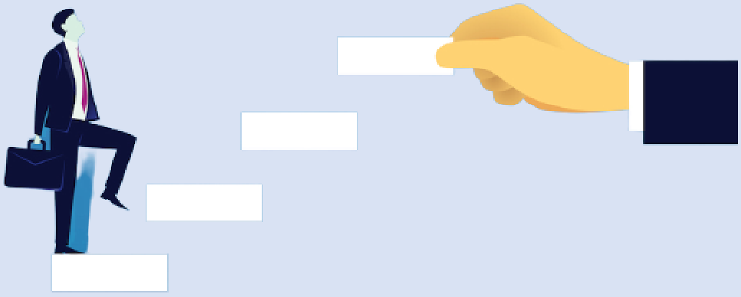
Degree-Job mismatch:Does it affect income?

DISC 203 - Group 2 - S1



Muhammad Ahmed Ehtisham - 24110167
Ammar Jawed Munshi - 24110071
Ali Mohammad Rana - 24110026
Muhammad Ahmad - 24110158
Sajid Hussain - 24110250

Research Question

"To what extent does the mismatch of your degree and the job market you work in affect your income?"

Table of Contents

Introduction

In today's technologically advanced world constituting a highly dynamic job market, horizontal mismatch between degrees and jobs has become prevalent (Piriya 2017). A horizontal mismatch means that with the same level of qualifications and skills, the type of degree and the job description are not consistent. A number of studies are available on this matter, but few are available in Pakistan. In a preliminary survey, it was observed that at least 34% of students in LUMS wanted to study for another degree. In another survey conducted, 13.6% of LUMS graduates currently claimed that they did not intend to study the degree they graduated with. A detailed survey was sent out to the LUMS alumni network to assess these inconsistencies. This enabled the report to use statistical tests and run regressions using R-programming to determine if inconsistencies between degree and job influence the income earned, and how big of a role do other factors like family background, delays and CGPA etc. play.

Methodology

Assumptions:

- Everyone is not able to study their preferred major
- The intended job market remains the same
- Gender disparity doesn't exist
- Lums data can be translated to other non-medical universities in Pakistan

Variables

The main aim of the research is to explore the relation between job-degree mismatch and the income one earns. Consequently, dependent, and independent variables had to be chosen. Consequently, the **dependent (Y) variable** was chosen as the **monthly income** earned. All factors chosen for the X variables were kept same for both models. To observe how the independent variables affect income when there is a consistency or an inconsistency between job and degree, **8 X-variables** were considered.

Data Collection and Cleaning

For this research, an online survey was created in order to collect the relevant data for the research. The survey was aimed to gather more information on independent factors such as CGPA, job experience, and family background, which would then be compared to the research's dependent variable, income. An online survey tool, Typeform was used to design the survey, as it allows the owner to use certain logic so that the follow-up questions that are not relevant to some respondents are not visible to them. Doing so makes it less likely for the respondents to default when filling the survey form and minimizes the response error. The survey was primarily made up of closed-ended questions. Before the survey was sent out, a sample of 20 alumni were asked to fill the form and their feedback gathered, in order to assess the average time taken to fill the form and make any changes necessary to make it more understandable and easier. The questionnaire was a blend of both nominal and ratio scale items such as, "Are you working in your preferred job market? (did you want to work in this sector?)" and "How many (if applicable) job(s)/part-time job(s) did you have?".

LUMS graduates from 2010 to 2020 were selected as the target respondents so that the information collected would be up-to-date. For the distribution of the survey, online and on-campus resources such as LUMS Discussion Group, UGPCO, and Office of Advancement were utilized. Moreover, the survey was also sent out to the respondents personally. In total, 179 responses were collected, which were then filtered using a stringent criterion to acquire only accurate data relevant to the model. In order to observe the link between different variables, data collected was evaluated using *R-programming* for an initial analysis. As quantitative data was required to verify the hypothesis and to conduct the regression, the qualitative data gathered was translated into quantitative data using scale assessment. 'Factor' was utilised for qualitative data, and dummy variables were employed to interpret and calculate the odds ratios. Values of 1 and 0 were assigned for the true/false responses, respectively. Likewise, different scales ranging from 0 to 10 were allocated based on the variables' needs. For example, the income bracket was assigned a scale ranging from 0 to 10, with 0 indicating the lowest and 10 indicating the highest income. Erroneous data was filtered out and extreme outliers were eliminated from the data.

Descriptive Statistics

Histograms

From the chosen sample size of 158 individuals, majority of the respondents were shown to be enrolled in an Accounting and Finance major, with respondents opting for a Management Sciences degree coming in at a close second. The remaining sample seems to be pretty evenly dispersed over a variety of disciplines, ranging from Economics to Computer Science. After data cleaning, of the 158 who filled the survey, the data set is interestingly divided between those either earning under PKR 200,000 or over PKR 300,000. Only a negligible number of respondents fall between the two categories, and this may be due to the differences in job categories of the sample set. On the higher end, more than 45 respondents are currently earning between the PKR 350,000 to PKR 400,000 income bracket, whereas on the lower end, there is a relatively even distribution, with 50 to 60 respondents being employed in jobs paying them between PKR 50,000 and PKR 150,000. None of the sample is earning below PKR 25,000.

Box plots

The graph for delay between graduation and employment has 3 outliers which stand out, with majority of the dataset being concentrated around little to no delay between graduation and employment, whereas each of the 3 outliers happens to have a 3-, 4-, and 5-year delay respectively. For analyzing family background type of respondents, the outliers lie on both extremes of the available range, with no representation from lower class and elite class backgrounds. The CGPA box plot shows a single outlier who earned a CGPA of 2.5 but is still earning a relatively high monthly income, possibly hinting at being self-employed or working for a family business. These outliers increase variability which means a higher P value in regression and a comparatively lower R-squared value for both models.

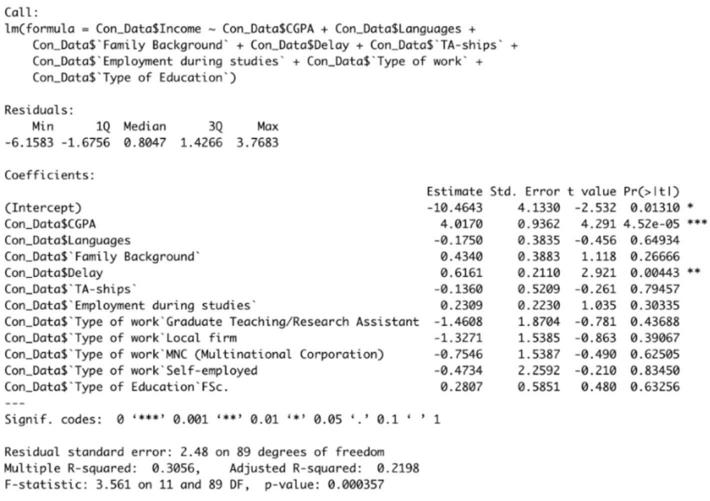
Two model division

The data sets were divided into 2 separate models - one where job and degree were consistent for the respondents and one where they were not. In simple words, if people worked a job consistent to their degree, their data was classified separately as 'Model 1- consistent data.' The screenshots for both models have been placed along with their regression equations. Two separate hypotheses are going to assess the impact of the independent variables on the dependent variable and give an insight on how well the mismatch between degree-job and income can be understood. The lower the p value, the stronger the variable is in explaining mismatch. Since there are two separate models, the research can show how strong the models are explained through the variables. Since both models are bound to have a different regression statistic, the analysis can also extend till the point where it can be explained which relationship is stronger compared to the other. (Compare both models and see which one is stronger/better explained by the variables.)

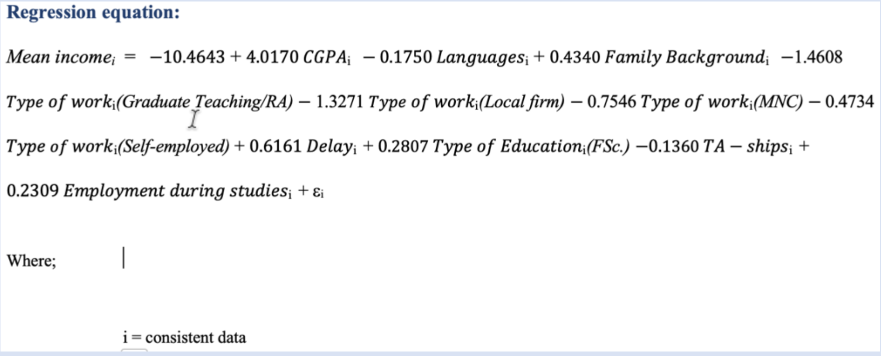
Regression

Model 1

Full regression summary



```
Call:
lm(formula = Con_Data$Income ~ Con_Data$CGPA + Con_Data$Languages +
    Con_Data$`Family Background` + Con_Data$Delay + Con_Data$`TA-ships` +
    Con_Data$`Employment during studies` + Con_Data$`Type of work` +
    Con_Data$`Type of Education`)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1583 -1.6756  0.8047  1.4266  3.7683

Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                               -10.4643     4.1330  -2.532  0.01310 *
Con_Data$CGPA                                               4.0170     0.9362   4.291 4.52e-05 ***
Con_Data$Languages                                         -0.1750     0.3835  -0.456  0.64934
Con_Data$`Family Background`                                0.4340     0.3883   1.118  0.26666
Con_Data$Delay                                              0.6161     0.2110   2.921  0.00443 **
Con_Data$`TA-ships`                                        -0.1360     0.5209  -0.261  0.79457
Con_Data$`Employment during studies`                        0.2309     0.2230   1.035  0.30335
Con_Data$`Type of work`Graduate Teaching/Research Assistant  -1.4608     1.8704  -0.781  0.43688
Con_Data$`Type of work`Local firm                          -1.3271     1.5385  -0.863  0.39067
Con_Data$`Type of work`MNC (Multinational Corporation)     -0.7546     1.5387  -0.490  0.62505
Con_Data$`Type of work`Self-employed                       -0.4734     2.2592  -0.210  0.83450
Con_Data$`Type of Education`FSc.                            0.2807     0.5851   0.480  0.63256
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.48 on 89 degrees of freedom
Multiple R-squared:  0.3056,	Adjusted R-squared:  0.2198
F-statistic: 3.561 on 11 and 89 DF,  p-value: 0.000357
```



Regression equation:

$$Mean\ income_i = -10.4643 + 4.0170\ CGPA_i - 0.1750\ Languages_i + 0.4340\ Family\ Background_i - 1.4608\ Type\ of\ work_i(Graduate\ Teaching/RA) - 1.3271\ Type\ of\ work_i(Local\ firm) - 0.7546\ Type\ of\ work_i(MNC) - 0.4734\ Type\ of\ work_i(Self\text{-}employed) + 0.6161\ Delay_i + 0.2807\ Type\ of\ Education_i(FSc.) - 0.1360\ TA - ships_i + 0.2309\ Employment\ during\ studies_i + \varepsilon_i$$

Where;

i = consistent data

Hypothesis

Null Hypothesis: $H_0 : \beta_1 = 0$

Alternate Hypothesis: $H_1 : \beta_1 \neq 0$

Overall significance

Since these three have fewer outliers, it means that they have fewer standard errors, for example the family background data only has a standard error of 0.38, CGPA has one of 0.93, while delay from graduation has a 0.21 standard error. Hence these are significant, as shown by their P value. All three of these have a P value lower than alpha, with family background at 0.26, delay at 0.004, and CGPA at 4.52x10^-5 P value. When P value lies below alpha these variables contribute to explaining the mismatch.

1. **Beta Coefficients**

The null hypothesis states that when type of degree and job are consistent, variables have no effect on income, while the alternates hypothesis claims that, when type of degree and job are consistent, variables influence income.

2. Adjusted R-squared

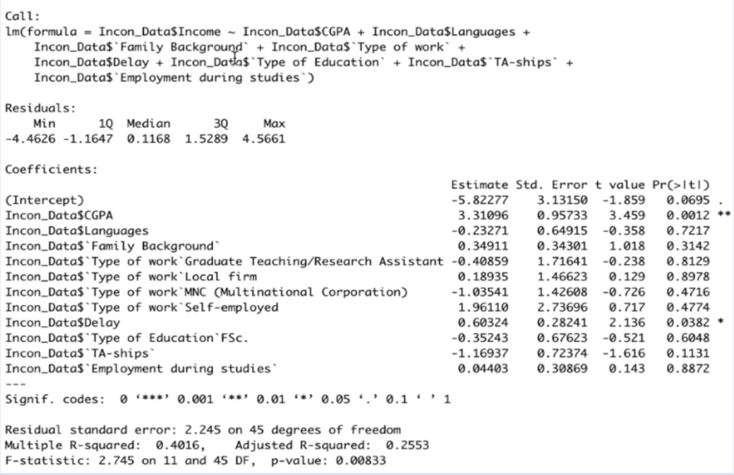
The R-squared is also low, amounting to 0.3056 which means that only 30.56% of the variation in income (Dependent variable) is explained by the change of all the independent variables.

3. F-testing

When a full multiple regression model is run, the p-value of the F-statistic is found to be 0.000357 which is lower than the significance level of 5%. Hence, it can be observed that the model is statistically significant and there is sufficient evidence to reject the null hypothesis. Therefore, it can be stated that the variables influence the income earned by a person given that their degree and job are consistent. The standard errors are observed to be relatively low, therefore, the model is less erroneous.

Model 2

Full regression summary

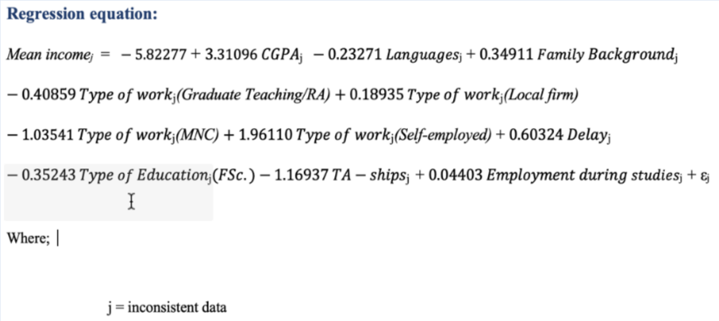


```
Call:
lm(formula = Incon_Data$Income ~ Incon_Data$CGPA + Incon_Data$Languages +
    Incon_Data$`Family Background` + Incon_Data$`Type of work` +
    Incon_Data$Delay + Incon_Data$`Type of Education` + Incon_Data$`TA-ships` +
    Incon_Data$`Employment during studies`)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4626 -1.1647  0.1168  1.5289  4.5661

Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                                          -5.82277    3.13150  -1.859   0.0695 .
Incon_Data$CGPA                                       3.31096    0.95733   3.459   0.0012 **
Incon_Data$Languages                                 -0.23271    0.64915  -0.358   0.7217
Incon_Data$`Family Background`                        0.34911    0.34301   1.018   0.3142
Incon_Data$`Type of work`Graduate Teaching/Research Assistant -0.40859    1.71641  -0.238   0.8129
Incon_Data$`Type of work`Local firm                   0.18935    1.46623   0.129   0.8978
Incon_Data$`Type of work`MNC (Multinational Corporation)      -1.03541    1.42608  -0.726   0.4716
Incon_Data$`Type of work`Self-employed                1.96110    2.73696   0.717   0.4774
Incon_Data$Delay                                      0.60324    0.28241   2.136   0.0382 *
Incon_Data$`Type of Education`FSc.                   -0.35243    0.67623  -0.521   0.6048
Incon_Data$`TA-ships`                                -1.16937    0.72374  -1.616   0.1131
Incon_Data$`Employment during studies`                0.04403    0.30869   0.143   0.8872
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.245 on 45 degrees of freedom
Multiple R-squared:  0.4016,	Adjusted R-squared:  0.2553
F-statistic: 2.745 on 11 and 45 DF,  p-value: 0.00833
```

Hypothesis

Null Hypothesis: $H_0 : \beta_1 = 0$

Alternate Hypothesis: $H_1 : \beta_1 \neq 0$



Regression equation:

$$\textit{Mean income}_j = -5.82277 + 3.31096\ \textit{CGPA}_j - 0.23271\ \textit{Languages}_j + 0.34911\ \textit{Family Background}_j$$

$$-0.40859\ \textit{Type of work}_j\textit{(Graduate Teaching/RA)} + 0.18935\ \textit{Type of work}_j\textit{(Local firm)}$$

$$-1.03541\ \textit{Type of work}_j\textit{(MNC)} + 1.96110\ \textit{Type of work}_j\textit{(Self-employed)} + 0.60324\ \textit{Delay}_j$$

$$-0.35243\ \textit{Type of Education}_j(\textit{FSc.}) - 1.16937\ \textit{TA} - \textit{ships}_j + 0.04403\ \textit{Employment during studies}_j + \varepsilon_j$$

Where;

j = inconsistent data

Overall significance

The following three variables have the least outliers, hence fewer standard errors, as can be seen below:

- CGPA - 0.95 standard error
- Family Background - 0.34 standard error
- Delay between Graduation and Employment - 0.28 standard error.

Due to these low errors, these variables are significant, as can also be seen by their P value. Each of these have P value lower than alpha (0.05), which is why they contribute to explaining the mismatch:

- CGPA - 0.0012 P value
- Family Background - 0.314 P value
- Delay between Graduation and Employment - 0.038 P value.

1. Beta Coefficients

The null hypothesis states that when type of degree and job are not consistent, variables have no effect on income, while the alternates hypothesis claims that, when type of degree and job are not consistent, variables influence income.

2. Adjusted R-squared

The R-squared is also low, amounting to 0.4016 which means that only 40.16% of the variation in income (Dependent variable) is explained by the change of all the independent variables.

3. F-testing

When a full multiple regression model is run, the p-value of the F-statistic is found to be 0.00833 which is lower than the significance level of 5%. Hence, it can be observed that the model is statistically significant and there is sufficient evidence to reject the null hypothesis. Therefore, it can be stated that the variables influence the income earned by a person given that their degree and job are not consistent. The standard errors are observed to be relatively low, therefore, the model is less erroneous.

Findings and Implications

It was observed that the p-value of the F-statistic for both models (Model 1 and Model 2) were significantly less than the significance level chosen (5% in this case). This means that for both models, there is sufficient evidence to reject the Null hypothesis of both models. Surprisingly, the variables effected in both the models were also the same. The CGPA variable and the delay variable were the ones most affected in both models with p-values less than alpha of 0.05. When individual regressions were run for all the variables, it was observed that both models, it was observed that the family background variable also proved to be significant. However, it was observed that these values were comparatively lesser for Model 1, including the F-statistic p-value as well as the individual p-values of the variables. The standard errors are lesser in Model 2 and the R-squared is also greater in Model 2. Furthermore, the degrees of freedom are higher for Model 1 than Model 2.

The research implies that monthly income is affected by the mismatch between the type of degree and job an individual has. Changes in factors like CGPA, Delay between employment and graduation and type of family background play a significant role in determining a person's monthly income, but these factors are more crucial when the job and degree are consistent. Since the research was carried out on LUMS alumni, the strength of the correlation in the differences that exists within the non-medical universities across Pakistan can be assessed. This also means that an individual studying in Pakistan should be well-aware of the dynamic job market and of the direction it is moving in, to avoid choosing fields of study or type of jobs that might be becoming obsolete. Students must plan out their degrees and intended job markets in advance in order to avoid mismatches. Moreover, CGPA is highly dependent on the performance of the individual, hence students should be advised to work hard in these years in order to earn a better monthly income. One variable, Delay, can be avoided, such as individual can work to decrease this time period since it does affect the monthly income. Companies do prefer fresh graduates as new employees since they bring in new perspectives, are highly adaptable and keen to learn (*Why Do Companies Hire Fresh Graduates?*, 2021). Furthermore, students can increase their networking and work experience as well as acquire other skills in order to diversify their major in order to reduce mismatch. Since a higher level social class is also related to better sources and wider networks, students can build that for themselves in order to increase the number of opportunities. With a diverse degree or with multiple skills and more knowledge, mismatch may decrease in other types of jobs, hence, affecting their monthly income positively.

Limitations

Even though both models turned out to be significant, some limitations were also observed. Firstly, the nature of the model demanded that the data be divided two parts, i.e., consistency between job income and inconsistency between them. Hence, the regression for consistent and inconsistent data could not be run together. Secondly, the models did not go through variable selection process, so variables that weren't strongly corelated i.e., languages were still considered. Therefore, the R-squared values observed were also quite low. Thirdly, there is an excellent possibility that the data cleaning may have had overlooked biases – such as through summary statistics, it was confirmed that there were significantly more responses from business school graduates than the others, hence there may be an availability bias. Fourthly, the survey was designed in such a way that it required subjective responses, consequently making the dataset more qualitative. Hence, the use of dummy variables and scales assigned to answers allowed regression to be run. Another factor to consider is that perception of social class (one of the independent variables) may be arbitrary and hence, allow some inaccuracy. Lastly, a significant limitation is that gender was not considered. The sample consisted of uneven responses from both males and females respondents. Gender may play an essential role in determining a person's pay, and so it might negatively affect the assessment of the income on job-degree mismatch. According to a global wage report of 2018/19, women earn 20 per cent less than men globally for the same kind of job. So, gender inequalities may have introduced further inaccuracies not accounted for.

Conclusion

By separating the responses collected through a carefully designed survey and separating it into two different datasets for the two models designed, interpretation of the results became clearer. This research concluded that the horizontal mismatch between job and degree does affect the income of a person, the main independent variables affecting the income (dependent variable) were CGPA, delay between graduation and employment, and the type of family background an individual belongs to. Hence, it can be concluded that for students studying in non-medical universities in Pakistan, CGPA, delay between graduation and employment, and the type of family background highly affects the monthly income, and the horizontal mismatch between job and degree also plays a role in the monthly income earned.

Bibliography

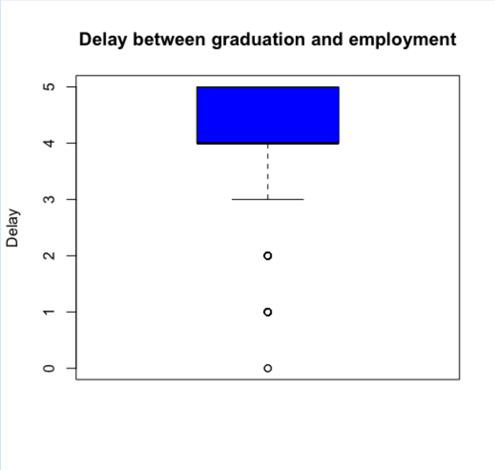
Why do Companies Hire Fresh Graduates? (2021, October 13). BMS Performance. https://bmsperformance.com/blog/graduates/why-do-companies-hire-fresh-graduates/

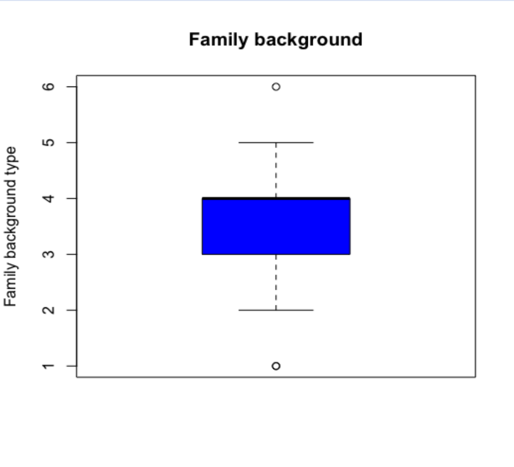
Global Wage Report 2018/19: What lies behind Gender Pay Gap. (2018, November 28). International Labour Organization. https://www.ilo.org/islamabad/info/public/pr/WCMS_651658/lang--en/index.htm

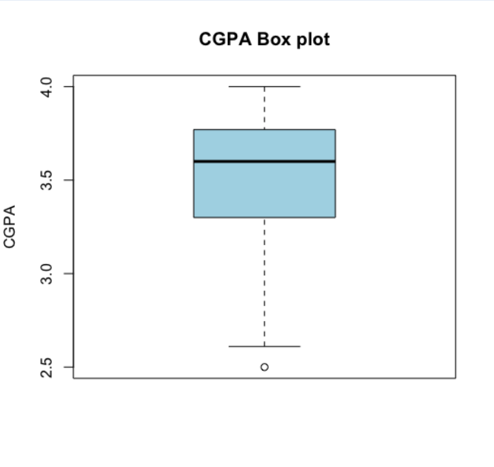
Pholphirul, P. (2017). "Educational Mismatches and Labor Market Outcomes: Evidence from Both Vertical and Horizontal Mismatches in Thailand". Emerald Group Publishing Limited. https://eric.ed.gov/?id=EJ1140344

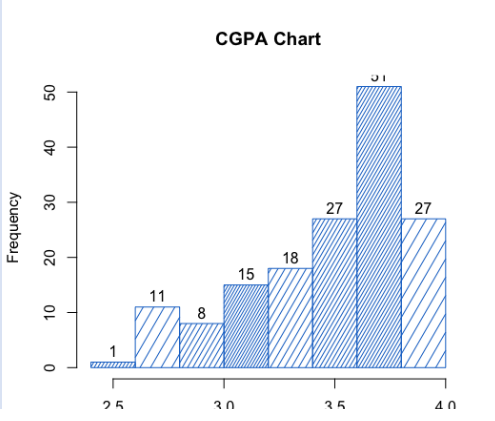
Index

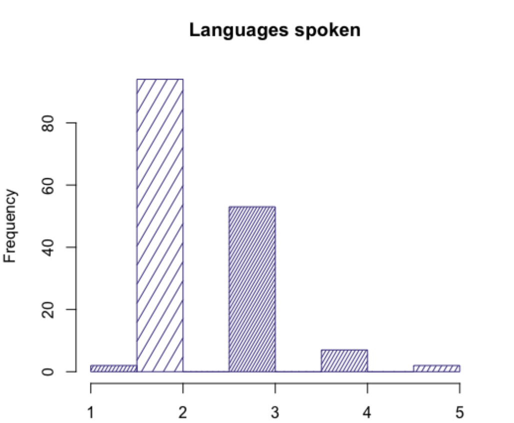
Box plots and Histograms

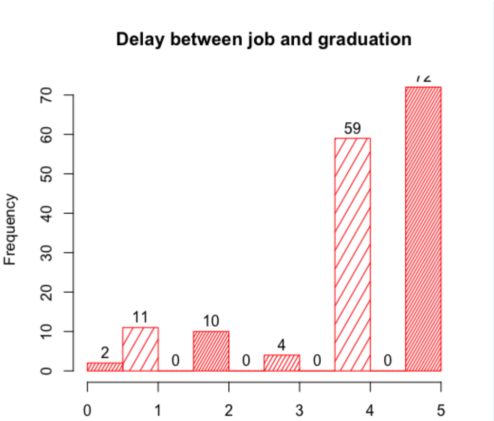


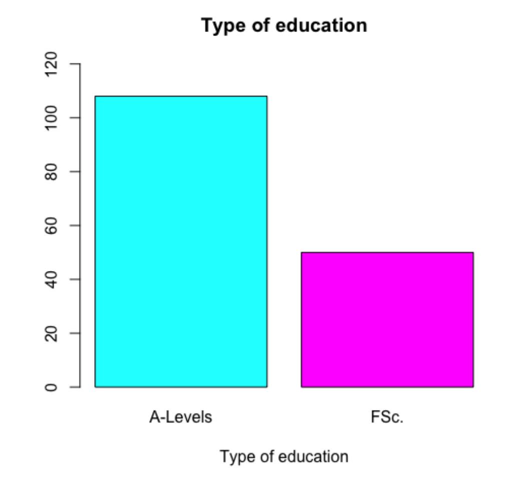


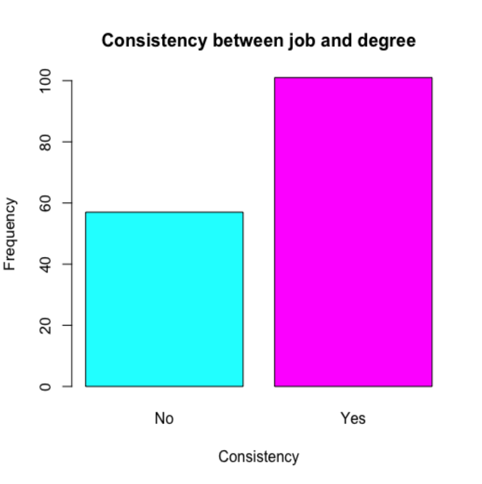


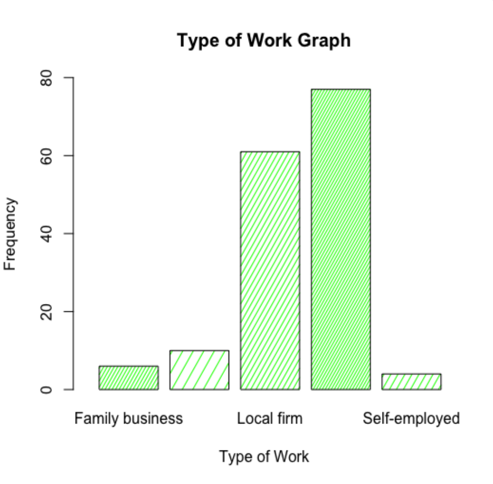


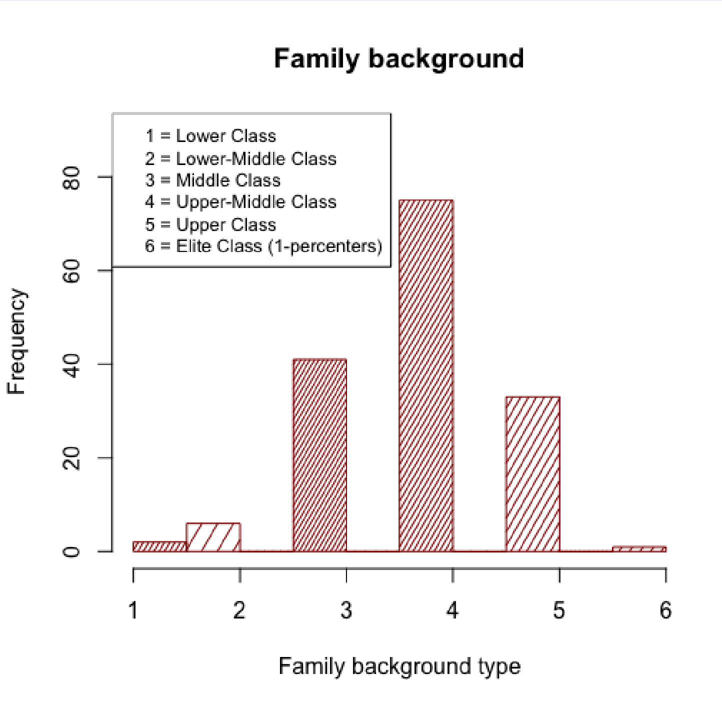


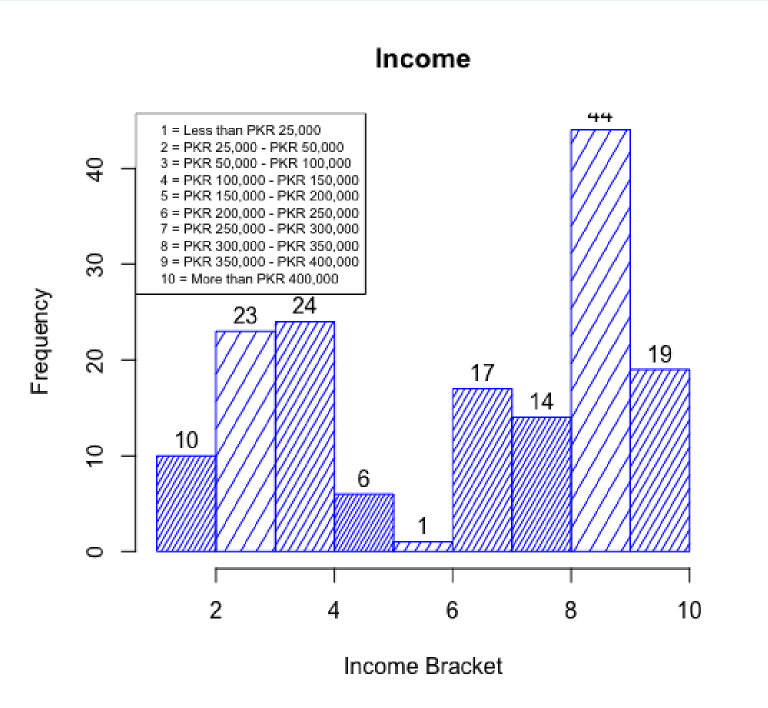


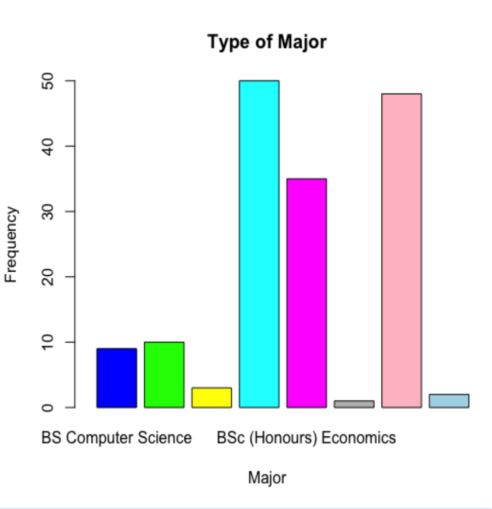


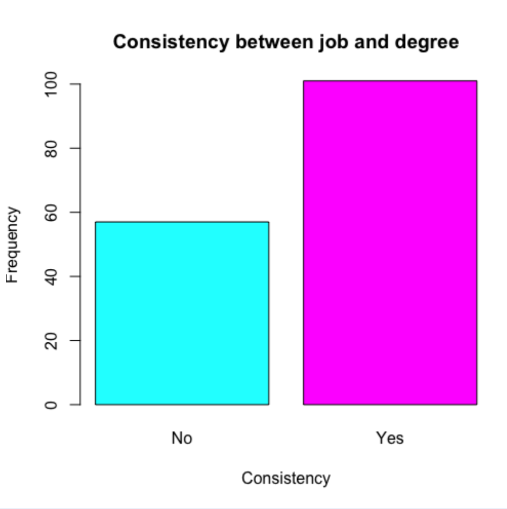


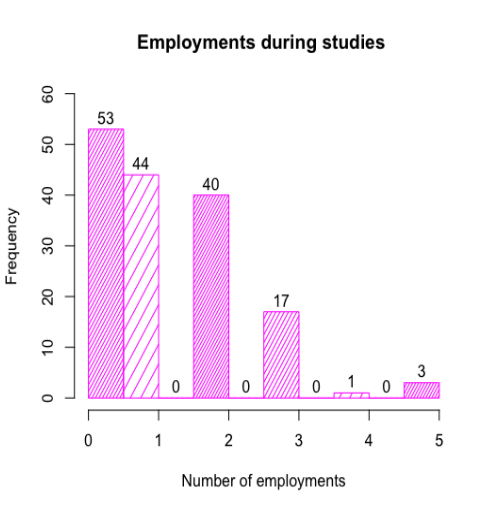



Family background
1 = Lower Class
2 = Lower-Middle Class
3 = Middle Class
4 = Upper-Middle Class
5 = Upper Class
6 = Elite Class (1-percenters)
Frequency
0
20
40
60
80
1
2
3
4
5
6
Family background type

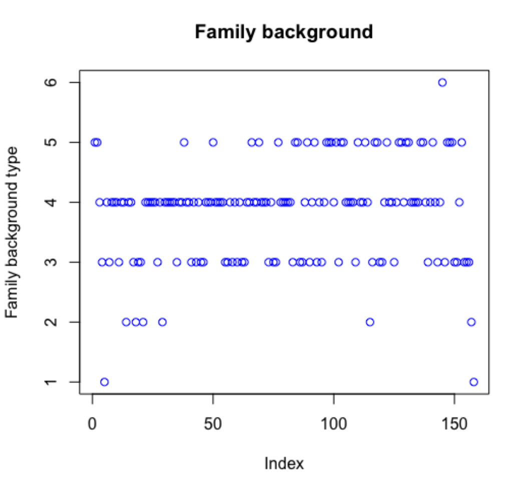

Income
1 = Less than PKR 25,000
2 = PKR 25,000 - PKR 50,000
3 = PKR 50,000 - PKR 100,000
4 = PKR 100,000 - PKR 150,000
5 = PKR 150,000 - PKR 200,000
6 = PKR 200,000 - PKR 250,000
7 = PKR 250,000 - PKR 300,000
8 = PKR 300,000 - PKR 350,000
9 = PKR 350,000 - PKR 400,000
10 = More than PKR 400,000
Frequency
0
10
20
30
40
10
23
24
6
1
17
14
44
19
2
4
6
8
10
Income Bracket

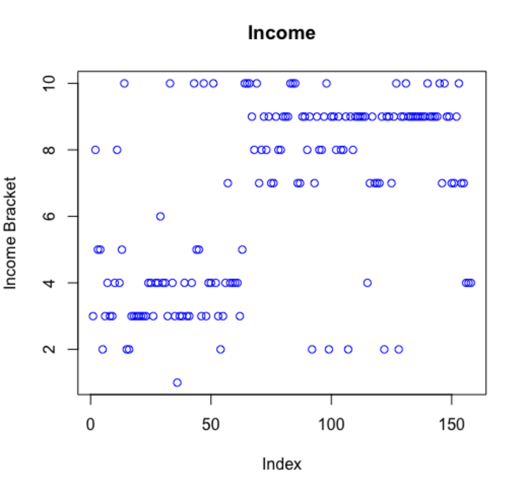

Type of Major
Frequency
0
10
20
30
40
50
BS Computer Science
BSc (Honours) Economics
Major

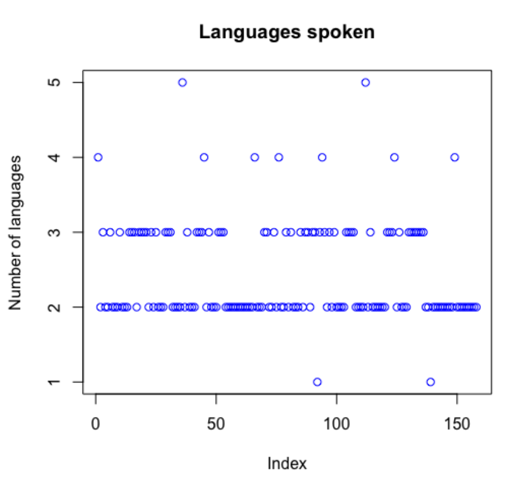

Consistency between job and degree
Frequency
0
20
40
60
80
100
No
Yes
Consistency

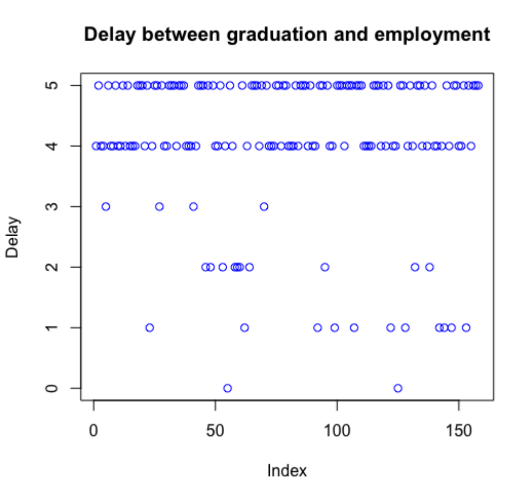

Employments during studies
Frequency
0
10
20
30
40
50
60
53
44
0
40
0
17
0
1
0
3
0
1
2
3
4
5
Number of employments

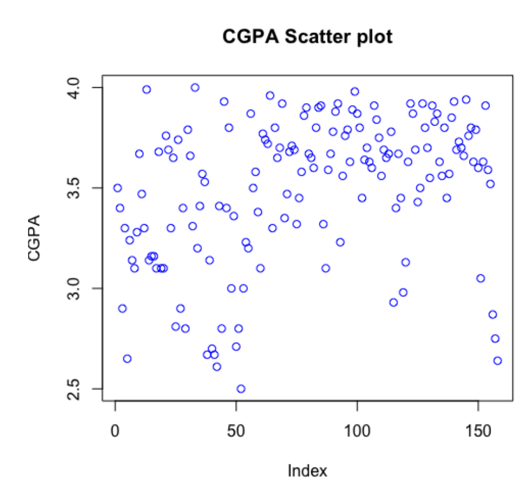
Scatter plots

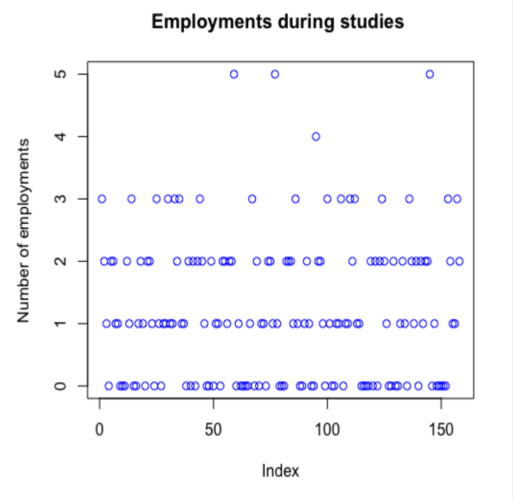


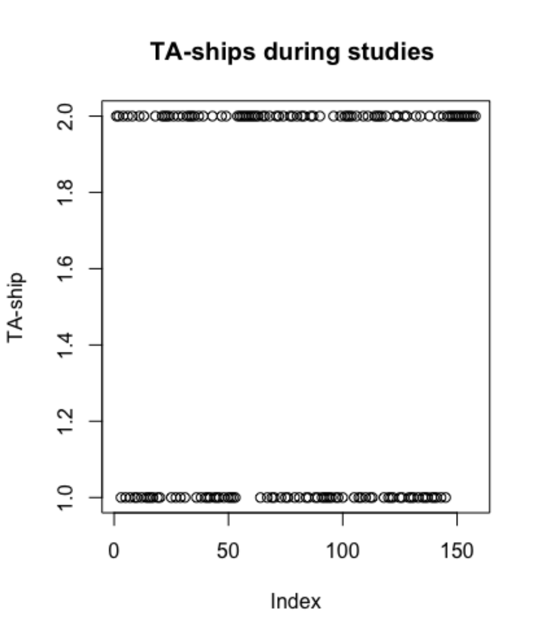




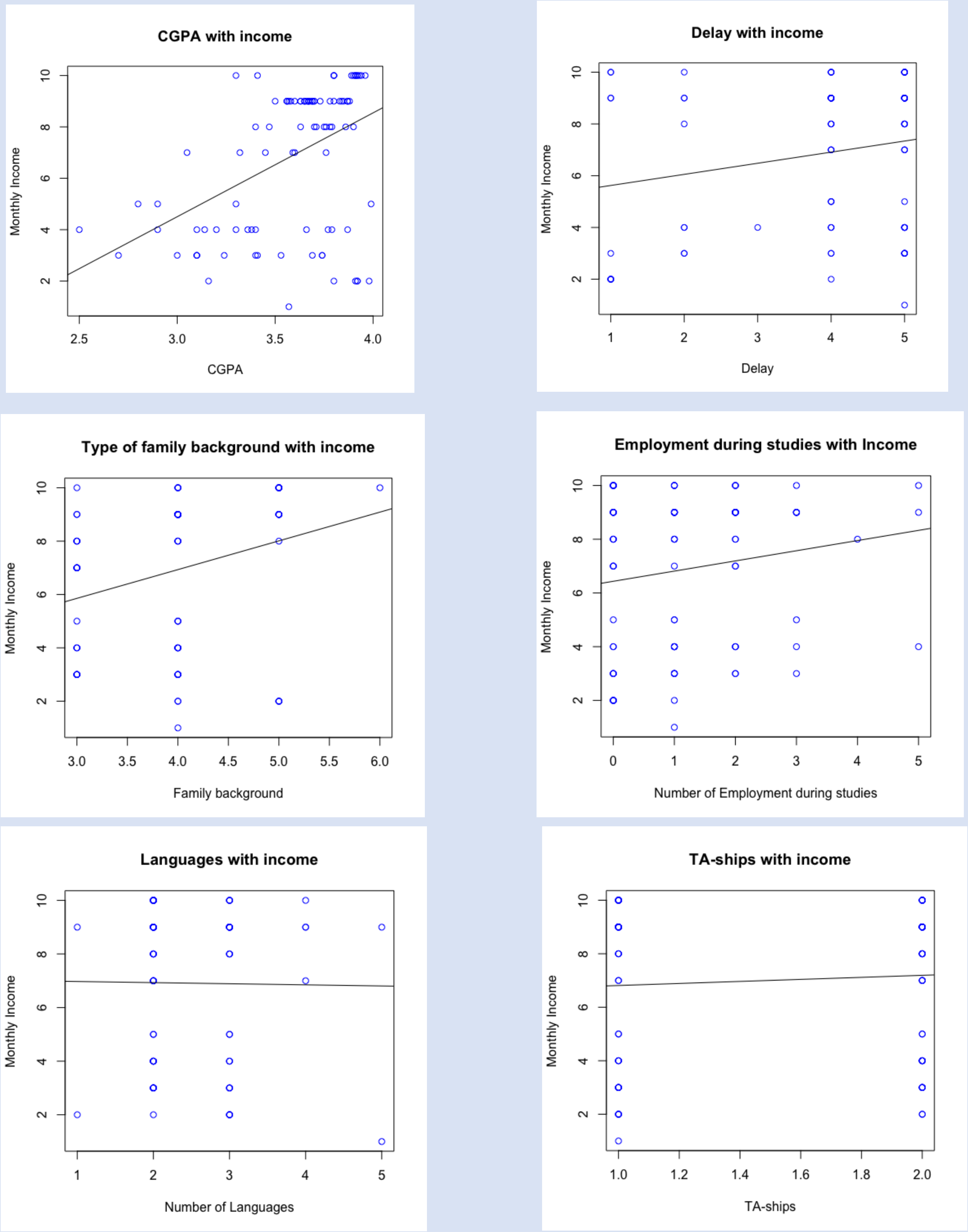




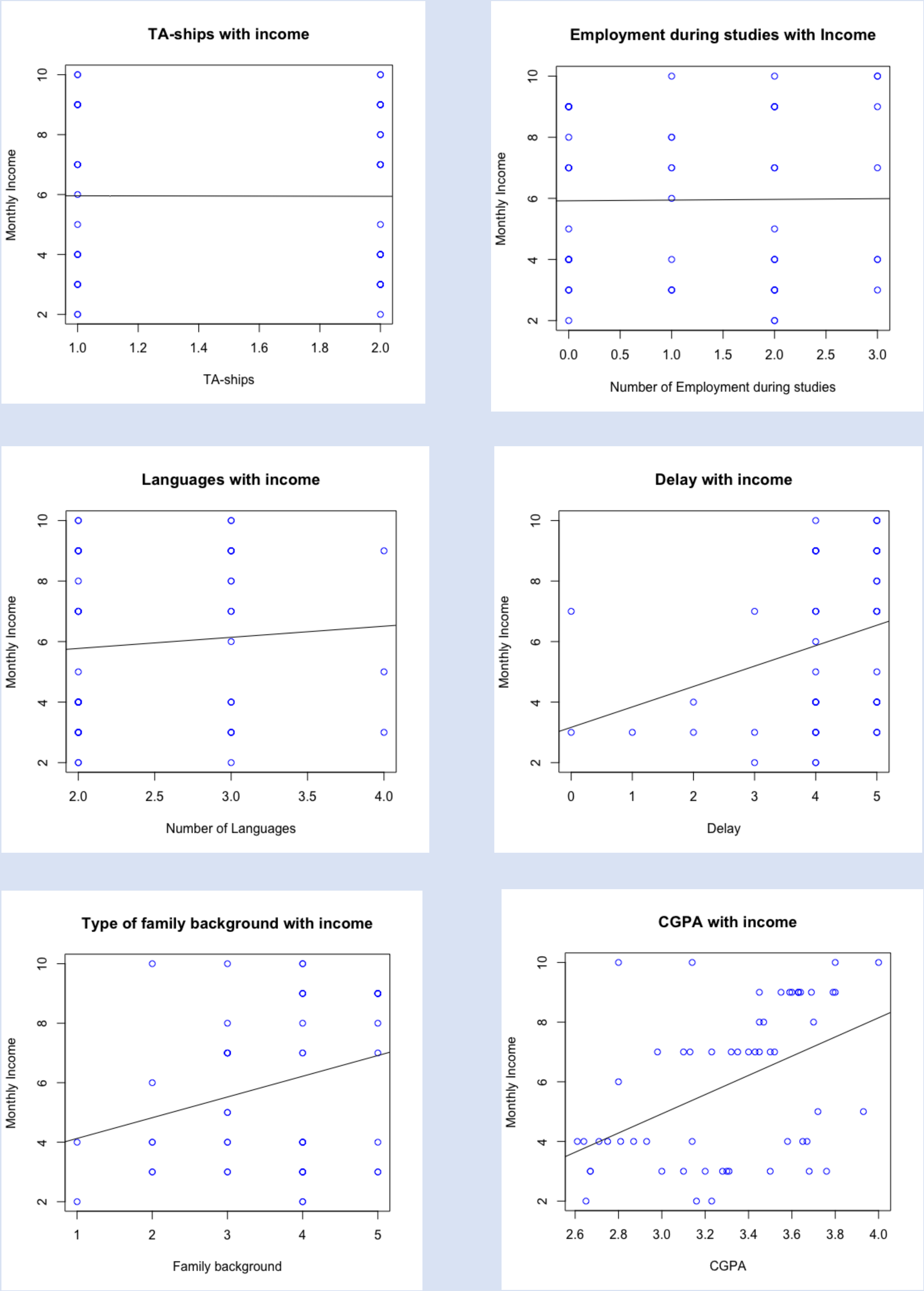




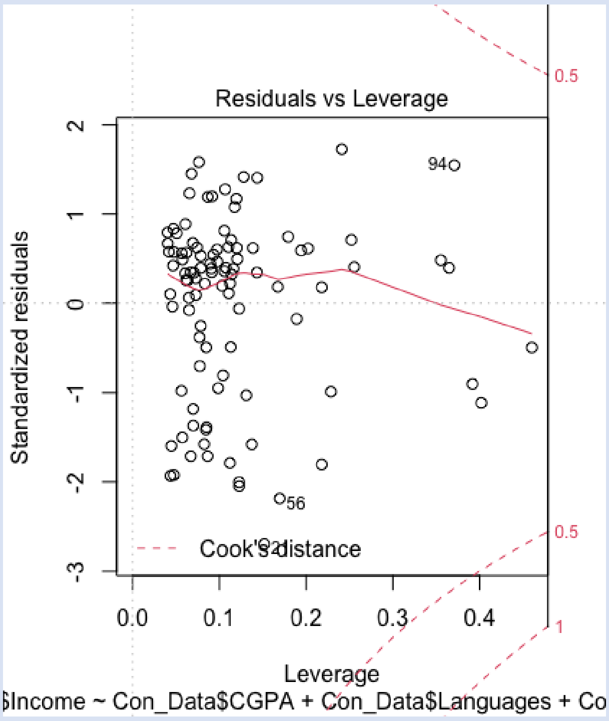
Scatter plots with Income – Model 1

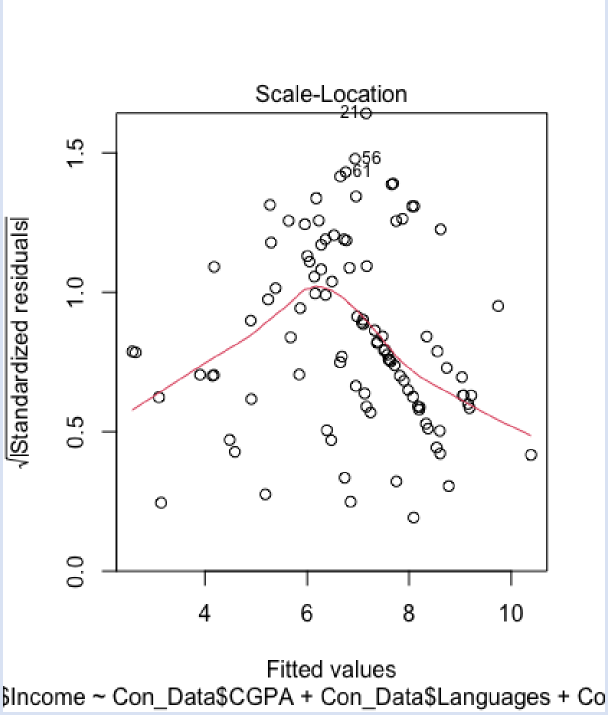


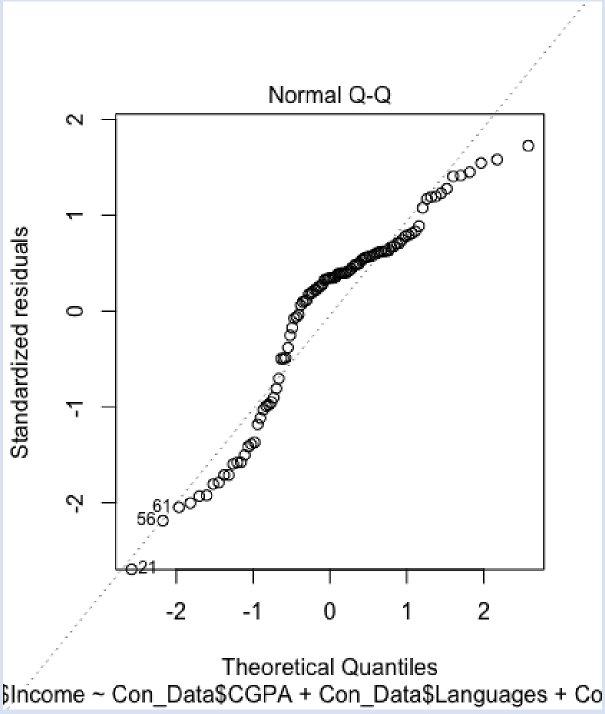
Scatter plots with Income – Model 2

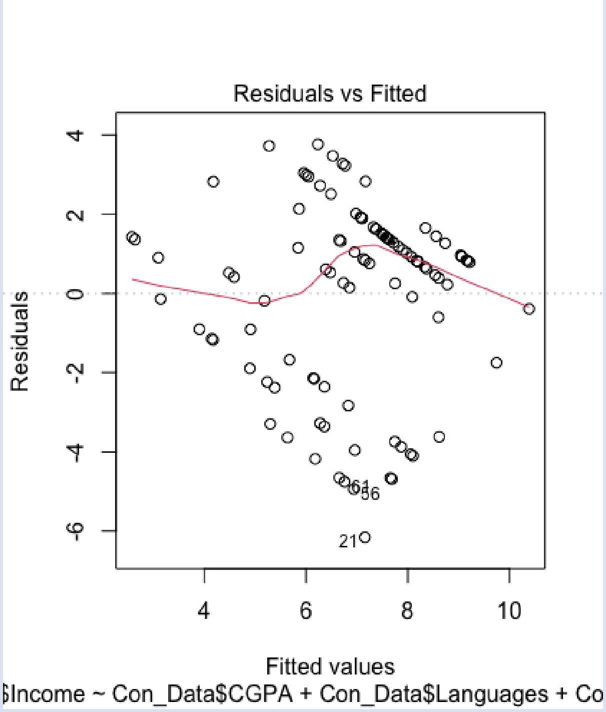


Multiple regression plots – Model 1

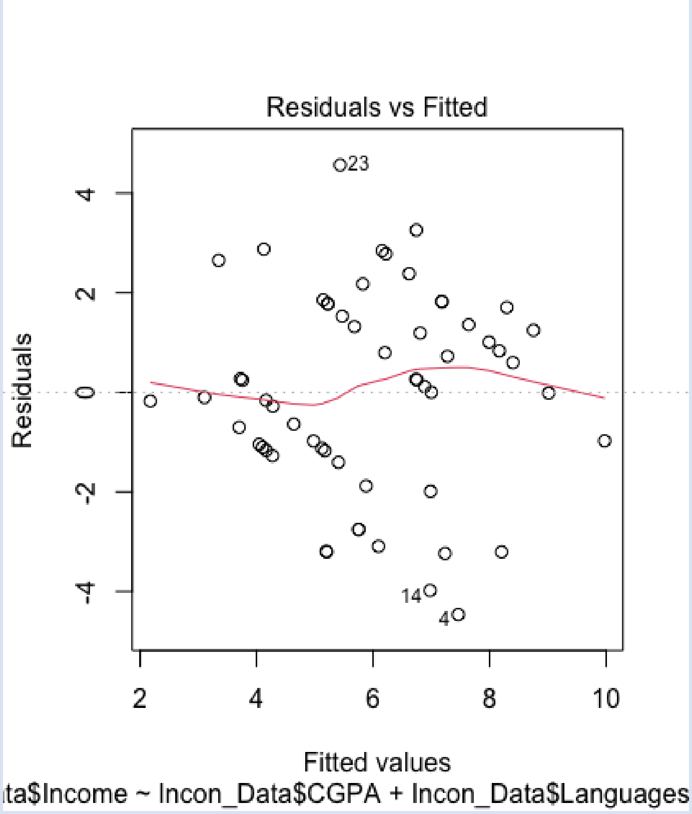


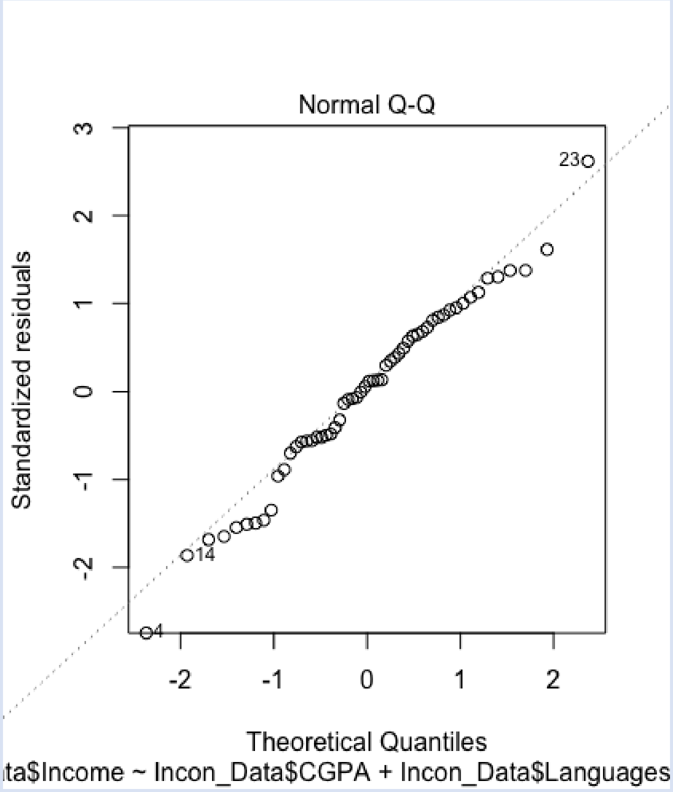


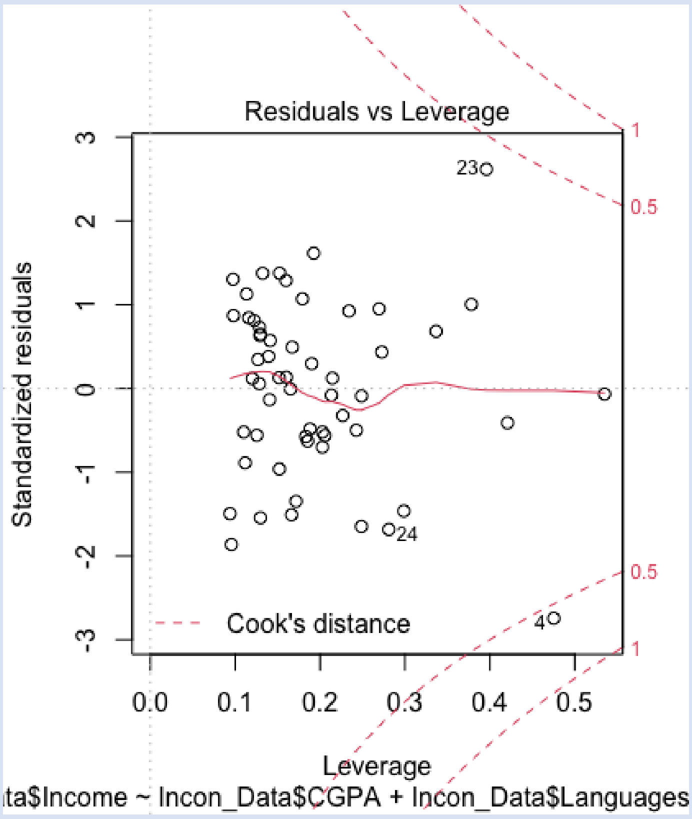


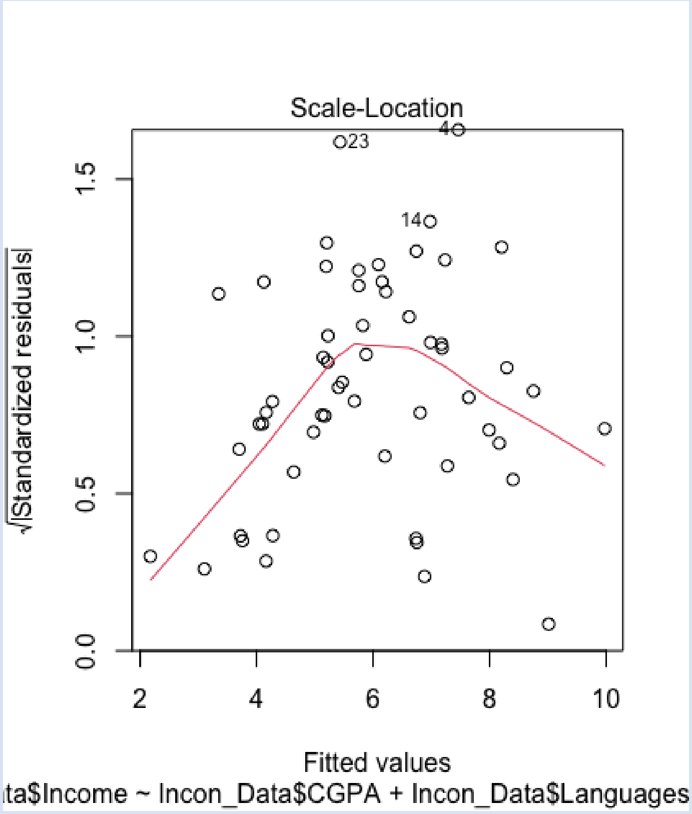


Multiple regression plots – Model 2









Links

Link to survey:

https://7uypgyy167i.typeform.com/Project-survey

Link to survey responses:

https://docs.google.com/spreadsheets/d/1QG7eFJkWc4kz20bu6x5XV5cLneN8GOtYvY_sABdorpU/edit#gid=42552357

R-Code

```
# install.packages("dplyr")
 library(dplyr)
 library(readxl)

# #Data Cleaning

# Dropping Columns not needed
Data_Collected_New <- Data_Collected[!names(Data_Collected) %in% c("Do you consent to participate in this survey?", "What year did you graduate from LUMS?", "Did you intend to complete this degree when you first joined LUMS?", "What was the reason?", "Are they (jobs, part-time jobs) related to your current job (if you are employed)?", "Do you have you a family business?", "Are you working in your preferred job market? (did you want to work in this sector?)", "Which company, do you work for?", "What is your job? (your position or a short job description)", "Why was there a delay?", "Please specify (name of the postgraduate degree and the institution)", "What is you age?", "What is your hometown?", "Would you be willing to give an interview for this research?", "What is your name and roll number?", "How do you wish to be contacted? (LUMS email address, or contact number) Please provide the information such as contact number if you wish to be contacted there.", "Token", "What is your gender?")]

# Filter rows of people unemployed and wrong values
Data_Collected_New <- Data_Collected_New[!Data_Collected_New$`Are you currently employed?` == "FALSE", ]
Data_Collected_New <- Data_Collected_New[!Data_Collected_New$`What was your CGPA at the time of graduation?` >4, ]

# Dropping Columns not needed

Data_Collected_New <- Data_Collected_New[!names(Data_Collected_New) %in% c("Are you currently employed?")]

# Change names of columns

Data_Collected_New <- rename(Data_Collected_New, "Major" = "What was your undergraduate degree/major?" )
Data_Collected_New <- rename(Data_Collected_New, "CGPA" = "What was your CGPA at the time of graduation?" )
Data_Collected_New <- rename(Data_Collected_New, "Employment during studies" = "How many (if applicable) job(s)/part-time job(s) did you have?")
Data_Collected_New <- rename(Data_Collected_New, "TA-ships" =  "Did you have any TA-ships during your undergraduate degree at LUMS?")
Data_Collected_New <- rename(Data_Collected_New, "Type of work" = "Where are you working?")
Data_Collected_New <- rename(Data_Collected_New, "Delay" = "Was there a time delay between your employment date and your graduation date?")
Data_Collected_New <- rename(Data_Collected_New, "Income" = "What is your monthly income bracket?" )
Data_Collected_New <- rename(Data_Collected_New, "Type of Education" = "What type of education did you complete before your undergraduate degree?")
Data_Collected_New <- rename(Data_Collected_New, "Family Background" = "How would you describe your family background?" )
Data_Collected_New <- rename(Data_Collected_New, "Languages" = "How many languages are you proficient in?")
Data_Collected_New <- rename(Data_Collected_New, "Consistency" = "Do you think that your degree and job are consistent? (for example an ACF major, working as an accountant)")

# Qualitative data to Quantitative data


Data_Collected_New$`TA-ships`[Data_Collected_New$`TA-ships` == "TRUE"] <- 2
Data_Collected_New$`TA-ships`[Data_Collected_New$`TA-ships` == "0"] <- 1

Data_Collected_New$Delay[Data_Collected_New$Delay == "None"] <- 5
Data_Collected_New$Delay[Data_Collected_New$Delay == "Less than a year"] <- 4
Data_Collected_New$Delay[Data_Collected_New$Delay == "1 year"] <- 3
Data_Collected_New$Delay[Data_Collected_New$Delay == "2 years"] <- 2
Data_Collected_New$Delay[Data_Collected_New$Delay == "3 years"] <- 1
Data_Collected_New$Delay[Data_Collected_New$Delay == "More than 3 years"] <- 0

Data_Collected_New$`Family Background`[Data_Collected_New$`Family Background` == "Lower class"] <- 1
Data_Collected_New$`Family Background`[Data_Collected_New$`Family Background` == "Lower-middle class"] <- 2
```

```
Data_Collected_New$`Family Background`[Data_Collected_New$`Family Background` ==
"Middle class"] <- 3
Data_Collected_New$`Family Background`[Data_Collected_New$`Family Background` ==
"Upper-middle class"] <- 4
Data_Collected_New$`Family Background`[Data_Collected_New$`Family Background` ==
"Upper class"] <- 5
Data_Collected_New$`Family Background`[Data_Collected_New$`Family Background` ==
"Elite class (1 percenters)"] <-  6

Data_Collected_New$Income[Data_Collected_New$Income == "Currently not employed"]
<- 0
Data_Collected_New$Income[Data_Collected_New$Income == "< PKR 25,000"] <- 1
Data_Collected_New$Income[Data_Collected_New$Income == "PKR 25,000 - PKR
50,000"] <- 2
Data_Collected_New$Income[Data_Collected_New$Income == "PKR 50,000 - PKR
100,000"] <- 3
Data_Collected_New$Income[Data_Collected_New$Income == "PKR 100,000 - PKR
150,000"] <- 4
Data_Collected_New$Income[Data_Collected_New$Income == "PKR 150,000 - PKR
200,000"] <- 5
Data_Collected_New$Income[Data_Collected_New$Income == "PKR 200,000 - PKR
250,000"] <- 6
Data_Collected_New$Income[Data_Collected_New$Income == "PKR 250,000 - PKR
300,000"] <- 7
```

```
Data_Collected_New$Income[Data_Collected_New$Income == "PKR 300,000 - PKR 350,000"] <- 8
Data_Collected_New$Income[Data_Collected_New$Income == "PKR 350,000 - PKR 400,000"] <- 9
Data_Collected_New$Income[Data_Collected_New$Income == "> PKR 400,000"] <- 10

# Summary

# Summary of Major

Data_Collected_New$Major <- as.factor(Data_Collected_New$Major)
summary(Data_Collected_New$Major)
plot(Data_Collected_New$Major, xlim = c(0,9), ylim = c(0,50), xlab = "Major", ylab = "Frequency", main = "Type of Major", col = c("Blue", "Green", "Yellow", "cyan", "magenta","grey", "Pink","lightblue"))



# Summary of CGPA

summary(Data_Collected_New$CGPA)
hist(Data_Collected_New$CGPA, main = "CGPA Chart", xlab = "CGPA", col = "dodgerblue3", labels = TRUE, density=c(30,10,20,35,15,25),angle=60)
plot(Data_Collected_New$CGPA, ylab = "CGPA", col = "Blue", main = "CGPA Scatter plot")
```

```
boxplot(Data_Collected_New$CGPA, ylab = "CGPA", col = "lightblue", main = "CGPA
Box plot")

# Summary of Employment during studies

Data_Collected_New$`Employment during studies` <-
as.numeric(Data_Collected_New$`Employment during studies`)
summary(Data_Collected_New$`Employment during studies`)
hist(Data_Collected_New$`Employment during studies`, xlab = "Number of employments",
main ="Employments during studies ", col = "magenta",labels = TRUE,
density=c(30,10,20,35,15,25),angle=60, ylim = c(0,60))
plot(Data_Collected_New$`Employment during studies`, ylab = "Number of employments",
main ="Employments during studies ", col = "Blue")

# Summary of the number of TA-ships

Data_Collected_New$`TA-ships` <- as.numeric(Data_Collected_New$`TA-ships`)
summary(Data_Collected_New$`TA-ships`)
hist(Data_Collected_New$`TA-ships`, xlab = "TA-ships", main ="TA-ship during studies ",
col = "black",labels = TRUE, density=c(30,10,20,35,15,25),angle=60, ylim = c(0,100))
plot(Data_Collected_New$`TA-ships`, ylab = "TA-ship", main ="TA-ships during studies ",
col = "black")
```

```
# Summary of type of work

Data_Collected_New$`Type of work` <- as.factor(Data_Collected_New$`Type of work`)
summary(Data_Collected_New$`Type of work`)
plot(Data_Collected_New$`Type of work`, xlim = c(0,6), ylim = c(0,80), xlab = "Type of Work", ylab = "Frequency", main = "Type of Work Graph", col = "green", density=c(30,10,20,35,15),angle=60)

# Summary of delay

Data_Collected_New$Delay <- as.numeric(Data_Collected_New$Delay)
summary(Data_Collected_New$Delay)
hist(Data_Collected_New$Delay,xlab = "Delay", main = "Delay between job and graduation", col = "red",labels = TRUE, density=c(30,10,20,35,15,25),angle=60)
plot(Data_Collected_New$Delay, col = "Blue", ylab = "Delay", main = "Delay between graduation and employment")
boxplot(Data_Collected_New$Delay, col = "Blue", ylab = "Delay", main = "Delay between graduation and employment")

# Summary of income bracket

Data_Collected_New$Income <- as.numeric(Data_Collected_New$Income)
summary(Data_Collected_New$Income)
hist(Data_Collected_New$Income, xlab = "Income Bracket", main = "Income", col = "Blue",labels = TRUE, density=c(30,10,20,35,15,25),angle=60)
```

```
legend("topleft", c("1 = Less than PKR 25,000","2 = PKR 25,000 - PKR 50,000", "3 = PKR
50,000 - PKR 100,000", "4 = PKR 100,000 - PKR 150,000", "5 = PKR 150,000 - PKR
200,000","6 = PKR 200,000 - PKR 250,000", "7 = PKR 250,000 - PKR 300,000", "8 = PKR
300,000 - PKR 350,000", "9 = PKR 350,000 - PKR 400,000", "10 = More than PKR
400,000"), cex = 0.6)
plot(Data_Collected_New$Income, ylab = "Income Bracket", main = "Income", col =
"Blue")

# Summary of type of education

Data_Collected_New$`Type of Education` <- as.factor(Data_Collected_New$`Type of
Education`)
summary(Data_Collected_New$`Type of Education`)
plot(Data_Collected_New$`Type of Education`,xlab = "Type of education", main = "Type of
education", col = c("cyan","magenta"), ylim = c(0,120))

# Summary of family background

Data_Collected_New$`Family Background` <- as.numeric(Data_Collected_New$`Family
Background`)
summary(Data_Collected_New$`Family Background`)
hist(Data_Collected_New$`Family Background`,xlab = "Family background type", main =
"Family background", col = "darkred", ylim = c(0,90),density=c(30,10,20,35,15),angle=60)
legend("topleft", c("1 = Lower Class","2 = Lower-Middle Class", "3 = Middle Class","4 =
Upper-Middle Class","5 = Upper Class", "6 = Elite Class (1-percenters)" ), cex = 0.8)
```

```
plot(Data_Collected_New$`Family Background`, col = "Blue", ylab = "Family background type", main = "Family background")
boxplot(Data_Collected_New$`Family Background`, col = "Blue", ylab = "Family background type", main = "Family background")

# Summary of the number of languages spoken

Data_Collected_New$Languages <- as.numeric(Data_Collected_New$Languages)
summary(Data_Collected_New$Languages)
hist(Data_Collected_New$Languages,xlab = "Number of languages", main = "Languages spoken", col = "slateblue4",density=c(30,10,20,35,15),angle=60)
plot(Data_Collected_New$Languages,ylab = "Number of languages", main = "Languages spoken", col = "Blue")

# Summary of job-degree consistency

Data_Collected_New$Consistency <- as.factor(Data_Collected_New$Consistency)
summary(Data_Collected_New$Consistency)
plot(Data_Collected_New$Consistency,xlab = "Consistency", ylab = "Frequency", main = "Consistency between job and degree", col = c("cyan","magenta"))
```

```
# Separating Consistent and Inconsistent data

Incon_Data <- Data_Collected_New[!Data_Collected_New$Consistency == "Yes", ]
Con_Data <- Data_Collected_New[!Data_Collected_New$Consistency == "No", ]

# Consistent Data

# CGPA with Income

firstreg = lm(formula = Con_Data$Income ~ Con_Data$CGPA)
summary(firstreg)
cor(Con_Data$Income,Con_Data$CGPA)

plot(Con_Data$Income ~ Con_Data$CGPA , main = "CGPA with income",ylab = "Monthly Income", xlab = "CGPA", col = "Blue")
abline(firstreg)

# Employment during studies with Income

secreg = lm(formula = Con_Data$Income ~ Con_Data$`Employment during studies`)
summary(secreg)
cor(Con_Data$Income,Con_Data$`Employment during studies`)
```

```
plot(Con_Data$Income ~ Con_Data$`Employment during studies` , main = "Employment
during studies with Income",xlab = "Number of Employment during studies", ylab =
"Monthly Income", col = "Blue")
abline(secreg)

# TA-ships with Income

thirdreg = lm(formula =  Con_Data$Income ~ Con_Data$`TA-ships`)
summary(thirdreg)
cor(Con_Data$Income,Con_Data$`TA-ships`)

plot(Con_Data$Income ~ Con_Data$`TA-ships`, main = "TA-ships with income",xlab =
"TA-ships", ylab = "Monthly Income", col = "Blue")
abline(secreg)

# Delay with Income

fourthreg = lm(formula =  Con_Data$Income ~ Con_Data$Delay)
summary(fourthreg)
cor(Con_Data$Income,Con_Data$Delay, use = "complete.obs")

plot(Con_Data$Income ~ Con_Data$Delay, main = "Delay with income",xlab = "Delay",
ylab = "Monthly Income", col = "Blue")
abline(fourthreg)
```

```
# Family background with income
sevreg = lm(formula =  Con_Data$Income ~ Con_Data$`Family Background`)
summary(sevreg)
cor(Con_Data$Income,Con_Data$`Family Background`)

plot(Con_Data$Income ~ Con_Data$`Family Background`, main = "Type of family
background with income",xlab = "Family background", ylab = "Monthly Income", col =
"Blue")
abline(sevreg)

# Languages with income

eigreg = lm(formula =  Con_Data$Income ~ Con_Data$Languages)
summary(eigreg)
cor(Con_Data$Income,Con_Data$Languages)

plot(Con_Data$Income ~ Con_Data$Languages, main = "Languages with income",xlab =
"Number of Languages", ylab = "Monthly Income",  col = "Blue")
abline(eigreg)

# MultiRegression

multireg = lm(formula = Con_Data$Income ~ Con_Data$CGPA + Con_Data$Languages +
Con_Data$`Family Background` + Con_Data$Delay + Con_Data$`TA-ships` +
Con_Data$`Employment during studies` + Con_Data$`Type of work`+ Con_Data$`Type of
Education`)
summary(multireg)

# Regression with CGPA - Family background - Delay with income

model1areg = lm(formula = Con_Data$Income ~ Con_Data$CGPA + Con_Data$`Family
Background` + Con_Data$Delay)
summary(model1areg)

# Inconsistent Data

# CGPA with Income

firreg = lm(formula = Incon_Data$Income ~ Incon_Data$CGPA)
summary(firreg)
cor(Incon_Data$Income,Incon_Data$CGPA)

plot(Incon_Data$Income ~ Incon_Data$CGPA , main = "CGPA with income",ylab =
"Monthly Income", xlab = "CGPA", col = "Blue")
abline(firreg)

# Employment during studies with Income

seccreg = lm(formula = Incon_Data$Income ~ Incon_Data$`Employment during studies`)
```

```
summary(seccreg)
cor(Incon_Data$Income,Incon_Data$`Employment during studies`)

plot(Incon_Data$Income ~ Incon_Data$`Employment during studies` , main = "Employment during studies with Income",xlab = "Number of Employment during studies", ylab = "Monthly Income", col = "Blue")
abline(seccreg)

# TA-ships with Income

thirddreg = lm(formula =  Incon_Data$Income ~ Incon_Data$`TA-ships`)
summary(thirddreg)
cor(Incon_Data$Income,Incon_Data$`TA-ships`)

plot(Incon_Data$Income ~ Incon_Data$`TA-ships`, main = "TA-ships with income",xlab = "TA-ships", ylab = "Monthly Income", col = "Blue")
abline(thirddreg)

# Delay with Income

fourreg = lm(formula =  Incon_Data$Income ~ Incon_Data$Delay)
summary(fourreg)
cor(Incon_Data$Income,Incon_Data$Delay)
```

```
plot(Incon_Data$Income ~ Incon_Data$Delay, main = "Delay with income",xlab = "Delay",
ylab = "Monthly Income", col = "Blue")
abline(fourreg)

# Family background with income
sevvreg = lm(formula =  Incon_Data$Income ~ Incon_Data$`Family Background`)
summary(sevvreg)
cor(Incon_Data$Income,Incon_Data$`Family Background`)

plot(Incon_Data$Income ~ Incon_Data$`Family Background`, main = "Type of family
background with income",xlab = "Family background", ylab = "Monthly Income", col =
"Blue")
abline(sevvreg)

# Languages with income

eiggreg = lm(formula =  Incon_Data$Income ~ Incon_Data$Languages)
summary(eiggreg)
cor(Incon_Data$Income,Incon_Data$Languages)

plot(Incon_Data$Income ~ Incon_Data$Languages, main = "Languages with income",xlab =
"Number of Languages", ylab = "Monthly Income",  col = "Blue")
abline(eiggreg)

# MultiRegression
```

```
multiireg = lm(formula = Incon_Data$Income ~ Incon_Data$CGPA +
Incon_Data$Languages + Incon_Data$`Family Background` + Incon_Data$`Type of work` +
Incon_Data$Delay + Incon_Data$`Type of Education` + Incon_Data$`TA-ships` +
Incon_Data$`Employment during studies`)
summary(multiireg)
plot(multiireg)

# Regression with CGPA - Family background - Delay with income

model2areg = lm(formula = Incon_Data$Income ~ Incon_Data$CGPA +
Incon_Data$`Family Background` + Incon_Data$Delay)
summary(model2areg)
```