KU Leuven

Project

Salary Gender Discrimination II

Analysis of Variance

[G0S76a]

Group 7

Gabriel Bénédict
gabriel.benedict@student.kuleuven.be - r0692805
Laura Juliana Guerrero
laurajuliana.guerrerovelasquez@student.kuleuven.be - r0689435
Daniel Gil Sanchez
daniel.gilsanchez@student.kuleuven.be - r0689432
Lyse Naomi Wamba Momo
lysenaomi.wambamomo@student.kuleuven.be - r0653272

Supervised by Prof. Dr. Ariel Alonso Abad
ariel.alonsoabad@kuleuven.be

December 21, 2017

Contents

List of Figures

List of Tables

1 Problem Statement

There is evidence to suggest that women now outperform men in many dimensions of educational achievement (such as college enrollment, high school graduation, attainment of college degrees) and that this tendency is likely to increase in the following years (Bobbitt-Zeher, 2007, p. 2). The purpose of this study is to assess whether those events that have been proven to take place for more than a decade now, do have an influence on the workplace later on; i.e. whether salary is different depending on gender and/or education. A random sample of 132 employees on a particular firm is selected and gender, salary and educational information is collected from them.

2 Descriptive Statistics

`Gender` is categorized as *males* and *females*, `Education` as *Degree* or *No degree*. For this sample, salaries go from 11.84 to 35.61 thousands of dollars per year.

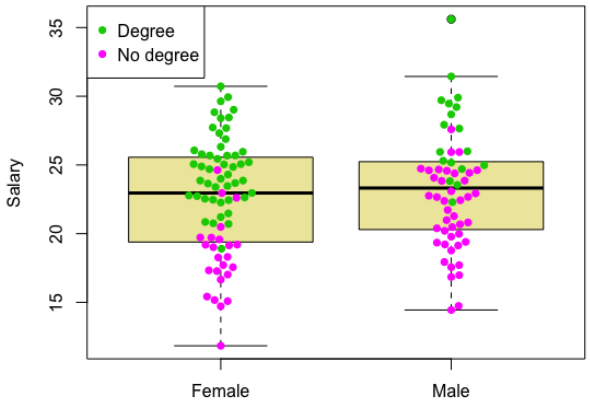

Figure 1: Boxplot by Gender

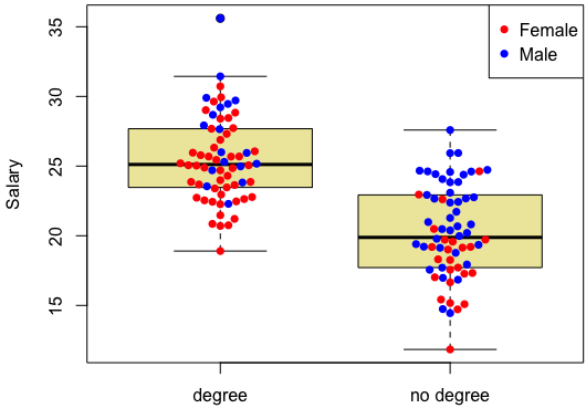

Figure 2: Boxplot by Education

On the perspective of the `Gender` boxplots (see figure 1[1]), the `salary` medians are similar: men slightly higher median is partly due to the large outlier (with degree) and the minimum for females is lower than for males. For both gender types, there is a clear segmentation for salary between individuals with a degree (D) and those without (ND). The overall mean and variability per gender seems to be similar as well.

Turning to the `Education` perspective (see figure 2), there is an evident gap for salary between those having a degree and those who don't. In this data set, there are more men without degree and more women with a degree. However, inside each group men tend to populate the higher tale of the salary distribution, while women dominate the lower tale.

Table 1 shows the descriptive statistics for each combinations of the variables `Gender` and `Education`, defined from now on as treatments. The first and probably the more important element that can be concluded from the table is that the study is unbalanced. There are more than two times females with degree than males and the opposite case is seen in No degree (as shown in figure 2). If a proper stratified random sampling was conducted, taking into account gender and education of the employees; it can be inferred that women are more educated than men in this firm.

[1]Note that figures 1 and 2 are computed with the `beeswarm` package that allows for displaying non-overlapping points respecting vertical scaling.

In addition, table 1 confirms what it is seen in the figures shown above, namely that average salary for educated subjects is larger than the non-educated and men average salary is larger than for women within each group. Low discrepancies between all cell standard deviations might imply homoscedasticity between educational levels and genders.

	Count		Mean		SD	
	Degree	No Degree	Degree	No Degree	Degree	No Degree
Female	48	24	24.81	18.28	2.68	2.83
Male	18	42	27.30	21.44	3.30	3.08

Table 1: Descriptive Statistics

Finally, the interaction between factors is analyzed in figure 3. No interaction seems to take place between factors, looking on the parallel behaviour. However, non-negligible average effects are likely to be observed for both gender and education. The general trend stays the same: males earn more than females and educated subjects earn more than non-educated ones.

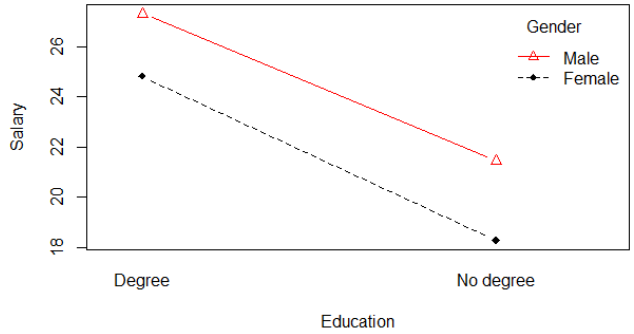

Figure 3: Interaction Plot

In order to draw conclusions on whether indeed no interaction exists and whether indeed the differences and similarities hinted in descriptive statistics are significantly relevant, Two-way Analysis of Variance is performed.

3 Assumptions

The Two-way Analysis of Variance model assumes constant variance among the groups, independent errors and univariate normal distribution of the response variable. It is further assumed that the sample sizes reflect the importance of the treatment means in the population.

4 Methodology

The principal purpose of this study is to find evidence for or against the fact that there is salary gender discrimination, i.e. **the average salaries received by the employees under the same conditions, educational levels in this case, are different for both genders**. In table 2 the notation of the experiment is defined.

	Degree	No Degree	
Female	μ_{FD}	μ_{FND}	$\bar{\mu}_{F.}$
Male	μ_{MD}	μ_{MND}	$\bar{\mu}_{M.}$
	$\bar{\mu}_{.D}$	$\bar{\mu}_{.ND}$	

Table 2: Two way ANOVA - Notation for the Treatment means

Taking into account the notation given in the table above, the following hypothesis is tested in order to conclude if the average salary is significantly different between genders:

$$H_0 : \ \bar{\mu}_{F.} = \bar{\mu}_{M.} \tag{1}$$

where the treatment means μ are estimated using the weighted means taking into account the `Education` factor.
As it was mentioned before, it is assumed that the sample size of each treatment reflects how the population of the firm is divided taking into account `Gender` and `Education`. Therefore, the weights are calculated as the proportion of the treatment sample sizes and so the hypothesis 1 can be rewritten as follows:

$$H_0 : \ \frac{48}{72}\bar{\mu}_{FD} + \frac{24}{72}\bar{\mu}_{FND} = \frac{18}{60}\bar{\mu}_{MD} + \frac{42}{60}\bar{\mu}_{MND} \tag{2}$$

The short-coming here is that this hypothesis only gives an idea of the average salary for each gender: which gender has on average the highest salary. But the effect of educational level is not taken into account in the estimation of these means (this issue raises the importance to use an alternative method in trying to answer if there is indeed salary gender discrimination).

The results are displayed in table 3 showing that the combined effects of gender and level of education do not have any influence on the average salary an employee receives (p-value $= 0.537 > 0.05$). This goes in line with the interaction plot displayed in figure 3. Also, the main effect of gender is not significant (p-value $= 0.271 > 0.05$) indicating that the average salary received by each gender type taking into account their sample proportion is not significantly different.

	Df	Sum Sq	F-value	Pr(>F)
Gender	1	10.5	1.224	0.271
Education	1	1112.1	130	$< 2e-16$
Gender:Education	1	3.3	0.383	0.537

Table 3: ANOVA Results (Type I SS)

It is important to notice that the validity of the results obtained in this model are based on the three assumptions mentioned. This diagnosis can be seen in the following figures.

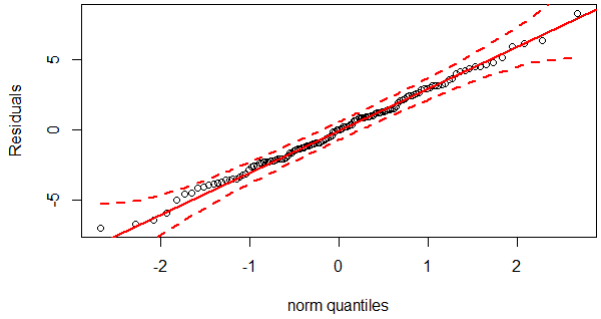

Figure 4: Normal Quantile plot

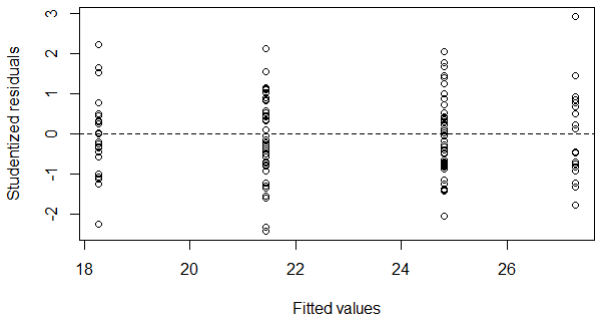

Figure 5: Residuals vs Fitted Values

Both qqplot in figure 4 and Shapiro Wilk test (W = 0.9956, p-value = 0.9614) do not give evidence to reject normality for the residuals[2]. In addition, from figure 5, it can be concluded that the range of the residuals is approximately the same for all the groups, indicating equal variance. Brown-Forsythe test corroborates with this initial guess (F-stat = 0.8222 and p-value = 0.4839). Finally, independence is addressed with the Durbin-Watson test, assuming the order on the data set is the collecting order. The test is not rejected (WD-stat = 2.345, p-value = 0.09), taking a significance level of 5%[3].

To sum up, the weighted means Two-way ANOVA gives evidence to conclude that there is a significant effect of education on salary received in this company. Furthermore, type I analysis does not indicate if there is significant differences between genders, which seems counterintuitive based on the descriptive findings. Using weighted means should allow to draw conclusions on the gender factor, but in this case it cannot be concluded if there is indeed gender discrimination. This occurs because the weighted approach implicitly is attributing some of the variance due to gender to the education factor. For this reason, an unweighted analysis could be more appropriate to conclude if there is salary gender discrimination.

5 Alternative Solution

In the previous section the data was analyzed using weighted means (Type I Sum of Squares) concluding that there is no significant difference in average `Salary` given `Gender` but only between levels of `Education`. However, given that the scientific question is to determine whether there is gender discrimination, unweighted means (Type III Sum of Squares) are used to estimate the difference between males and females salary, of course not taking into account the sample cell sizes.

As before and using the same notation mentioned, the hypothesis of interest is:

$$H_0 : \ \bar{\mu}_{F.} = \bar{\mu}_{M.} \tag{3}$$

But given that now unweighted means are used, this hypothesis can be rewritten as follows:

$$H_0 : \ \frac{1}{2}\left(\bar{\mu}_{FD} + \bar{\mu}_{FND}\right) = \frac{1}{2}\left(\bar{\mu}_{MD} + \bar{\mu}_{MND}\right) \tag{4}$$

It is easy to see that now the effect of `Education` factor is controlled in the analysis and so it is expected that the results are consistent with what has been found in the descriptive section. Table 4 shows the main results.

	Df	Sum Sq	F-value	Pr(>F)
Gender	1	225	26.32	1.05E-06
Education	1	1082	126.50	$< 2.2e-16$
Gender:Education	1	3	0.38	0.54

Table 4: ANOVA results (Type III SS)

Unlike for the weighted means analysis, `Gender` presents significant effect in the Salary. As before, the interaction between both factors is not significant and `Education` is significant. This goes inline with the interaction plot displayed in figure 1.

These results lead to important conclusions. First and the most important conclusion is that there is no gender discrimination for salary (see table 4). Second, men have higher salaries than women and the difference between both groups is significant. Finally, degree holders have on average higher salaries than people who do not have a degree and this difference is significant (see the interaction plot for the first intuition and table 5 for a confirmation). It can be counterintuitive that men are receiving more salary than women and that there is no gender discrimination. However, a plausible explanation can be that, men can have higher positions that imply more responsibility accompanied by a higher salary.

[2]In order to be completely sure about this, normality is also tested with Anderson-Darling and Lilliefors tests.

[3]It is important to mention that the p-value of this test is very close to the significance value. However, this can be explained by how the data is organized in the dataset, which in this case is clearly ordered by gender and education.

Once again, these conclusions are valid if the model assumptions are accomplished. This diagnosis its checked with the same tests for normality, homoscedasticity and independence from the last section. Also the diagnosis is showed in the following figures.

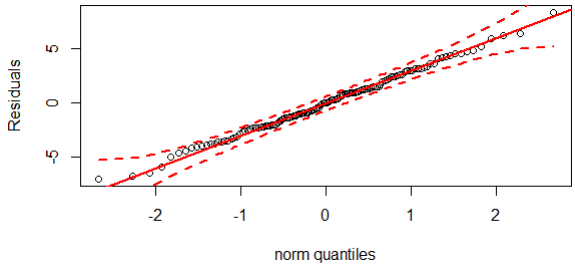

Figure 6: Normal Quantile plot

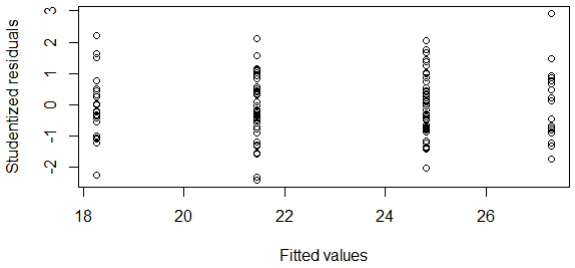

Figure 7: Residuals vs Fitted Values

Now given that the factors present significant differences it is important to test how these levels are affecting salary. Thus, because each factor has only two levels, the following hypothesis are tested using Tukey contrasts.

$$H_0 : NoDegree - Degree = 0 \tag{5}$$

$$H_0 : Male - Female = 0 \tag{6}$$

The results are displayed in table 5. It can be seen that both tests show significant differences regarding salary average. In fact, the second and third conclusion given above are true simultaneously, because the probability of making a type I error is controlled by the Tukey multiple comparisons procedure.

Hypothesis	Tukey	
	Estimate	CI
No degree - Degree == 0	-6.20	[-7.29,-5.11]
Male - Female == 0	2.83	[1.74,3.92]

Table 5: Education and Gender Contrast

To sum up, the unweighted means Two-way ANOVA gives evidence to say that there is no salary gender discrimination given the levels of education (no interaction). However, there are significant differences between gender and education levels. In fact, it can be concluded that a man with a degree potentially receives more salary than any other combinations of gender and education level. These results are consistent with the descriptive section. Note again that the assumption of sample sizes reflecting the population sizes per level had to be discarded for this interpretation. It is important to mention that even though the results suggest that men are receiving higher salaries than women, gender discrimination cannot be concluded. So if the researchers want to be completely sure about the results, more variables should be taken into account such as position, years of experience or labor hours.

6 Conclusion

A group of researchers were interested in whether or not there is salary gender discrimination at a particular firm. A random sample of 132 individuals is selected and salary, gender and education level is collected. In the first place, it is assumed that the sample size of each treatment reflects the importance of the treatment means, so Type I sum of squares of a two-way ANOVA is analyzed. In contradiction with the descriptive section, the results suggest that there is not significant difference between gender. Using weighted means leads to conclude who received more salary in term of gender but it cannot be concluded if there is indeed gender discrimination, considering education.

As an alternative solution, it is considered to analyze Type III sum of squares of the same two-way ANOVA. Instead of confounding the effect of the education factor, Type III allows to control for it. Defining

salary gender discrimination as the difference between males and females' salary under the same conditions, which in this case is under the same education level, it can be concluded that there is no gender discrimination because the interaction between factors is not significant. There is though a significant difference between males and females salary, but this can be explained by men in higher positions that imply higher salaries. For more conclusive results, more variables, like position and years of experience, should be taken into account.

References

Bobbitt-Zeher, D. (2007). The gender income gap and the role of education. *Sociology of Education*, *80*(1).

R Code

```
#####################
### Set Directory ###
#####################
setwd("...")

############################
### Load/Install packages ###
############################
packages <- c("ggplot2", "gplots", "dplyr", "multcomp",
              "car", "MASS", "FSA", "beeswarm","Hmisc",
              "sandwich","nortest")

package.check <- lapply(packages, FUN = function(x) {
  if (!require(x, character.only = T)) install.packages(x)
  if (! (x %in% (.packages() )))  library(x, character.only = T)
})

#################
### Read Data ###
#################
data =  read.table("salary.txt",header=T,sep=",")
str(data)
summary(data)


#################################################
### Initial Graphical exploration of the data ###
#################################################
#   Boxplot + beeswarm [per Education]
# inspired from
↪  https://www.r-statistics.com/2011/03/beeswarm-boxplot-and-plotting-it-with-r/
boxplot(data$Salary~data$Education, names = c("Degree","No Degree"),
↪  col=rgb(238,232,170, maxColorValue = 255))
beeswarm(data$Salary~data$Education, add = T,
         method = 'swarm',
         pch = 16, pwcol = as.numeric(data$Gender)*2,
         xlab = '', ylab = 'Salary'
)
legend('topright', legend = levels(data$Gender), pch = 16, col = c(2,4))

#   Boxplot + beeswarm [per Gender]
# inspired from
↪  https://www.r-statistics.com/2011/03/beeswarm-boxplot-and-plotting-it-with-r/
boxplot(data$Salary~data$Gender, names = c("Female","Male"),
        col=rgb(238,232,170, maxColorValue = 255),
        pars = list(boxwex = 0.8, staplewex = 0.5, outwex = 0.5))
beeswarm(data$Salary~data$Gender, add = T,
         method = 'swarm',
         pch = 16, pwcol = as.numeric(data$Education)*2,
         xlab = '', ylab = 'Salary'
```

```
)
legend('topright', legend = levels(data$Education), pch = 16, col = c(2,4))

#   Confidence interval plot [by Education]
plotmeans(data$Salary~data$Education, n.label = F, xlab="", ylab="Salary", main="Mean
↪  Plot with 95% CI", connect = F)
abline(mean(data$Salary),0, lty = 2)

#   Confidence interval plot [by Gender]
plotmeans(data$Salary~data$Gender, n.label = F,xlab="", ylab="Salary", main="Mean
↪  Plot with 95% CI", connect = F)
abline(mean(data$Salary),0, lty = 2)


##############################
### Descriptive Statistics ###
##############################

# mean summary tables and plots
tab = table(data$Gender, data$Education)
mean = tapply(data$Salary, list(data$Gender, data$Education), mean)
sd = tapply(data$Salary, list(data$Gender, data$Education), sd)
cbind(tab,mean,sd)

# Interaction Plots
interaction.plot(data$Education,data$Gender,data$Salary,xlab="Education", ylab =
↪  "Salary",
                 col = c(1:2), pch = c(18, 24),type="b",trace.label = "Gender")

################################
### Unbalanced Model Type I ###
################################
# Weighted Means = Type I SS
contrasts(data$Gender) = contr.treatment
contrasts(data$Education) = contr.treatment

# Hypothesis testing of gender
data.fit = aov(lm(Salary~Gender*Education,data = data))
data.fit.sum = summary(data.fit)
data.fit.sum

# Diagnostics
#  Homoscedasticity
#   Plot residuals vs fitted values
plot(fitted.values(data.fit),rstandard(data.fit),xlab="Fitted values",
     ylab="Studentized residuals",main="Residuals vs fitted values plot")
abline(h=0,lty="dashed")

#   Brown-Forsythe test
leveneTest(Salary~Education*Gender, data=data) ##The variance is constant

#   Independence of error terms
plot(rstandard(data.fit)[-c(1)],rstandard(data.fit)[-c(132)],
```

```
     xlab="Studentized residuals at 1 lag",
     ylab="Studentized residuals",
     main="Sequence plot")
abline(a=0,b=1,lty="dashed")

#   Durbin Watson test
durbinWatsonTest(data.fit, alternative="two.sided", data=data)

#   Normallity of error terms
#   QQ plot
qqnorm(residuals(data.fit), main="")
qqline(residuals(data.fit)) #Seems normal
qqPlot(residuals(data.fit), grid=F, ylab="Residuals")

#   Normallity Test
shapiro.test(residuals(data.fit)) # Not reject normality
ks.test(residuals(data.fit),"pnorm",alternative="two.sided") #Reject normality
ad.test(residuals(data.fit))
lillie.test(residuals(data.fit))

#   Outliers
n_T = dim(data)[1]
r = 4
pvalue_outliers = NULL
for(i in 1:n_T)
  pvalue_outliers[i]=1-pt(abs(rstudent(data.fit)[i]),
                         + n_T-r-1)
pvalue_outliers[pvalue_outliers>(0.05/(n_T))]=1
Stud.Deleted.Res=rstudent(data.fit)
Outlier.p.value=pvalue_outliers
out.data<-data.frame(Stud.Deleted.Res,Outlier.p.value)
out.data
table(out.data$Outlier.p.value)


##################################
### Unbalanced Model Type III ###
##################################
data.fit.type3= lm(Salary~Gender*Education,
                     contrasts=list(Gender='contr.sum', Education='contr.sum'),
                     data = data)
summary(data.fit.type3)
Anova(data.fit.type3, type='III')

# Model assumptions for second model
#  Homoscedasticity
# plot residuals vs fitted values
plot(fitted.values(data.fit.type3),rstandard(data.fit.type3),xlab="Fitted values",
     ylab="Studentized residuals",main="Residuals vs fitted values plot")
abline(h=0,lty="dashed")
#  Brown-Forsythe test (no normality needed and no equal sample size)
leveneTest(Salary~Gender*Education, data=data)
```

```r
#  Independence of error terms
# Durbin Watson test
durbinWatsonTest(data.fit.type3, alternative="two.sided", data=salary)

#  Normality of error terms
# QQ plot
qqPlot(residuals(data.fit.type3), grid=F, ylab="Residuals")

shapiro.test(residuals(data.fit.type3)) # Not reject normality
ks.test(residuals(data.fit.type3),"pnorm",alternative="two.sided") #Reject normality
ad.test(residuals(data.fit.type3))
lillie.test(residuals(data.fit.type3))

#  Outliers
pvalue_outliers = NULL
for(i in 1:132)
  pvalue_outliers[i]=1-pt(abs(rstudent(data.fit.type3)[i]),
                          + 132-4-1)
pvalue_outliers[pvalue_outliers>(0.05/(132))]=1
Stud.Deleted.Res=rstudent(data.fit.type3)
Outlier.p.value=pvalue_outliers
out.data<-data.frame(Stud.Deleted.Res,Outlier.p.value)
table(out.data$Outlier.p.value)

# Tukey contrasts for Gender and Education
sg_type3_Gender= glht(data.fit.type3, linfct = mcp(Gender= "Tukey"))
summary(sg_type3_Gender)
confint(sg_type3_Gender)

sg_type3_Education = glht(data.fit.type3, linfct = mcp(Education= "Tukey"))
summary(sg_type3_Education)
confint(sg_type3_Education)
```