Sum of Sq      RSS     AIC
- stratio  1     16327  4323573  881.81
- ind      1     62053  4369300  882.65
- a985     1     88425  4395672  883.13
<none>                  4307247  883.50
- b211     1    242530  4549777  885.88
- inc2015  1  73720310 78027557 1113.24

Step:  AIC=881.81
inc2017 ~ inc2015 + ind + a985 + b211

          Df Sum of Sq      RSS     AIC
- ind      1     66867  4390441  881.03
- a985     1     77138  4400711  881.22
<none>                  4323573  881.81
- b211     1    284393  4607967  884.90
- inc2015  1  74119695 78443268 1111.67

Step:  AIC=881.03
inc2017 ~ inc2015 + a985 + b211

          Df Sum of Sq       RSS     AIC
<none>                   4390441  881.03
- a985     1    295673   4686114  884.25
- b211     1    421008   4811449  886.36
- inc2015  1 133457800 137848241 1154.77
> library(leaps)
```

Figure 4 AIC test

The result shows that the smaller AIC it has, the more better this model is.

3.1.2. Regresubsets

After Stepwise regression and All-subsets regression, the following picture can be got.

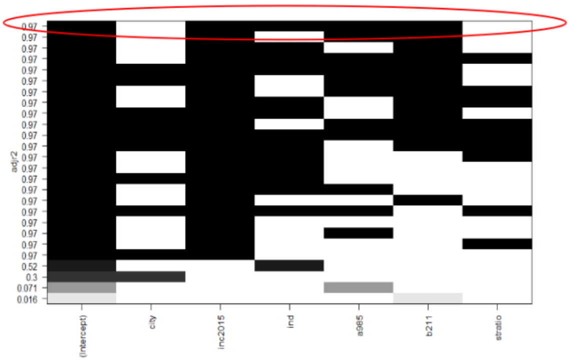


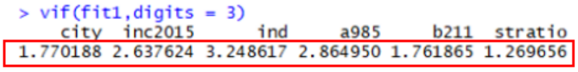
Figure 5 Regresubsets

Finally, the following model was selected.

$$inc2017 = \beta_0 + \beta_1 inc2015 + \beta_2 ind + \beta_3 a985 + \beta_4 b211 \quad (1)$$

3.2. Modelling building

3.2.1. Variable selection



```
> vif(fit1,digits = 3)
    city  inc2015      ind     a985     b211  stratio
1.770188 2.637624 3.248617 2.864950 1.761865 1.269656
```

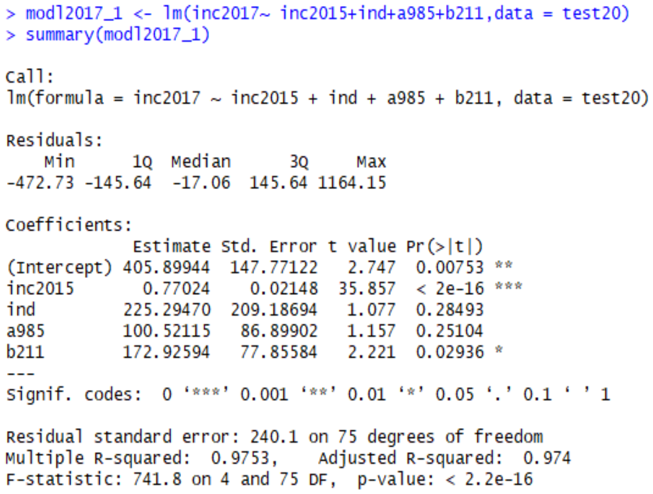
The above picture shows the result. If VIF<10, explanatory variables have no relative or multicollinearity issues.

3.2.2. Variables Transformation

(1) Original form function

Build the Original form function as following:

$$inc2017 = \beta_0 + \beta_1 inc2015 + \beta_2 ind + \beta_3 a985 + \beta_4 b211 \quad (2)$$



```
> modl2017_1 <- lm(inc2017~ inc2015+ind+a985+b211,data = test20)
> summary(modl2017_1)

Call:
lm(formula = inc2017 ~ inc2015 + ind + a985 + b211, data = test20)

Residuals:
    Min      1Q  Median      3Q     Max 
-472.73 -145.64  -17.06  145.64 1164.15 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 405.89944  147.77122   2.747  0.00753 ** 
inc2015       0.77024    0.02148  35.857  < 2e-16 ***
ind         225.29470  209.18694   1.077  0.28493    
a985        100.52115   86.89902   1.157  0.25104    
b211        172.92594   77.85584   2.221  0.02936 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 240.1 on 75 degrees of freedom
Multiple R-squared:  0.9753,	Adjusted R-squared:  0.974 
F-statistic: 741.8 on 4 and 75 DF,  p-value: < 2.2e-16
```

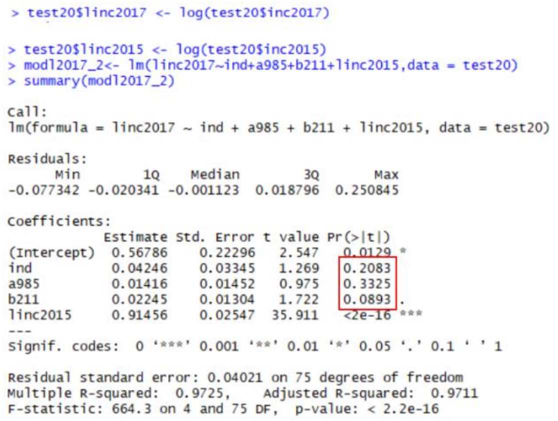
Figure 6 Liner regression1

According to the p-value, a985 and b211 are not very significant then the author decided to change variables.

(2) Take log of inc2017 and inc2015

While talking about income we always take log, in order to get a smaller number which can be easier to observe.

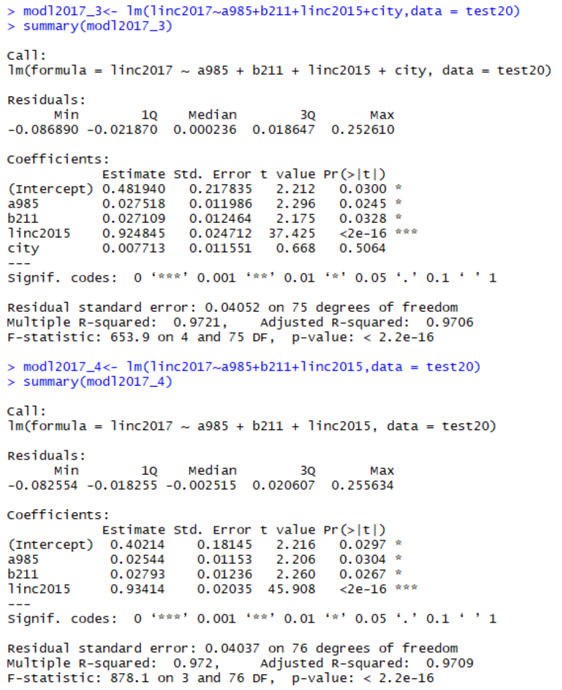
$$linc2017 = \beta_0 + \beta_1 ind + \beta_2 a985 + \beta_3 b211 + \beta_4 linc2015 \quad (3)$$



```
> test20$linc2017 <- log(test20$inc2017)

> test20$linc2015 <- log(test20$inc2015)
> modl2017_2<- lm(linc2017~ind+a985+b211+linc2015,data = test20)
> summary(modl2017_2)

Call:
lm(formula = linc2017 ~ ind + a985 + b211 + linc2015, data = test20)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.077342 -0.020341 -0.001123  0.018796  0.250845 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.56786    0.22296   2.547   0.0129 *  
ind          0.04246    0.03345   1.269   0.2083    
a985         0.01416    0.01452   0.975   0.3325    
b211         0.02245    0.01304   1.722   0.0893 .  
linc2015     0.91456    0.02547  35.911   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04021 on 75 degrees of freedom
Multiple R-squared:  0.9725,    Adjusted R-squared:  0.9711 
F-statistic: 664.3 on 4 and 75 DF,  p-value: < 2.2e-16
```

Figure 7 Linear regression after take log1

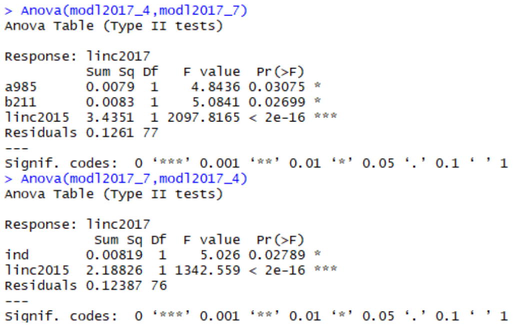
After taking log of the income, those three variables, ind, b211, a985 are becoming insignificant, so there was an idea that a school whether it is a985 or b211 may have a relationship with ind (a good school will also collect the best students with the higher rank of the examination), thus the author modeled them separately and then analyzed them by using ANOVA.



```
> modl2017_3<- lm(linc2017~a985+b211+linc2015+city,data = test20)
> summary(modl2017_3)

Call:
lm(formula = linc2017 ~ a985 + b211 + linc2015 + city, data = test20)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.086890 -0.021870  0.000236  0.018647  0.252610 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.481940   0.217835   2.212   0.0300 *  
a985        0.027518   0.011986   2.296   0.0245 *  
b211        0.027109   0.012464   2.175   0.0328 *  
linc2015    0.924845   0.024712  37.425   <2e-16 ***
city        0.007713   0.011551   0.668   0.5064    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04052 on 75 degrees of freedom
Multiple R-squared:  0.9721,    Adjusted R-squared:  0.9706 
F-statistic: 653.9 on 4 and 75 DF,  p-value: < 2.2e-16

> modl2017_4<- lm(linc2017~a985+b211+linc2015,data = test20)
> summary(modl2017_4)

Call:
lm(formula = linc2017 ~ a985 + b211 + linc2015, data = test20)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.082554 -0.018255 -0.002515  0.020607  0.255634 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.40214    0.18145   2.216   0.0297 *  
a985         0.02544    0.01153   2.206   0.0304 *  
b211         0.02793    0.01236   2.260   0.0267 *  
linc2015     0.93414    0.02035  45.908   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04037 on 76 degrees of freedom
Multiple R-squared:  0.972,     Adjusted R-squared:  0.9709 
F-statistic: 878.1 on 3 and 76 DF,  p-value: < 2.2e-16
```

Figure 8 Linear regression after take log2

Then ANOVA was used,



```
> Anova(modl2017_4,modl2017_7)
Anova Table (Type II tests)

Response: linc2017
          Sum Sq Df   F value  Pr(>F)    
a985      0.0079  1    4.8436 0.03075 *  
b211      0.0083  1    5.0841 0.02699 *  
linc2015  3.4351  1 2097.8165 < 2e-16 ***
Residuals 0.1261 77                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> Anova(modl2017_7,modl2017_4)
Anova Table (Type II tests)

Response: linc2017
           Sum Sq Df  F value  Pr(>F)    
ind       0.00819  1    5.026 0.02789 *  
linc2015  2.18826  1 1342.559 < 2e-16 ***
Residuals 0.12387 76                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9 ANOVA

The author found both of them are all satisfied, then plotted them separately.

Plot(modl2017_4)

After choosing a better model, the residuals analyse are shown though the following.

$$linc2017 = \beta_0 + \beta_1 a985 + \beta_2 b211 + \beta_3 linc2015 \quad (4)$$

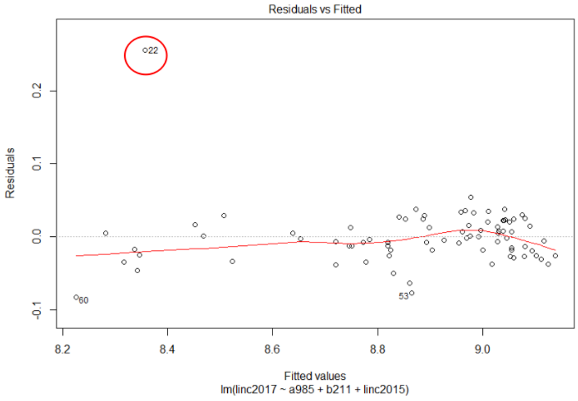


Figure 10 Residuals1

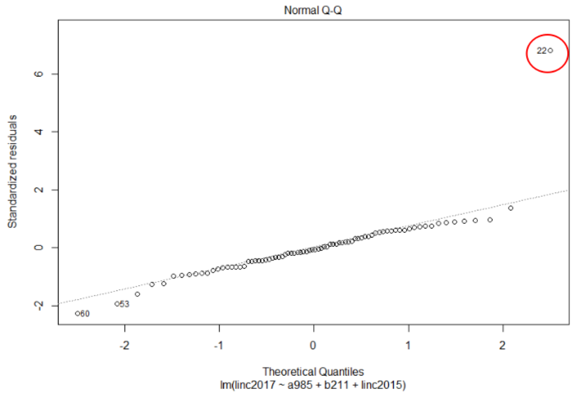


Figure 11 QQ-plot

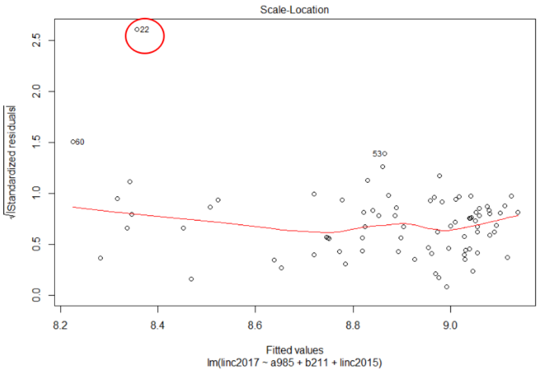


Figure 12 Scale-Location

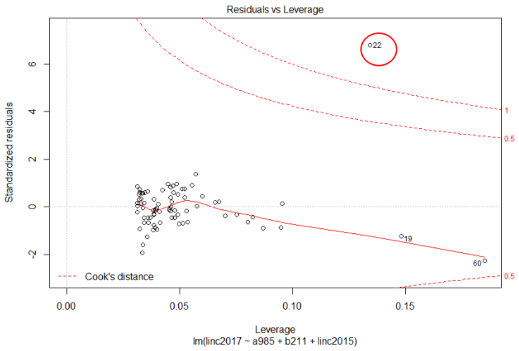


Figure 13 Residuals1

Plot (modl2017_7)

The model only contain the lagging term.

$$linc2017 = \beta_0 + \beta_1 ind + \beta_2 linc2015 \quad (5)$$

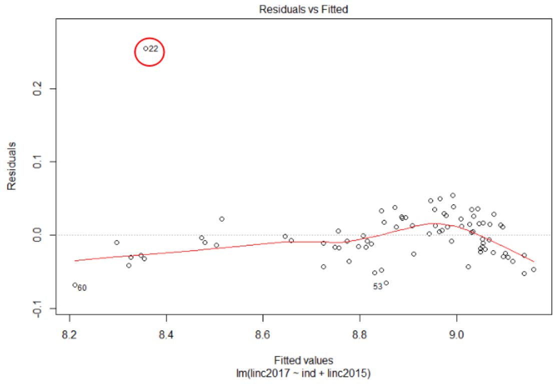


Figure 14 Residuals2

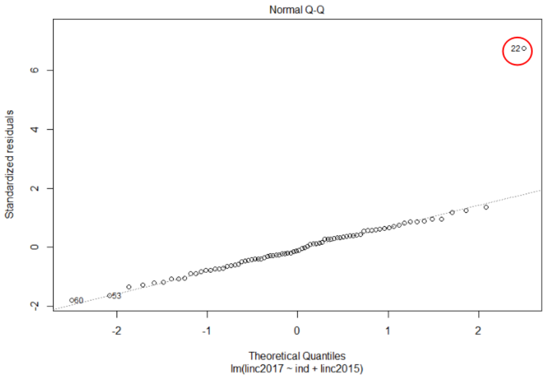


Figure 15 QQ-plot

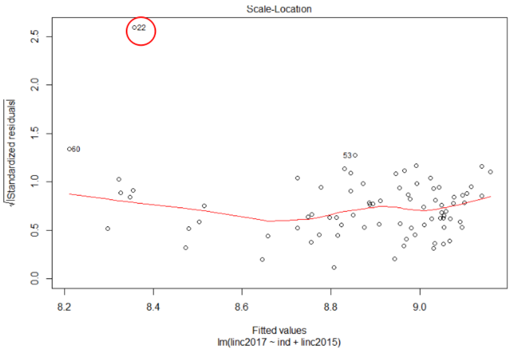


Figure 16 Scale-location

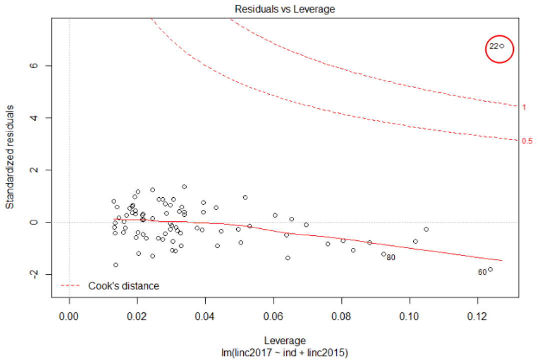
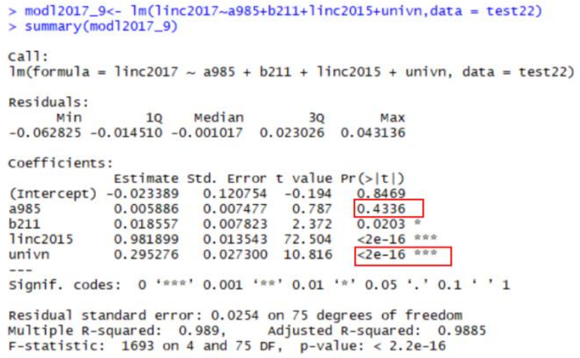


Figure 17 Residuals vs Leverage

3.2.3. Analysis outlier point

Comparing both of those two models, the author found that first model is better. $(linc2017 = \beta_0 + \beta_1 linc2015 + \beta_2 a985 + \beta_3 b211)$

However, the author found that Sichuan University is an outlier point in both of those two models, then what should we do about the outlier? Firstly the Sichuan University was changed to a dummy variable, then a new regression was conducted, finally the author got the result as the picture blow.



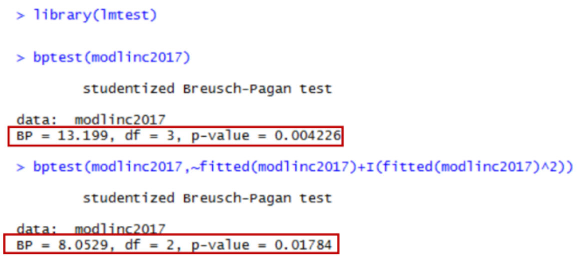
```
> modl2017_9<- lm(linc2017~a985+b211+linc2015+univn,data = test22)
> summary(modl2017_9)

Call:
lm(formula = linc2017 ~ a985 + b211 + linc2015 + univn, data = test22)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.062825 -0.014510 -0.001017  0.023026  0.043136 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.023389   0.120754  -0.194   0.8469    
a985         0.005886   0.007477   0.787   0.4336    
b211         0.018557   0.007823   2.372   0.0203 *  
linc2015     0.981899   0.013543  72.504   <2e-16 ***
univn        0.295276   0.027300  10.816   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0254 on 75 degrees of freedom
Multiple R-squared:  0.989,	Adjusted R-squared:  0.9885 
F-statistic:  1693 on 4 and 75 DF,  p-value: < 2.2e-16
```

Figure 18 Dummy variable

Although the new dummy variable is statistically significant, and when this dummy variable added into regression, it affected the other a985 dummy variable. This was the result that the author did not want to have. So the author would not add Sichuan University as a dummy variable into our regression. Next the author tries to find that why this outlier will appear in our regression, except this number was a wrong number.

From the picture which showed the residual, it can be known that this number was over estimation.

3.2.4. Homoscedasticity Test

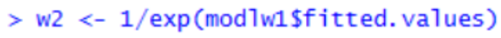


```
> library(lmtest)

> bptest(modlinc2017)

	studentized Breusch-Pagan test

data:  modlinc2017
BP = 13.199, df = 3, p-value = 0.004226

> bptest(modlinc2017,~fitted(modlinc2017)+I(fitted(modlinc2017)^2))

	studentized Breusch-Pagan test

data:  modlinc2017
BP = 8.0529, df = 2, p-value = 0.01784
```

Figure 19 Homoscedasticity Test

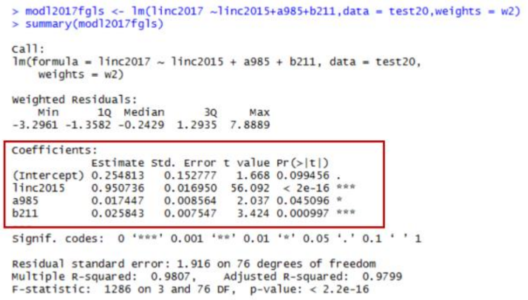
The author did the BP Test (Breusch-Pagan Test), $\mathrm{p}-\mathrm{value}_{\mathrm{BP}}=$ 0.004226, and there is some evidence of heteroskedasticity at the 5% level. In general, the White Test is more able to see the general bias of heteroscedasticity, so the author also did the white Test $\mathrm{p}-\mathrm{value}_{\mathrm{white}}=$ 0.01784. Not surprisingly the result is very similar with BP Test. In conclusion, the author realized that the regression model had heteroskedasticity.

Next, the author can not do robust standard error + OLS and did not know the form of weight. So, the feasible GLS should be selected to continue the whole process.



```
> w2 <- 1/exp(modlw1$fitted.values)
```

The following is the regression,



```
> modl2017fgls <- lm(linc2017 ~linc2015+a985+b211,data = test20,weights = w2)
> summary(modl2017fgls)

Call:
lm(formula = linc2017 ~ linc2015 + a985 + b211, data = test20, 
    weights = w2)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-3.2961 -1.3582 -0.2429  1.2935  7.8889 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.254813   0.152777   1.668 0.099456 .  
linc2015    0.950736   0.016950  56.092  < 2e-16 ***
a985        0.017447   0.008564   2.037 0.045096 *  
b211        0.025843   0.007547   3.424 0.000997 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.916 on 76 degrees of freedom
Multiple R-squared:  0.9807,	Adjusted R-squared:  0.9799 
F-statistic:  1286 on 3 and 76 DF,  p-value: < 2.2e-16
```

Figure 20 GLS

4. RESULT AND DISCUSSION

$$inc2017=\beta_0+\beta_1 inc2015+\beta_2 a985+\beta_3 b211+\beta_4 stratio+\beta_5 ind+\beta_6 total \tag{6}$$

$$inc2017=405.8994+0.7702inc2015+100.5212a985+172.9259b211+225.2947ind \tag{7}$$

$$linc2017=0.5679+0.9146linc2015+0.0142a985+0.0225b211 \tag{8}$$

The explanatory variables selected were the total income expected by the university in 2017 (the school’s funding status reflects the size and strength of the university) and the school’s teacher-student ratio (teaching resources) without a significant impact, whether the school location is a first-tier city (location factor), the overall admission score index (student’s own level) and the four or six average scores (students' ability to study during college) also had small impact in 2014, but in 2015 the graduation salary of college students and whether the university was 985, 211 and other key universities had a significant impact.

Through this analysis, it can be known that the fame of a school directly affects the salary of college graduates. Therefore, students should study hard in high school and try their best to be offered by famous universities so that they can get high salaries after graduation.

There are some limitations as well. This article just analyzed 80 universities in China, the findings only have an impact on college graduates in China. At the same time, this paper did not rule out that there were more effective factors affecting the salary of graduates. This paper only listed the most intuitive reflection of graduates salary elements.

5. CONCLUSION

The conclusion is that wages depend more on the school and the average wage in society. Through these analysis, students can be offered a direction for their own academic career and employment salary advance. High school students should take choice seriously when deciding the entering college. The attribute of college should not only be famous, but also have a reasonable expectation of their post-graduation employment wages. Therefore, the model can provide a certain basis for students to choose a proper university, and the school should also pay more attention to the problem of school fame in order to solve the problem better.

ACKNOWLEDGMENT

First and foremost, I would like to show my deepest gratitude to my teachers and professors in my university, who have provided me with valuable guidance in every stage of the writing of this thesis. Further, I would like to thank my friends and parents for their encouragement and support. Without all their enlightening instruction and impressive kindness, I could not have completed my thesis.

REFERENCES

[1] Xinchou.Com. (2018). [Online] Available:https://www.xinchou.com/ [Retrieved on April 14th]

[2] Research team of associate professor kuang chunwei (2014). Ranking of admission scores of Chinese universities (2014 edition). Social investigation center of east China normal university.

[3] Chinese Academy of Social Sciences magazine. (2019, 12 (7). Retrieved from Chinese social sciences net: www.cssn.cn/ on April 14th)

[4] Zhang, J. , Zhao, Y. , Park, A. , & Song, X. (2005). Economic returns to schooling in urban china, 1988 to 2001. Journal of Comparative Economics, 33(4), 0-752.

[5] Liu Xujie, & Yue Changjun. An empirical study on the influencing factors of college graduates' salary change. Friends of accounting 2010(5), 53-57.