Data Analysis Report
Survey on US workers

Ambrosi Andrea
Bonaldi Helena
Munari Chiara
Pakler Marco
Papa Bruno

February 15, 2019

Abstract

In this data analysis report a subset of a US survey on workers is analysed, in order to study the relation between the actual wage of people and some socio-demographic characteristics, such as the years of education, the marital status and several others. Moreover, the presence of discrimination patterns is investigated through the use of statistical methods, such as hypothesis testing applied to the linear models built over the dataset. In conclusion some behaviours outside the assumptions of the standard linear model will be noticed.

1 Introduction

The aim of this data analysis report is to study statistically relevant patterns in a subset of a US survey on workers. In particular, the goal of this analysis is to investigate what are the socio-demographic factors that influence the workers' wage, what is the magnitude of their influence, and how they behave and interact with other factors. Three main questions will be assessed:

A whether above-average looking women earn more than average looking women;

B whether the effect of physical appearance on the wage is the same for women and men;

C whether the education exerts the same effect on the wage of both black and white workers.

Then other interesting relations will be investigated, such as those arising from the extension of the city, the type of work and the membership in a union.

1.1 Dataset and descriptive statistics

In this chapter we are going to introduce the 12 variables analysed in the wages dataset (composed by 1260 observations):

Wage: hourly wage. The most common hourly wage is 3$ per hour, and the mean value is 6$. A narrow peak in the distribution around this value can be noticed, with an exponential-like decrease in the number of people with a wage above that value (see fig 1).

Exper: years of workforce experience. It can be seen that the distribution of the experience variable is positively asymmetrical, with the 50% of the sample falling under 15 years (see fig 2).

Looks: ranking made by an interviewer for physical attractiveness, using five categories (homely, quite plain, average, good looking, and strikingly beautiful or handsome) coded from 1 to 5, respectively.

Union: if union member (yes/no). In the sample 27% of people are member of a union, 73% are not.

Goodhlth: if good health (yes/no). 7% of respondents reported not to be in good health condition, while 93% in good health condition.

Educ: years of schooling. The variable takes 8 different values, ranging from 5 to 17, and the mode value is 12 years (corresponding to the standard U.S. educational program).

Ethnicity: if the person is black or white. The sample has 7% of black people and 93% of white people surveyed.

Gender: if the gender of the person is male or female. The percentages in the dataset are 65% male and 35% female.

Marital: if the person is married (69% of the sample) or single/divorced(31%).

Region: if a person lives in a northern or southern state. The observation for this variable are divided in 83% of people living in the North and 17% in the South.

City: if the person lives in a small, medium or big city. In the sample are collected responses of workers coming for 22% from a big city, 31% from a medium one, and 47% from a small one.

Industry: if the person works in the service(27% of the sample) or manufacturing industry(73%).

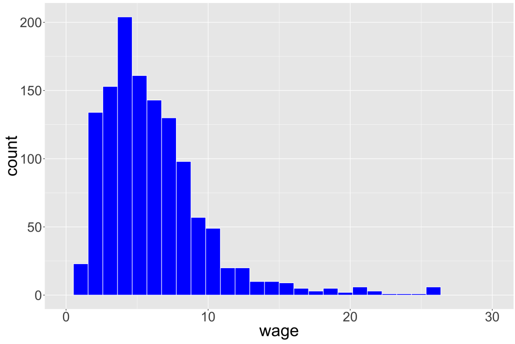

Figure 1: Distribution of wage by hourly amount x range limited to 30$.

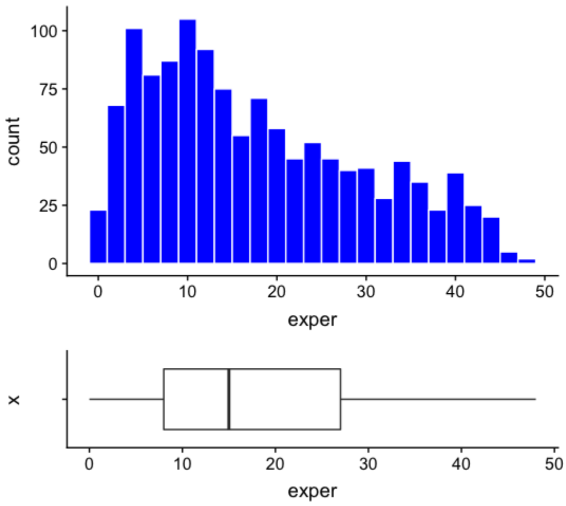

Figure 2: Distribution of people with a given number of years of experience with underneath its box-plot .

1.2 Data Exploration

Firstly here is reported the correlation matrix, which gives us low correlation among the numerical variables, see fig 16 for better visualization (in appendix)

	wage	educ	exper
wage	1.0000000	0.2123328	0.2346322
educ	0.2123328	1.0000000	-0.1861999
exper	0.2346322	-0.1861999	1.0000000

Figure 3: Correlation matrix for numerical variables in *wages* dataset

Then, various plots have been investigated in order to grasp the contents of the datasets, and to obtain some hints for performing the analysis procedures.

The distribution of wages has been seen again, but in a different perspective: in fig 4, in each bin of the histogram is also portrayed the proportion of women and men. Even though we limited the plot by the x axis, few points have been excluded and it can be seen that large part of the women occupies the bins in the lowest pay zones.

To better visualize this concept, in fig 7 we compared the number of people falling above or below the mean wage value. Even though the disproportion in terms of percentage of male and female represented in our sample is significant, we can observe that female workers are commonly less paid than male ones.

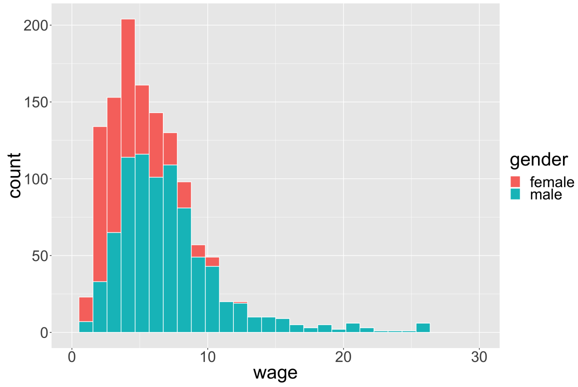

Figure 4: Distribution of wage
by hourly amount
the filling colour represents the proportion of men and women in each bin.

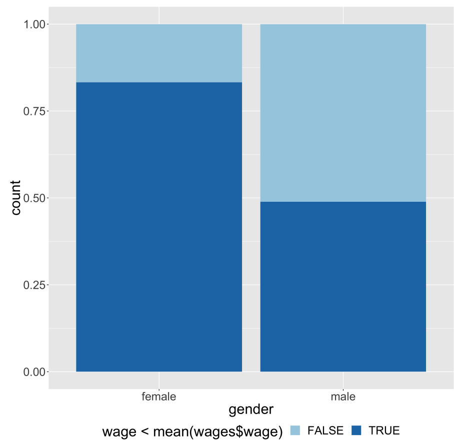

Figure 5: proportion of male and female workers falling above or below the mean wage threshold,
dark blue is lower than the mean wage value, light blue is higher.

In order to see how this phenomenon is different for the various levels of education of the respondents, we use a discretized variable in place of *educ*, reporting whether the respondent has an average level of education above or under that threshold fig 6. It can be noticed that in the highest education tier is reached the highest equality, which however is still not perfect.
As a final observation, the graph in figure 7 shows the experience represented according to the four quartiles of its distribution in the sample, and within every quartile

the distribution of people that has a wage above or under the mean.

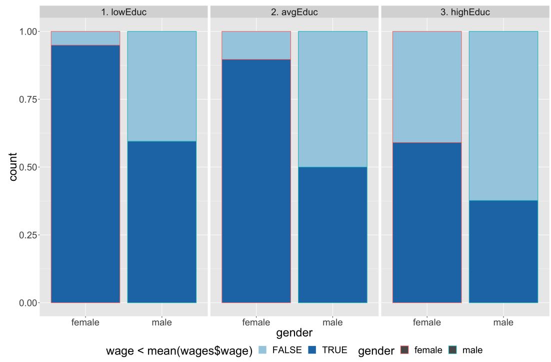

Figure 6: proportion of male and female workers falling above or below the mean wage threshold of $6.3,
dark blue is lower than the mean wage value, light blue is higher.

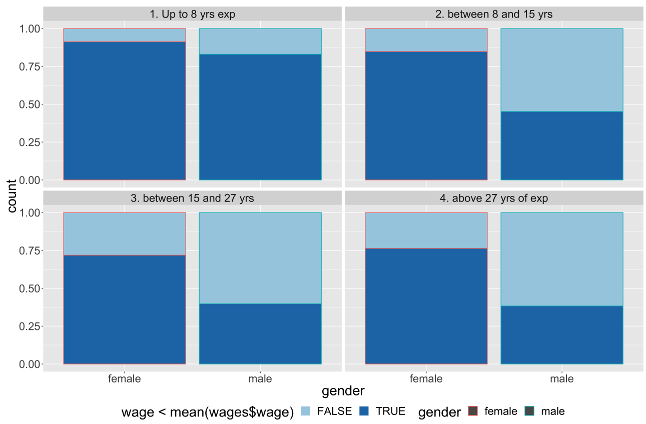

Figure 7: proportion of male and female workers falling above or below the mean wage threshold
dark blue is lower than the mean wage value, light blue is higher.

It can be observed a trend in line with the previous graph in terms of pay, but here the highest equality corresponds to the category with the lowest level of experience.

This discrepancy may coincide with a specific category of workers, young-adults with high level of education but few years of experiences. This can be interpreted as a new trend in the society, maybe predicting a future decreasing in the gender gap.

2 Methods and Analysis

2.1 Data Cleaning and dataset re-organization

reLook is a variable we built: it is a recategorization of the variable *looks*, performed to obtain less categories and thus reduce the noise in the analysis. We have reduced the variables from 5 to 3 possible values by unifying the variables in *underAvg* (composed by 1 & 2), *average* and *aboveAvg* (composed by 4 & 5)

looks	**reLook**
1	underAvg
2	
3	Average
4	aboveAvg
5	

This approach is justified by the observation that few people belong to the external categories, so by unifying them the result will not change much. In this way, eventual biases in the interviewer observation can be compensated. In fact, these may have emerged by the difficulty in differentiating such a complex characteristic as attractiveness into too many categories. In conclusion, in our analysis we will only take into account these three levels of attractiveness, and we will base our observations on them and on their overall effect on the wage.

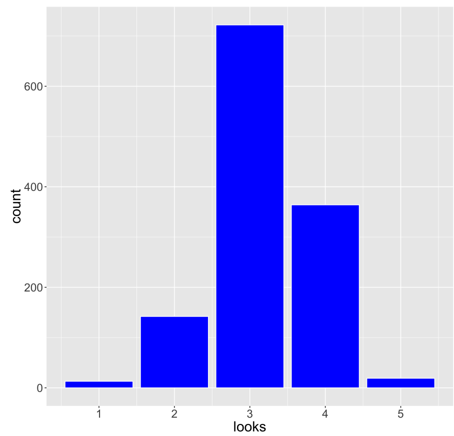

Figure 8: Amount of people belonging to various *looks* categories. As can be seen, a small number belongs to the external categories.

Since logarithms will be used in developing our model, 0 values in the numerical columns of the dataset would cause errors. To fix this problem, in the analysis and in developing the model it will be used a new data frame in which every 0 value will be substituted with 1.

2.2 Model building approach

Before starting the inferential analysis it is necessary to find the best model specification to study the current dataset.

The main factors we considered in this procedure are the following:
- The smallest model that fits the data is the best one (principle of Occam's Razor).
- Unnecessary predictors have to be dropped from the model, because they can be a source of noise to the estimation of the other quantities we are interested in.

- Parameters can be inserted in the model in a linear way or by applying different transformations (square root, logarithmic or exponential).
- The value of the *coefficient of determination* (R^2), that tell us about the goodness of fit of our model.
- The significance of the parameters of interest and the p-value, that indicates if the relationship found is significant.

Firstly, we created a model with all the variables specified in the dataset linearly included, as a baseline in order to compare every subsequent model specification. According to our intuition and some insights from literacy about the relationships among the considered parameters, we started including or excluding them one by one in the model, looking for improvements in the above mentioned factors.

By means of the anova function in R, it was tested whether a model was better than another one, which was different only for the addiction of an extra parameter. The anova function indeed, performs the F-test statistic and shows if the addiction of the parameter increases the percentage of the variance, explained or not by the model.

2.3 Current model description

After having performed the steps described in section 2.2 , it has been determined the subsequent model specification,

```r
lm(log(wage) ~ log(exper) + educ +
    union + region + city +
      educ:ethnicity + industry:city +
       reLook*gender, data = wagesL)
```

Here in figure 9 is reported the summary of the model.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.113395   0.096294   1.178   0.2392
log(exper)                    0.216964   0.016205  13.389  < 2e-16 ***
educ                          0.060990   0.007257   8.404  < 2e-16 ***
unionyes                      0.163282   0.029692   5.499 4.62e-08 ***
regionsouth                   0.060513   0.034568   1.751   0.0803 .
citymedium                   -0.188133   0.042866  -4.389 1.24e-05 ***
citysmall                    -0.091780   0.039615  -2.317   0.0207 *
reLookAverage                -0.064390   0.049034  -1.313   0.1894
reLookUnderAvg               -0.178152   0.071185  -2.503   0.0125 *
gendermale                    0.340910   0.049588   6.875 9.80e-12 ***
educ:ethnicitywhite           0.008094   0.004340   1.865   0.0624 .
citybig:industryservice       0.002184   0.064854   0.034   0.9731
citymedium:industryservice   -0.161504   0.055029  -2.935   0.0034 **
citysmall:industryservice    -0.185583   0.043148  -4.301 1.83e-05 ***
reLookAverage:gendermale      0.082730   0.060816   1.360   0.1740
reLookUnderAvg:gendermale     0.058798   0.090201   0.652   0.5146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.458 on 1244 degrees of freedom
Multiple R-squared:  0.4136,	Adjusted R-squared:  0.4065
F-statistic: 58.49 on 15 and 1244 DF,  p-value: < 2.2e-16
```

Figure 9: Summary of the model *modBig*

When looking at the model specification and its summary, the first thing that has to be noted is the transformation of the response variable as well as of an independent one. This transformation enables us to treat the variables as elasticities and also improves the smoothness of the model function.

The variable **exper** has proved to be quite significant as reported from literature, and it highlights that more year of experience have a positive return on the wage, in particular with the transformations performed we interpret the parameter of *exper* in percentages.

educ is here reported in the normal form, as we hypothesize that the years of education have a return on the wage amount that is exponential, in particular they have a positive effect.

union , **region** and **city** all prove to be really significant predictors, and have entered in the model for they are traits of recognised socio-economical relevance, an explanation of their effect is reported in section 2.5 .

educ:ethnicity is the interaction term between those two variables, and even though they did not proved to be significant as single variables, their interaction did, and this is also one of the aspect that we wanted to investigate in the first place, their effect is investigated in more detail in section 2.4.

industry:city is a relation that arose to be significant during the model building process and that brings some interesting clues about the organization of society and resource distribution, an explanation of their effect is reported in section 2.5 .

reLook*gender enter the model both as single variables and as their interaction. The aim is to determine if there is a different return on the wage between diverse looks of the people, in case they are female or male workers. It was also intriguing to determine the general effect of the gender of an individual(the gender gap issue) which will be described more in detail in section 2.5.

2.4 Investigation of possible discrimination patterns

With the given model we performed some analysis on the data, to answer the proposed questions.
To make inferences on them we used the output in 10.

A The relationship between being an above-average looking woman compared to an average looking woman is not significantly related to receiving an higher wage.

B For men, the effect on the wage of being above-average looking instead of average looking, even if statistically significant, is very small and thus not relevant.

Both these conclusions are supported by the confidence intervals analysis, which shows with a confidence of 95% that the relation is not statistically significant. See figure 11 (see fig 15 in appendix for better visualization).

In order to see those two effects, we used the summary from the original model and from another one obtained changing the reference level of the gender variable, this way all the information needed: the value of the parameter and the p-value were obtained only from the term of $reLook = Average$, as the reference level would be in one case Female-AboveAverageLooking and in the second is Male-AboveAverageLooking (see section Models in the R script).

C By investigating on whether education exerts the same effect on the wage of black and white workers, it emerges that the significance level is small ($p - value < 0.05$) and the effect is substantially significant (it's really small: 0.008)

```
t test of coefficients:

                              Estimate Std. Error t value  Pr(>|t|)
(Intercept)                  0.0780140  0.0917490  0.8503  0.395323
log(exper)                   0.2283761  0.0165338 13.8127 < 2.2e-16 ***
educ                         0.0613820  0.0071032  8.6415 < 2.2e-16 ***
unionyes                     0.1614871  0.0276211  5.8465 6.407e-09 ***
regionsouth                  0.0476509  0.0315203  1.5118  0.130851
citymedium                  -0.1764452  0.0391176 -4.5106 7.074e-06 ***
citysmall                   -0.0831880  0.0350565 -2.3730  0.017797 *
reLookAverage               -0.0584909  0.0521215 -1.1222  0.261992
reLookUnderAvg              -0.1651369  0.0678044 -2.4355  0.015011 *
gendermale                   0.3289701  0.0492368  6.6814 3.562e-11 ***
educ:ethnicitywhite          0.0082721  0.0048733  1.6974  0.089865 .
citybig:industryservice      0.0455514  0.0883914  0.5153  0.606409
citymedium:industryservice  -0.1666281  0.0545359 -3.0554  0.002296 **
citysmall:industryservice   -0.1922442  0.0437499 -4.3942 1.207e-05 ***
reLookAverage:gendermale     0.0748105  0.0619943  1.2067  0.227765
reLookUnderAvg:gendermale    0.0482542  0.0844023  0.5717  0.567618
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: Model parameters estimation after the application of the FWLS method

	2.5 %	97.5 %
(Intercept)	-0.1029150431	0.25894310
log(exper)	0.1988546789	0.25789761
educ	0.0475520780	0.07521194
unionyes	0.1090137677	0.21396040
regionsouth	-0.0145827380	0.10988455
citymedium	-0.2538094167	-0.09908097
citysmall	-0.1515162816	-0.01485976
reLookAverage	-0.1524195824	0.03543774
reLookUnderAvg	-0.2924834134	-0.03779036
gendermale	0.2376162521	0.42032387
educ:ethnicitywhite	0.0004173509	0.01612691
citybig:industryservice	-0.0821962729	0.17329901
citymedium:industryservice	-0.2801645283	-0.05309159
citysmall:industryservice	-0.2817083518	-0.10278010
reLookAverage:gendermale	-0.0413399862	0.19096098
reLookUnderAvg:gendermale	-0.1128449719	0.20935332

Figure 11: Confidence intervals at $\alpha = 0.05$ for model parameters

2.5 Further analysis on interesting patterns

Aside from the main purpose of this report, other interesting and significative patterns were found.

According to the expectations, the variable *gender* influences heavily the hourly wages, with a coefficient of 0.33 in favour of men and proved to be highly statistically significant.

The variable *city* is divided into three different categories: "Big", "Medium" and "Small". The model shows that there is a different impact on wages depending on whether a person works in a big city rather than in a medium one (-0.18 for a medium city with respect to a big one, with a high level of significance). From the model output, it cannot be inferred any statistically significant difference on the impact on wages between workers living in small and big cities.

Another interesting relation has been observed between the socio-demographic variables and the hourly wage, in the interaction between the city size and the type of work in which the respondents are involved. From the table in fig 10 it can be noticed that in bigger cities the kind of work does not really affect the wage of an individual, whereas in smaller cities the difference is more remarkable, and statistically significant as well.

To join a union affects the hourly wage of a worker: according to the output of the model there is a statistically significant relation (+0.16 if member).

2.6 Standard linear model assumptions and observed violations

To check the presence of multicollinearity, firstly it is observed that the value of R^2 is lower than 0.8 and there are enough significant indipendent variables too; every pair-wise linear correlation coefficient between two numerical regressors is low as well(see R script, Appendix A , section 4.1, line 217).

Another approach that has been used to exclude the presence of multicollinearity involved the Variance Inflation Factor (VIF), where every factor has indeed a value lower than 10 (see script in appendix A section 4.1 , line 217)

Since we are analysing cross-sectional data, violation of the exogeneity assumption is unlikely. This can be further confirmed by looking at the plot of the residuals, where no particular pattern can be noticed.

The normality assumption states that the errors are normally distributed, with zero mean and constant variance. According to the central limit theorem, a number of observations greater than thirty is sufficient to assess it. To further prove that, we look at the plot of the raw residuals and observe that they lie along a 45° degree line (see R script in appendix A , section 4.2, line 230), it can also be noted a quite heavy tail on the right side of fig 12. This is not really problematic but can inflate confidence interval and slightly modify inferential conclusion.

Furthermore, we have to check the presence of heteroskedasticity and autocorrelation of errors across observations. It can be observed the distribution of the residuals: no pattern can be noticed, so , in a preliminar way it can be excluded a problem of autocorrelation. However, the average distance of the residuals from zero is not constant, which is a sign of heteroskedasticity.

We then proceed with the formal Breusch-Pagan test: the p-value thus obtained is around $7 \cdot 10^{-4}$, so low enough to reject the null hypothesis about homoskedasticity (see R script in appendix A , section 4.3, line 237).

BP = 47.91, df = 21, p-value = 0.0007

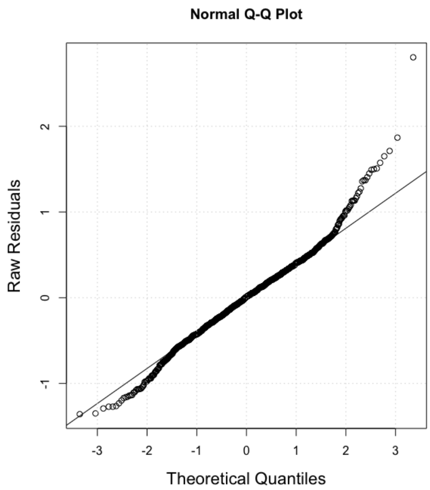

Figure 12: qqplot of Raw residuals vs theoretical quantiles , residuals lying on along the line confirm normality hypothesis

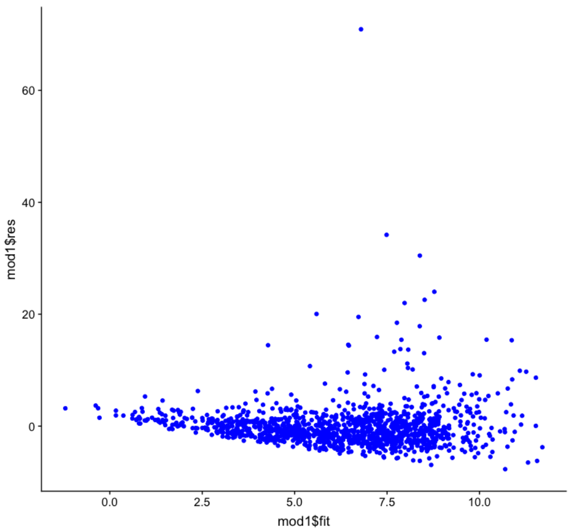

Figure 13: Residual plot of *modBig* , before logarithmically transforming the response variable, a larger spread in the distribution of points can be observed in the right region

2.6.1 Heteroskedasticity

The use of the logarithmic transformation in the model already reduced the problem of heteroskedasticity. However, non constant variance still seems a problem after performing the Breush-Pagan test at each step. A first approach in order to reduce the issue is the feasible weighted least squared (FWLS) estimator. We take the logarithms of the squared residuals, and regress them to the other variables. We then use the fitted values w as the weights that will be used in the correct model (modBigFWLS). With those errors, the estimator will no longer be unbiased but will provide a correct input for the inferential procedures, and will also be more asymptotically efficient than the ordinary β_{OLS} estimator (see R script in appendix A , section 4.4 , line 261)

As the heteroskedasticty speficiation used in the FWLS could probably be wrong it has been used the White's heteroskedasticity-robust standard errors as a final approach. With those errors, the estimator will no longer be unbiased but will provide asymptotically valid inferences about the model parameters. This is allowed because we have a large enough sample (see R script, section 4.5 , line 283)

2.7 Further improvements - Outliers

Outliers can be difficult to detect: some points can be just highly influential and far from the other observations, but still important.

To attempt an outlier analysis on the model described in section 2.3 we used the R function plot(<model>) , which among other useful plots enables the user to easily get the ID of the three most distant points from the fitted values.

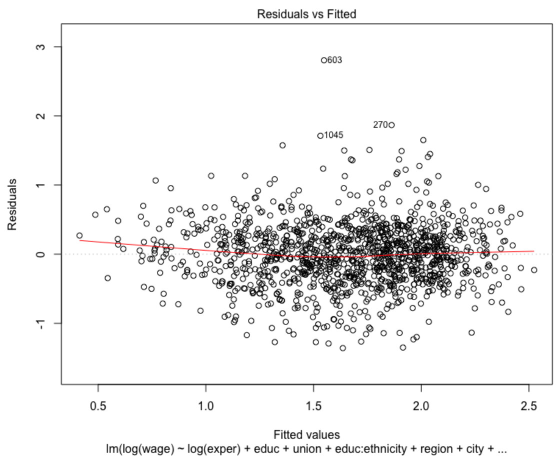

Figure 14: plot of residuals, with candidate outliers points highlighted

Moreover, it has been used an R function to find the most distant points from the fitted values (see R script in appendix A, section 3 . Outlier Test , line 196) , which confirmed the record 603 as an unusually distributed residual, giving a low Bonferroni's p-value , a hint of the point being mean-shifting.

The record number 603 has then been inspected: it belonged to a black woman, who scored 5 in the original *looks* category, with standard years of education (13), 9 years of experience and an hourly wage of 77$. By removing from the dataset that record, which was the only true possible outlier, a little improvement in the overall significance of the model and a higher R^2 value have been registered.

It has to be considered that we do not have enough evidence to conclude that there was any error in that record, so even though the value is far from our model prediction, it cannot be excluded from the analysis.

3 Conclusions

After the analysis performed, it has been developed a model with a satisfying reliability (see section 2.3)

built by taking into account the relevance of every variable and by trying to use the minimum number of explanatory variables as possible.

This model was then used to verify and determine the absence of discrimination in the return to education in terms of wage for black and white workers.

Moreover, there was no evidence of discrimination between *above average* and *average* looking workers (both female and male). even though a statistically relevant difference is registered in the overall salary between the two genders.

All these results were obtained accounting also for intrinsic problems in the dataset, as it did not satisfy some of the standard linear model assumptions (i.e. : homoskedasticity), this problem was partly solved by the logarithmic transformation of the numerical variables and by using the methods of FWLS and then the White's *heteroskedasticity-robust standard errors* for β_{OLS}.

Finally, other interesting trends were observed, such as the differences between the salary of men and women, the relation between the wages of workers and the type of job (service or industry), its location in a big or small city and the membership of the worker in an union.

3.1 Final thoughts and perspectives

It was rather surprising the absence of discrimination based on physical appearance or ethnicity. On the other hand, a significant discrimination based on the gender has been observed.

Another observable peculiarity is the disproportion of the respondents in the survey: for example female workers were just half of the male ones, and black workers were represented in a equally low proportion too. It may have been interesting to view our findings in the context in which this survey took place, to compare the percentage distribution in the sample with the one of the population. This could have been useful to better interpret the values obtained, and to understand if some unexpected values in the proportion of the sample are a normal feature in the population or a distortion due to bad sampling methods. Moreover, some other attributes could have been useful, such as the age of the respondents, or to take into account other ethnic groups such as the Latin American one, present in large number in the US but totally ignored in the sample. Finally, knowing the year when the survey took place, it may have been helpful to make other consideration accounting for some characteristics of the socio-economic reality during that year.

Appendices

A R Script

```
######################################
#####STATISTICAL LEARNING PROJECT#####
#####         Based on R 3.5.1   #####
######################################
# Date:    14/02/2019                #
# Authors: Ambrosi Andrea            #
#          Bonaldi Helena            #
#          Munari  Chiara            #
#          Pakler  Marco             #
#          Papa    Bruno             #
######################################

library(ggplot2)
library(faraway)
library(lmtest)
library(corrplot)
library(gpairs)
library(car)
library(coefplot)
library(gridExtra)
library(cowplot)

#Set the directory
setwd("/Users/bru_008/Documents/Corsi/Stat_Learning/ProgettoLinearModel_StatistikaLMaximo/
    versoLaFine")
#Import the wages file
wages < read.csv("wages.csv", header = T)

#########################################
#     1.Cleaning and modifying         #
#   dataset for further analysis       #
#########################################

#Re categorization of the column look
#(1,2)   > 1
#(4,5)   > 3
#For low accuracy in determining marginal values, reduce noise in analysis
wages$reLook[wages$looks == 1 | wages$looks == 2] < "UnderAvg"
wages$reLook[wages$looks == 3] < "Average"
wages$reLook[wages$looks == 4 | wages$looks == 5] < "AboveAvg"

#Build a new column with different interval of hourly wage
#from 1 (low) to 5 (very high)
binWidth = ( max(wages$wage)   min(wages$wage) )/ 5
wages$wageCat[ wages$wage < 2] < "1"
wages$wageCat[ 2 <= wages$wage & wages$wage< 4] < "2"
wages$wageCat[ 4 <= wages$wage & wages$wage< 6] < "3"
wages$wageCat[  wages$wage >=6] < "5"

#Create categories for education year under avg above
wages$educCat[wages$educ < 11] < "lowEduc"
wages$educCat[wages$educ < 13.5 & wages$educ> 11.5] < "avgEduc"
wages$educCat[wages$educ >13.5] < "highEduc"

#Create categories for exper novice   standard   high   veteran
wages$expCat[wages$exper <= 8] < "Up to 8 yrs exp"
wages$expCat[wages$exper <= 15 & wages$exper> 8] < "between 8 and 15 yrs"
wages$expCat[wages$exper <= 27 & wages$exper> 15] < "between 15 and 27 yrs"
wages$expCat[wages$exper >= 27] < "above 27 yrs of exp"

##############################
#pre 2 numerical exploration#
##############################

#information about the dataframe dimension and type of data in columns
```

```r
str(wages)

#correlation among numerical variables
wagesNumericals < wages[c("wage" , "educ" , "exper")]
cor(wagesNumericals) #that confirm low correlation |0.2| or lower.. it's ok
corrplot.mixed(cor(wagesNumericals), upper = "ellipse" ) #un modo pi?? carino di vederlo

#visually plot every num variable in respect of every other
gpairs(wagesNumericals)


#function to find the mode of a sample distribution
getmode < function(v) {
  uniqv < unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

getmode(wages$wage)
getmode(wages$educ)



#####################################
# 2.DATA EXPLORATION AND PLOTTING #
#####################################




#Histogram of the wage distribution
ggplot(data = wages) + geom_histogram(mapping = aes(x = wage , fill = reLook), color = "white")+ xlim(0,30) + theme( text = element_text(size = 30))
ggsave("wageDistroReLookLimX30.png")

#histogram plot for experience distribution
ggplot(data = wages) + geom_histogram(mapping = aes(x = educ) , fill = "blue", color = "white")

#experience years distribution: histogram plot
ggplot(data = wages) + geom_histogram(mapping = aes(x = exper) , binwidth= 5 , fill = "blue", color = "white")

#experience years distribution: histogram plot + density + boxplot
a < ggplot(wages, aes(x = exper)) +
  geom_histogram(binwidth = 2, color = "white", fill = "blue" )

b < ggplot(wages, aes(x = "", y = exper)) +
  geom_boxplot() +
  coord_flip()

plot_grid(a,b,nrow=2 ,align="v", rel_heights = c(2/3 , 1/3))


#Histogram counts for the different looks categories
ggplot(data = wages) +
  geom_bar(mapping = aes(x = reLook),fill = "blue", color = "white")+
  theme( text = element_text(size = 20))
ggsave("reLookCount.png")

#histogram plot, counts of respondents in three educational level
ggplot(data = wages) +
  geom_bar(mapping = aes(x = educCat),fill = "blue", color = "white")+
  theme( text = element_text(size = 20))


###Histogram for educ and exp recathegorized investigating the gender gap####

#male vs female vs condition of being above or below the mean wage threshold
ggplot (data = wages) + geom_bar(mapping = aes(x = gender , fill = wage < mean(wages$wage)), position = "fill")+
  theme( text = element_text(size = 24) , legend.position = "bottom") + scale_fill_brewer(palette = "Paired")
```

```r
#ggsave("genderVsmeanWageNorm.png")

#mean wage threshold plot for different educational level
ggplot(data = wages, aes =(col = gender), position = "dodge") +
  geom_bar(mapping = aes(x = educCat  ,fill = wage < mean(wages$wage)))+
  theme( text = element_text(size = 20))

#mean wage threshold plot for different educational level, focus on different gender
ggplot(wages , aes( x = gender)) + geom_bar(aes(fill = wage<mean(wages$wage), color = gender), stroke = 1 , position = "fill") +
  facet_wrap(~educCat) +   theme( text = element_text(size = 24) , legend.position="bottom")  + scale_fill_brewer(palette = "Paired")
ggsave("genderVsEducVsWageMean.png")

#mean wage threshold plot for different experience level, focus on different gender
ggplot(wages , aes( x = gender)) + geom_bar(aes(fill = wage<mean(wages$wage), color = gender), stroke = 1 , position = "fill") +
  facet_wrap(~expCat) +   theme( text = element_text(size = 24) , legend.position="bottom")  + scale_fill_brewer(palette = "Paired")



#Histogram plus density funciton, both normalized to see the distribution of the wage value among population
ggplot(data = wages , aes(x=wage)) + geom_density() + geom_histogram(aes(y = ..density..), alpha = 0.3)

#Histogram of the educ distribution
ggplot(data = wages) + geom_bar(mapping = aes(x = educ))

#Histogram of years of experience accounting for looks
ggplot(data = wages) + geom_histogram(mapping = aes(x = exper , fill = reLook), binwidth = 5)

#Histogram of years of experience fill colors prop to reLook variable normalized
ggplot(data = wages) +
  geom_histogram(mapping = aes(x = exper , fill = reLook), position = "fill", color = "white", binwidth = 5)+
  theme( text = element_text(size = 20))
#ggsave("experLookProp.png")


#Histogram that has on the x the reLook columns for both the genders.
#Each columns is filled by colors based on the wage category.
ggplot(wages , aes( x = reLook)) + geom_bar(aes(fill = wageCat), position = "fill" , color = "white") +
  facet_wrap(~gender) + scale_fill_brewer(palette = "Set1")

#Dotted graph for ethnicity and education
ggplot(data = wages) + geom_point( mapping = aes( x = ethnicity , y = educ, color = educ) , position = "jitter")

#Histogram with the education discretized for ethnicity
ggplot(wages) + geom_histogram(aes(x = educ), binwidth = 3, color = "white", fill = "blue") + facet_wrap(~ethnicity)

#############
# 3. MODELS #
#############

#Clone the dataset replacing the 0 with 1 in the column
#of the experience in order to perform the analysis with the log function.
wagesL <  wages
wagesL$exper[wagesL$exper == 0] <  1

# Best model #
modBig <  lm( log(wage) ~ log(exper) + educ + union + educ:ethnicity
              + region + city  + industry:city + reLook*gender, data = wagesL)

# same model but with releveled gender variable
#modBig <  lm( log(wage) ~ log(exper) + educ + union + educ:ethnicity
#              + region + city  + industry:city + reLook*relevel(gender , "male")
```

```
#                , data = wagesL)

#We use relevel in order to change the reference level for the relative variable.
#It is used to see the difference in reLook between a nice and a wonderful women.

#With the summary function we can see and verify the patterns of possible discrimination
summary(modBig)


####################################
#       Outliers Test              #
####################################
#find and test for outlier status with Bonferroni critical value from library car
outlierTest(modBig)

outliersDF < wagesL[c(603),]
wagesNoOut < wagesL[ c(603),]

modBigNoOut < lm( log(wage) ~ log(exper) + educ + union + educ:ethnicity
                + region + city  + industry:city + reLook*gender, data = wagesNoOut)
summary(modBigNoOut)




############################################
# 4.VIOLATIONS OF THE MODEL ASSUMPTIONS #
############################################

#We assume the normal distribution according to the large numbers law

# 4.1 FULL RANK / COLLINEARITY / MULTICOLLINEARITY #

round(cor(wagesL[, c(1,2,6)]), digits = 3)
#As we can see the correlation between the numeric variables is low.

#Now let's control the VIF (variance inflator factor)
#If it is > 10 we'll have a problem
vif(wages[, c(1,2,6)])
#But the result is less than 10 so there is no high multicollinearity.



aux_reg1 < lm(log(exper)~ + educ + union + educ:ethnicity
                + region + city  + industry:city + reLook*gender, data = wagesL)

aux_reg2 < lm( educ ~  log(exper) + union + educ:ethnicity
                + region + city  + industry:city + reLook*gender, data = wagesL)

summary(aux_reg1)
summary(aux_reg2)

# 4.2 NORMALITY DISTRIBUTION #

qqnorm(modBig$res, ylab = 'Raw Residuals')
qqline(modBig$res)
grid()


# 4.3 HOMOSCEDASTICITY AND NONAUTOCORRELATION #

#Informal method:
#Now we plot the residuals over the fitted values
#in order to visually detect if there are patterns
plot(modBig$fit, modBig$res, xlab = "Fitted", ylab = "Residuals")
grid()
#and in the plot we can see a pattern on the bottom

#Formal method:
#Let's do the BP test using the F test
auxmodBig < lm(modBig$res^2  ~ log(exper) + educ + union + educ:ethnicity
```

```r
                  + region + city  + industry:city + reLook*gender, data = wagesL)
summary(auxmodBig)
#Our p value is: 0.00438

#Another way to perform the BP test is with the lmtest library:
bptest(modBig)
#Result
#BP = 33.029, df = 15, p value = 0.004651
#Given H0: homoscedasticity
#As result we have that the p value is small enough so we can reject the null hypothesis.
#It seems there is the heteroscedasticity problem but we can try to correct it with the FWLS.

# 4.4 FWLS AGAINST HETEROSCEDASTICITY #

#Here we take the log of the square of the residuals
logRes2 < log(modBig$res^2)

#And here we fit the log of the square of the resuduals to the rest of the variables
varMod < lm( logRes2 ~ log(exper) + educ  + union + educ:ethnicity
                + region + city  + industry:city + reLook*gender, data = wagesL)

#Then we take the fitted values that we are going to use as weights for the corrected model
w < exp(varMod$fit)

#This is the corrected model, like the original one but weighted using the w
modBigFWLS < lm( log(wage) ~ log(exper) + educ  + union + educ:ethnicity
                     + region + city  + industry:city + reLook*gender, weight = 1/w , data = wagesL)
summary(modBigFWLS)

confint(modBigFWLS)

coefplot(modBigFWLS, intercept=FALSE, outerCI=1.96,
          xlab="Association with hourly wage")

# 4.5 HETEROScEDASTICCTY ROBUST STANDARD ERRORS

coeftest(modBigFWLS , vcoc = hccm)

######## EOF #########
```

SLScript.r

B Supplementary plots

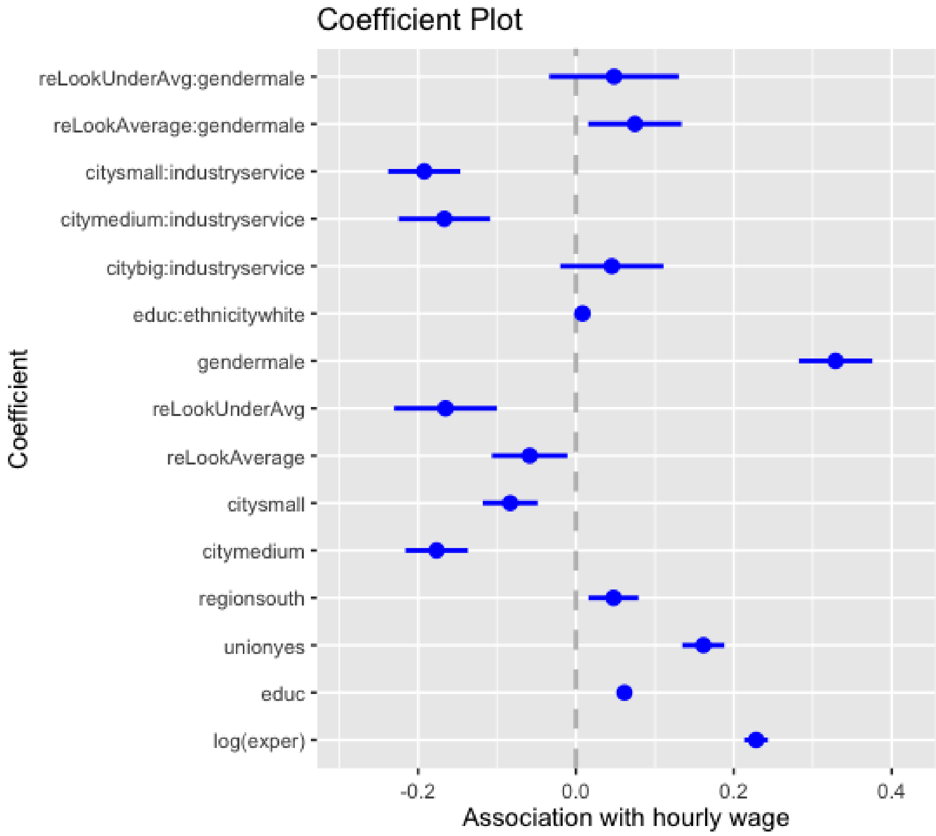

Figure 15: Output of the function *coefplot*, con α significance level of 0.05

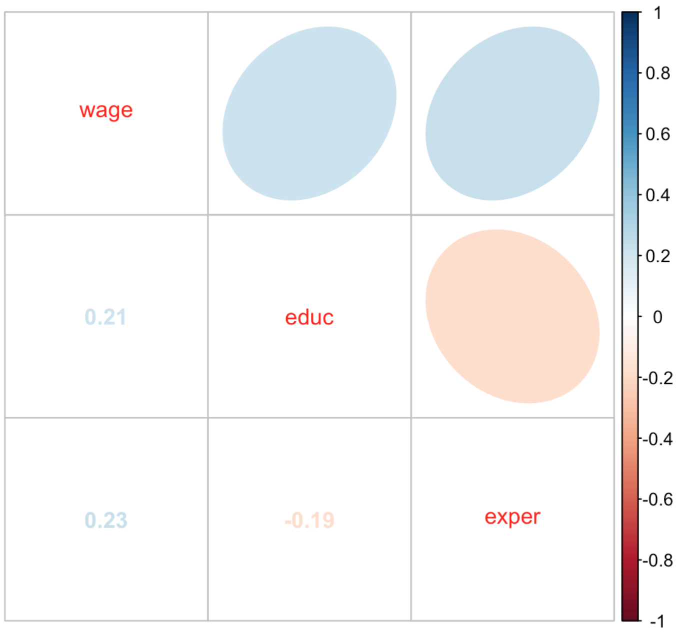

Figure 16: Graphical display of correlation among numerical variables. The colour of the ellipses is related to the respective correlation coefficients