Bi Variate Analysis

Habibullah, Perozi
Statistics 1/ITC 255
Fall Semester 2020

1. Introduction:

For this assignment, we use a collection of data from Wooldridge on 786 economics faculty at a university. According to Wooldridge, the data were available for each faculty member for as many as three years. The main purpose of this assignment is to study the associations between selected variables.

The data is collected by Professors Baser and Pema on 30 variables. But for the sake of our own assignment here, we just extracted 5 variables of them which are *gende*r, *salary, fullTimePro, yearPhdObtained,* and *totalArticlePages*. This collection of data includes all economics faculty at the professor Pema's university. We implement some bi-variate analysis techniques on this dataset to find a precise answer for the following questions.

1. Is there any association between full time professors and their gender among all economics faculty of the university?
2. Is the year of getting Ph.D. effects being a full-time professor at the faculty?
3. Is there any relation between the number of article pages that faculty members wrote and their annual salary?

2. Methodology:

2.1. Population and data:

The population includes all of the economic faculty of the university from professors to assistants and other faculty members. This university has other faculties and staff that are not included in this data set. The total amount of observers in this data through a period of time in 1992, 1995, and 1999 were 786. We extracted the needed dataset for our assignment and renamed some variables. Although the data seems clean there were NA values that were cleaned before analysis.

2.2. Analysis:

The techniques that are used for the analysis of the data through univariate techniques are specifically graphical methods and numerical methods. Furthermore, we use bi var techniques as well. Two-way table, graphical methods, t-stat, numerical methods, and liner relation model are used to analyze this dataset.

3. Results:

3.1. Summary statistics:

In this section, we discuss the population characteristics. Table 1 presents the summary statistics of the quantitative variables. Figure 1 and figure 2 shows the graphical representation of qualitative variables.

Table 1. Summary statistics of the quantitative variables

	salary	yearPhdObtained	totalArticlePages
Min.	18670	38	0
1st Qu	63051	70	86.25
Median	80297	77	144.59
Mean	83786	76.56	179.97
3rd Qu.	98737	83.75	233.16
Max.	191000	96	1060.5
St.Dev	27566.64	10.1194	146.4973

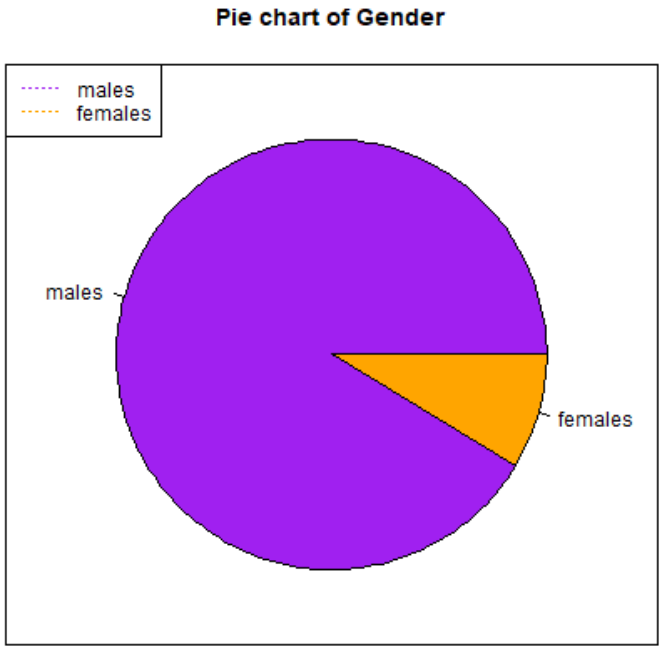


Figure 1. Distribution of Gender

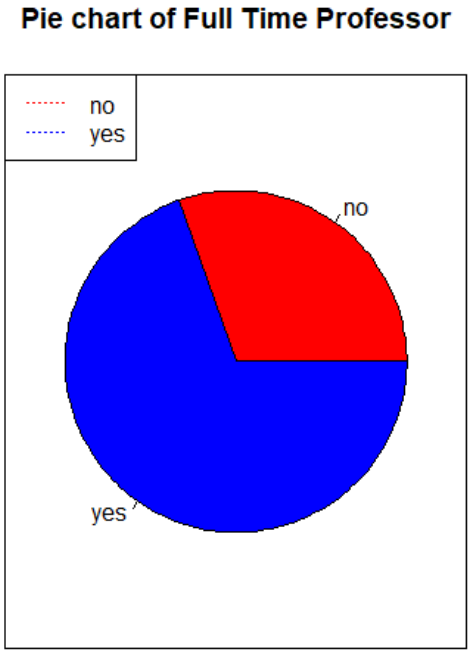


Figure 2. Distribution of Full Time Professors

The summary of quantitative variables indicates that the average salary among faculty is around 80 to 84 thousand. Also, some faculty members are very old and got their Ph.D. early which the earliest among all is in 1938. But in general, most of them got their Ph.D. around 18 years ago from the time this data was aggregated. Some of the faculty members haven't written any articles that they may be assistant or other employees within the economics faculty. The summary also indicates that there are high variations from mean for all quant vars except the year of their Ph.D.

As we can see on the Figure 1 and Figure 2, most of the economics faculty are males. There are part time professors a little more than ¼ that most of the economics faculty are males and most of the professors are full time.

3.2. Bi variate Analysis

This section discusses the association of the concerned variable by applying different bi variate analysis.

First, we want to find the association between full-time professors and their gender through a two-way table.

	0	1	marginal_gender
f	31	16	47
m	134	361	495
marginal_fullTimePro	165	377	542

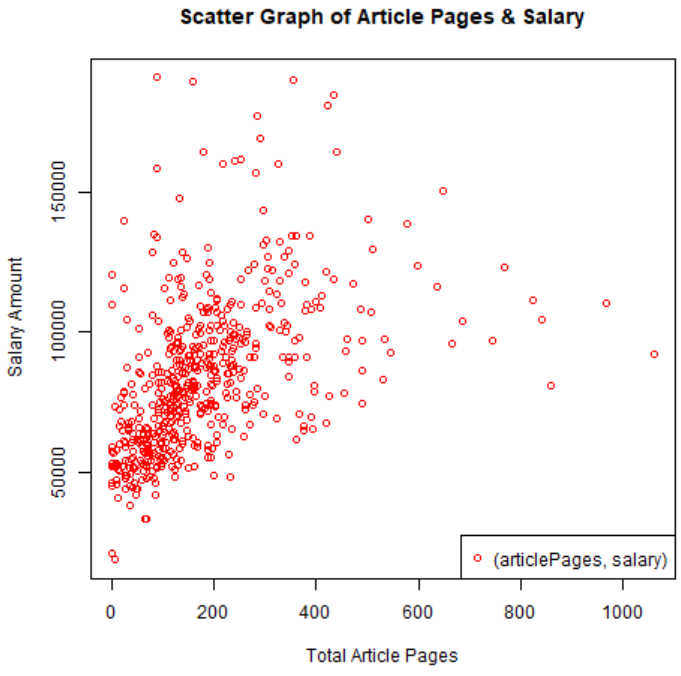
As it indicates, 1/3 of female faculty members are full-time professors and the rest are part-time. 361 which is a big number among all are full-time male professors which are near 2/3 of all of the professors at the economics faculty. This table declares a huge relation in terms of employment among male and female professors and in general, most of the professors are full-time when it comes to gender males, however, it's vice versa for the female professors.

From our dataset, we want to calculate the t-stat among the full-time professors and the year that they obtained their Ph.D.

We find that the t-stat is 17.9 which indicates a very strong association. It also shows that those professors who got their Ph.D. earlier are full time now, and their average year of obtaining a Ph.D. is 1972. On the other hand, those professors that aren't full-time got their Ph.D. with the average year of 1985. Therefore, we should consider the year of getting a Ph.D. when it comes to a full-time professor at the economics faculty.

Another point to analyze is the relation between the faculty member's number of articles that they wrote and their annual salary. Here is their scatter plot.

The scatter plot shows a medium linear relationship between the total article pages and the salary of the faculty members. The correlation value is 0.483 that indicates a medium association among these two quant variables. Although there are some values very far from the least square line, it's okay to build its liner model.



TotalArticlePages are independent values as x, and the *salary* var is dependent as y. Using y = a+bx and lm function in R, we find a=67406 and b=91. Here is its linear model:
*salary = 67406+91*totalArticlePages*

According to this model, for each one-page article, there will be 91 dollars added to the economics faculty member's salary.

Appendix:

```
setwd('F:/AUAF FALL 2020/ITC255/R')
library(reshape)
library(wooldridge)
wool = data.frame(wooldridge::big9salary)

#extracting our needed data
d = data.frame(wool[12],wool[3], wool[8], wool[11], wool[5])
#rename our vars
d=rename(d, c(female="gender",salary ="salary",
prof="fullTimePro",yearphd="yearPhdObtained",totpge="totalArticlePages" ))

#cleaning NA values
d_clean = na.omit(d)
#so now we analyze our clean data or implement our Tasks
summary(d_clean$salary)
sd(d_clean$salary)
sd(d_clean$yearhdObtained)
sd(d_clean$totalArticlePages)

aFre = table(d_clean$fullTimePro)
rFre = prop.table(aFre)
ecFre = cumsum(rFre)
tg = cbind(aFre, rFre, ecFre)
disTable = data.frame(tg)
show(disTable)
View(d_clean)

pie(rFre
   ,labels = c("no", "yes")
   ,col = c("red","blue")
   ,main = "Pie chart of Full Time Professor")
legend("toplef"
      ,legend = c("no","yes")
      ,col = c("red","blue")
      ,lty = 3)
box(which = "plot", lty = 1)

###TASKS:
##1. Consider two QL vars from your dataset and construct their two-way table.
#for this task I selected "gender" and "fullTimePro" as my Qual vars

t = table(d_clean$fullTimePro, d_clean$gender)
t
#both values of both variables are 0s and 1s;
#therefore we change the gender values into "f" & "m"
gen = c()
for(i in 1 :length(d_clean$gender)){
  if(d_clean$gender[i]==1){
    gen[i]="f"
```

```
  }else{
    gen[i]="m"
  }
}


d_clean = cbind(d_clean, gen)

#so now good to go
t = table(d_clean$gen, d_clean$fullTimePro)
t
marginal_fullTimePro = colSums(t)
marginal_fullTimePro
t1 = rbind(t,marginal_fullTimePro)
t1
marginal_gender = rowSums(t1)
t2 = cbind(t1, marginal_gender)
View(t2)
write.csv(t2, file = "gender&fullTimePro_two-way-table.csv")

##2. Consider 1 QL nd 1 QN vars in your dataset and calculate their t-stat/F-stat.
#for this task we consider the "fullTimePro" Qual var and the "yearPhdObtained" Quant var
t.test(d_clean$yearPhdObtained~d_clean$fullTimePro)
#we should consider fullTimePro as a factor when it comes as to yearsthephd was obtained by

##3. Consider 2 QN vars in your dataset: plot their scatter graph, calculate their core and build thei linear
model.
#for this task we consider these Quant vars:  "totalArticlePages" & "salary"
#plotting their scatter graph
png(filename = "Scatter Graph of Article Pages & Salary.png")
plot(d_clean$totalArticlePages, d_clean$salary
    , main = "Scatter Graph of Article Pages & Salary"
    , xlab = "Total Article Pages"
    , ylab = "Salary Amount"
    , col = ("red")
    , pch = 1)
legend("bottomright", legend = "(articlePages, salary)", col = "red", pch=1)
dev.off()
cor(d_clean$totalArticlePages, d_clean$salary)
#hence its liner y = a+bx
# x = totalArticlePages   y = salary

l_s_Salary = lm(d_clean$salary~d_clean$totalArticlePages)
summary(l_s_Salary)

# a = 67406  b = 91.014
#salary =  67406+91*totalArticlePages
```