Statistical Learning Project

1. Introduction

In this paper we will analyse data from the dataset wages. It is a subset from a U.S. survey on the labour-market which accounts not only for socio-demographic aspects but also for personal characteristics like physical appearance and health of the interviewed. There is a vast literature in social sciences on the influence of education and experience over wages. Starting from this dataset we can explore more deeply the presence of a link between wages and other attributes thanks to the structure of our dataset. Since we are interested in relationships between variables and interpretation of the results in our case is a relevant part, a linear regression will be conducted on our data. Moreover, being this an exercise to model a linear regression on "real-life" data, we will tackle problems related to the violation of the assumptions of a linear regression model based on the Ordinary Least Squares approach. We will, then, develop several models evaluating the performance of each of them. To conclude, we will show which is the most statistically sound model and, starting from it, we will draw the conclusion of our analysis.

2. Violations detection and data manipulation

After having fit a simple linear model on the whole dataset and having explored the results it is immediately clear that adjustments are necessary to improve the performance of the linear model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.00717039	1.04275323	-1.9248757	5.447217e-02
exper	0.07897499	0.01067212	7.4001229	2.492787e-13
looks	0.41424067	0.17425827	2.3771650	1.759642e-02
unionyes	0.61117361	0.26702630	2.2888143	2.225673e-02
goodhlthyes	-0.05165029	0.47552715	-0.1086169	9.135238e-01
educ	0.42472330	0.05009991	8.4775270	6.412430e-17
ethnicitywhite	0.11015402	0.46128174	0.2387999	8.113000e-01
gendermale	2.12806702	0.27636788	7.7001242	2.754602e-14
maritalsingle/divorced	-0.82169852	0.27441057	-2.9944128	2.803945e-03
regionsouth	0.35674662	0.31164836	1.1447088	2.525495e-01
citymedium	-1.71995666	0.33633605	-5.1138040	3.651630e-07
citysmall	-1.13294855	0.30810863	-3.6771075	2.459280e-04
industryservice	-0.47885515	0.28821162	-1.6614706	9.687032e-02

Figure 1: coefficients of the basic linear model provided by the summary function.

Firstly, in *Figure 1*, we can see that three variables are not satisfactory in terms of linear correlation. In fact, being caucasian, being healthy and coming from south seem to be not so helpful in describing our dataset with a linear model. We can suggest that those variables are not helpful by looking at the probability of getting values as extreme as the ones we are getting in a random distribution where the variable of interest has no effect (p-value). In other words, the p-value is the probability of observing a value of the t-statistic different from the null hypothesis, that is H0 -> t-statistic equal to 0. As result, a small p-value, generally lower than 0.05, means low probability of having the coefficient of a specific variable equal to 0 and therefore a high probability of an influence on our

model of the corresponding variable. The interpretation of this results is that, in the United States, a common aspect of discrimination present in the literature such as ethnicity seems not to be present. However, this, as we said earlier, is a basic model, built without considering the structure of our data. This means that we do not know yet if those data satisfy the assumptions of linear regression. Additionally, the R^2, that is the proportion of the explained variance of the outcome variable by the predictors, is very low (0.21), meaning that the model is not describing data as it should. To understand the reasons of this poor performance, an effective method is to try exploratory data analysis. R provides four pre-built plots for linear model objects to explore the results of the models.

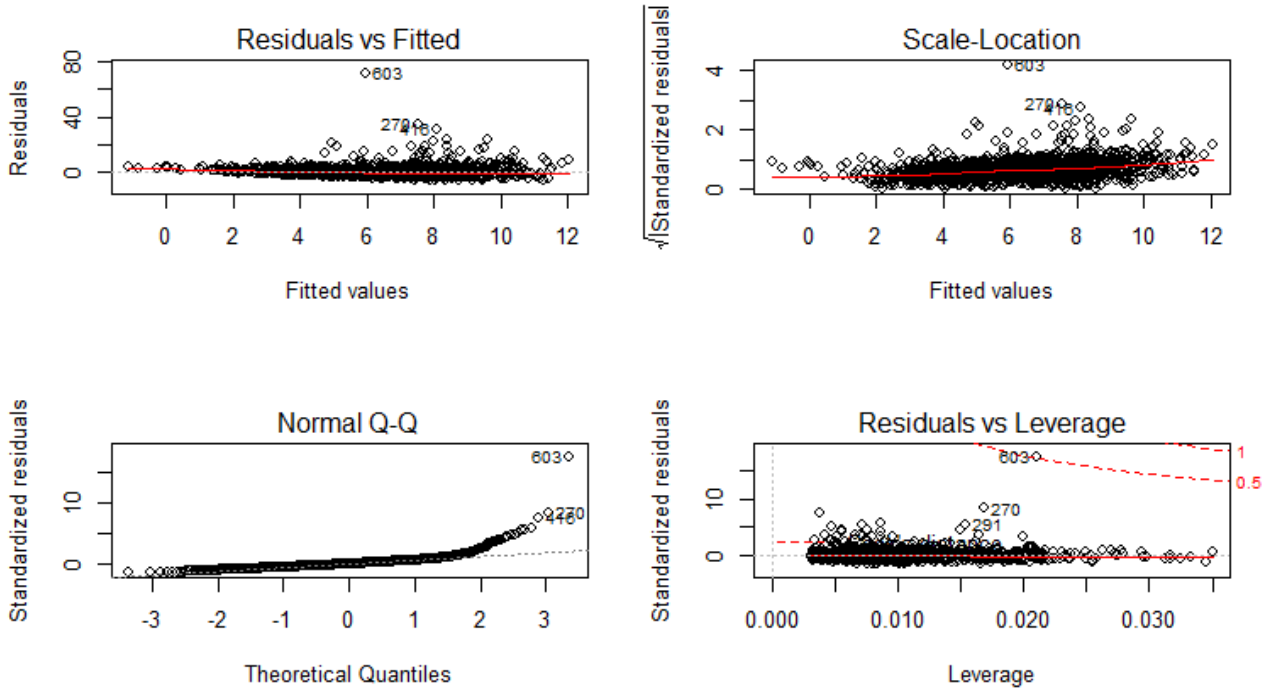

Figure 2: plots of the basic linear model

The plot in top-left of *Figure 2* is called Residual vs Fitted and it gives hints about the assumptions of a linear model. It shows the distribution of the errors for every (predicted) value. In our case we can see that errors instead of being sparse tend to diverge from the error mean (horizontal red line) as we move on higher values of the x axis, denoting no equal variance of the errors. Indeed, this is an evidence against one assumption of the Ordinary Least Squares regression, that is Homoscedasticity. Homoscedasticity relates to the error terms. In the ideal case, where we can assess that our estimators are the most efficient ones according to Gauss-Markov theorem, those errors must present the same finite variance and expectation equal to 0 independently from the value of the regressors. It can be seen that in our case lots of observation present higher errors than most observations. This last problem is even more evident in the scale location plot: here instead of plotting Residuals themselves, are plotted the root of the absolute values of the standardized errors. In fact, errors increase moving right on the x axis showing a “megaphone” shape. Another assumption of linear regression is that errors must follow a Normal distribution. The Normal Q-Q plot deals with this specific assumption: on the x axis are plotted the points as they belong to a normal distribution and in the y axis are plotted the residuals of our model. Errors are normally

distributed if they lay on the dashed line. We can see that in the right tale, the distribution of our errors differs from the theoretical one, indicating that high wages are not well described by the predictors of our model. Linear models are heavily biased by outliers since they affect the calculation of the variance-covariance matrix. The bottom-right plot of the *Figure 2* shows influential points of our dataset according to Cook's distance. In our case there is only one influential point with Cook's distance greater than 0.5 that lays outside the dashed line. Other points are far from the main cloud of points, such as observations 270 and 291. As result these points will be discussed later. Once again, to be confident in assessing the performance of our linear model we must check if errors (or residuals) are not correlated to the variables used as predictors. However, we are comfortable in saying that this assumption is respected because, performing a correlation test against residuals, variables experience and education showed a correlation near zero with p-value equal to 1 in both circumstances.

Now that we have accounted for the violations of our linear model, we can try to manipulate the structure of our dataset. Hints were given to us to guide the manipulation. For example, variables education and experience may not exert a linear relationship with the output variable wage. In addition, another action suggested is to re-encode the variable look. As first step we will analyse the outliers.

	wage	exper	looks	union	goodhlth	educ	ethnicity	gender	marital	region	city	industry
270	41.67	16	4	no	no	13	white	male	married	north	small	manufacturing
416	38.86	29	3	no	yes	13	white	male	married	north	small	manufacturing
603	77.72	9	4	yes	yes	13	black	female	married	north	big	service

Figure 3: outliers of the dataset wages

The wage mean in the dataset is 6.3 dollars per hour. In these observations the wage per hour is, at least, of 38$. Particularly interesting is the number 603. This case is a black woman. Her wage per hour is 77.72 $, that is more than 10 times higher than the average of people in the dataset. This could be not surprising itself but if we look at her experience and education it is quite surprising. In fact, people with the same characteristics of education and experience gain an average of 11.3 $ per hour. The other two cases (270,416) are relatively common. It is not unusual that some people gain more money than the rest of people. And it is even more common that only few people perceive a very high wage if compared to the mode of wages. Additionally, considering that the highest wage come from a female, it could hide differences in terms of wage. Since there are strong evidence in assessing that the case number 603 is an outlier, and there are weaker evidences for the other two observations to be outliers, we decided to remove from the dataset only the first one. The model resulting from this new dataset is more powerful. In fact, now, it describes 27% of the total “variance” of our dataset. In addition, variables that were statistically not significant are now closer to be significant. By removing the other two outliers the model performs even better in terms of R squared and regressors significance, but we do not have strong evidence to be confident in removing them.

Had the first problem been tackled, let us exert the nature of the relationship between the outcome and the continuous regressors education and experience. To verify if the relationship is non-linear, we ran the white neural network test. White neural network test assumes as null hypothesis

linearity between variables so, if the p-value is lower than 0.05 we are 95% confident in assessing that the variables tested are not linearly correlated. In both two cases, i.e. experience over wage and education over wage, p-value is lower than 0.05, which means non-linear relationship between predictors and depended variable. What we can do now is to find the best transformation for each continuous regressor. R provides a function called Box-Tidewell within the package car. This specific function tries iteratively, with the maximum likelihood approach, the best power transformation for each specified regressor, that are, in our case, education and experience. Considering our dataset, the best transformation for experience and education are respectively logarithmic transformation and 2.75 power transformation with a confidence interval over 95%. The model resulting from these transformations however is almost unchanged. Since we have already fixed problems related to outliers and non-linearity, the next step of our pipeline is to tackle heteroscedasticity highlighted in the diagnostic plots. We executed several tests for homoscedasticity (Breusch-Pagan, Score test, F test, Bartlett test) and the response of each of them was a statistical evidence against homoscedasticity. As consequence, known that our outcome variable is strictly positive, we performed another algorithm based on maximum likelihood looking for the best possible transformation for the variable wage. The algorithm is called Box-Cox transformation and the best possible transformation resulted once again to be a logarithmic transformation. The model improved its performance sensibly.

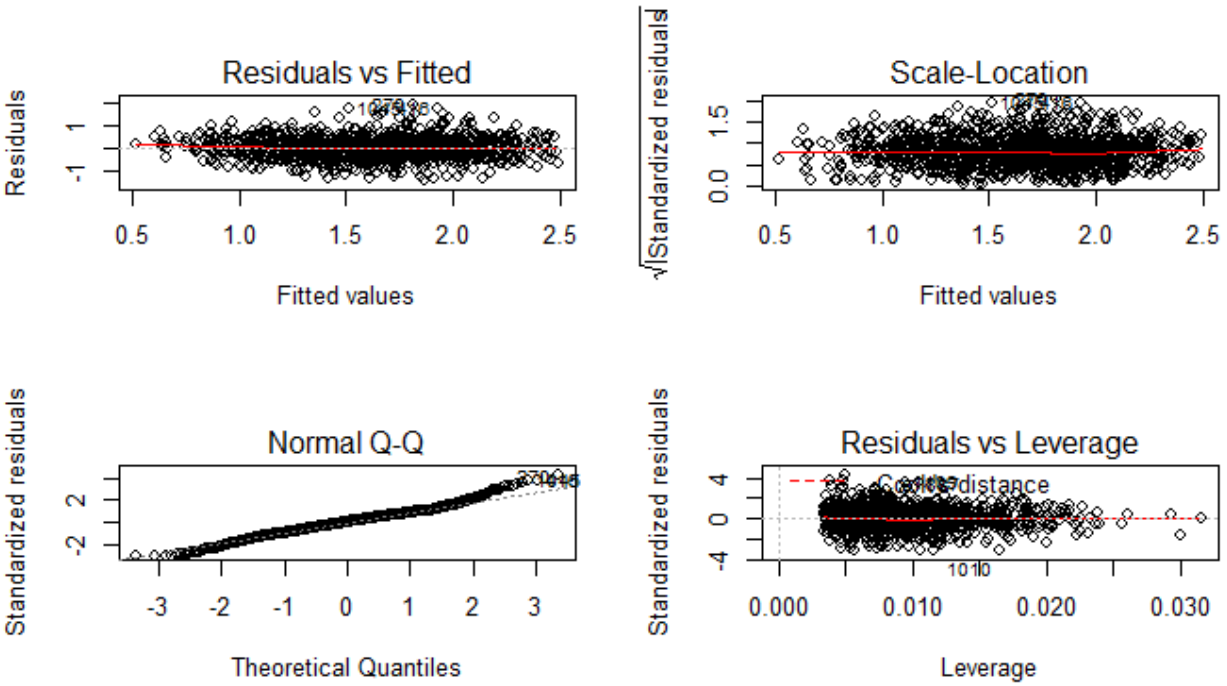

Figure 4: plots of the final model

By looking at the plots of the final model in Figure 4 we can see were the model has improved. In fact, the scale location plot shows a sparser scatter of the residuals against the fitted values, hence a lower presence of heteroscedasticity. In addition, influential points are now reduced or even not present since almost every point is under the red dotted line of Cook's distance. Finally, the normal Q-Q Plot points out that residuals are almost normally distributed amongst observations. Finally,

below are reported the coefficients of the final model. Since we are interested in discriminatory patterns, we changed the baselines of our model to easily have access to them.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2679053905	6.831901e-02	18.5586040	4.411816e-68
log(exper)	0.1984812839	1.652560e-02	12.0105361	1.592689e-31
educ275	0.0002571863	2.187273e-05	11.7583110	2.396875e-30
looksGood Looking	0.0034296355	2.938349e-02	0.1167198	9.071009e-01
looksHomely	-0.1452897607	4.056481e-02	-3.5816698	3.545446e-04
unionno	-0.1549387720	2.942597e-02	-5.2653743	1.646209e-07
genderfemale	-0.4270801210	2.907403e-02	-14.6894013	3.559693e-45
aritalsingle/divorced	-0.0271843328	3.071729e-02	-0.8849848	3.763354e-01
citymedium	-0.2466646641	3.677150e-02	-6.7080386	2.984170e-11
citysmall	-0.1445211217	3.381958e-02	-4.2732979	2.072590e-05
ethnicityblack	-0.1391752054	5.078302e-02	-2.7405855	6.220540e-03
regionsouth	0.0617887452	3.425542e-02	1.8037654	7.150935e-02

Figure 5: estimate of the coefficients of the final model

From the table above we can see the significance level of the coefficients of our final model and their values. P-values of almost all the variables are lower than 0.05, meaning that they are statistically significant. Only in two of them, that are marital status and the top level of variable look, the p-value is higher than 0.05.

3. Inferential analysis.

Table 5 shows the estimates of the coefficients of the final model. The intercept relates to a married white average-looking man with one year of experience, minimum education, living in a big city in the north of the United States of America and member of a union. His log wage per hour is of 1.2 dollars per hour. Exploring the other coefficients one can notice that some groups of people are penalized in terms of wage. Despite good looking-men seem not to be more payed than average-looking men, this is not valid in the opposite case, where homely men are discriminated with a decrement of 0.15 dollars. Let us remind that the variable wage has been transformed by a logarithmic function, hence even if the differences we are pointing out may seem low, the real effect is exponential. A quantitatively similar discrimination is present in the dichotomous variable union. A man that is not member of a union gains a lower wage then a man who is member of a union. The highest difference between groups is present between the two classes of variable gender. In fact, the two sex register, within same conditions, a difference of 0.42 dollars per hour. Given that, we can state that discrimination between women and men were present and strong in the United States of America at the time of this survey. To conclude, we fit the last linear model to highlight possible discrimination patterns by adding interaction terms between regressors. We were particularly interested in differences between gender and ethnicity.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.137288e+00	8.057865e-02	14.1140083	4.603893e-42
looksGood Looking	-1.661500e-02	3.640329e-02	-0.4564147	6.481716e-01
looksHomely	-1.580171e-01	5.095394e-02	-3.1011743	1.970962e-03
unionno	-1.560344e-01	2.952724e-02	-5.2844208	1.488165e-07
maritalsingle/divorced	-1.398387e-02	3.094248e-02	-0.4519313	6.513974e-01
citymedium	-2.447256e-01	3.665309e-02	-6.6768065	3.673001e-11
citysmall	-1.402950e-01	3.370334e-02	-4.1626433	3.362731e-05
regionsouth	5.667337e-02	3.412629e-02	1.6606953	9.702705e-02
ethnicitywhite:educ275	2.479894e-04	2.695518e-05	9.2000646	1.468925e-19
ethnicityblack:educ275	2.378223e-04	7.419043e-05	3.2055656	1.382301e-03
ethnicitywhite:log(exper)	2.509997e-01	2.152121e-02	11.6629004	6.693906e-30
ethnicityblack:log(exper)	1.988986e-01	3.267741e-02	6.0867295	1.533098e-09
genderfemale:looksAverage	-2.100789e-01	1.061469e-01	-1.9791345	4.802143e-02
genderfemale:looksGood Looking	-1.609179e-01	1.072718e-01	-1.5000958	1.338437e-01
genderfemale:looksHomely	-1.828770e-01	1.240406e-01	-1.4743316	1.406458e-01
educ275:genderfemale	4.504041e-05	4.511779e-05	0.9982849	3.183357e-01
log(exper):genderfemale	-1.157882e-01	3.284437e-02	-3.5253611	4.382955e-04

Figure 6: estimates of the final model with interactions

Almost every coefficient related to the variable looks is not statistically significant. Anyway, we can see that being an average-looking woman penalizes the perceived wage with respect of being an average-looking man. This is also valid for good-looking woman. As consequence, focusing on the variables gender and look, we can report two types of discrimination: the first type is between different gender, the second type is within the gender being female. In fact, prettiest women are less penalized. Another traditional source of discrimination is the ethnicity. In our dataset this source is present. The same increment in education or experience leads to a greater increment in wage, for white people, than for black people. The last type of discrimination is between the dimension of the city: people living in a small city or in a medium one can be paid less.

4. Conclusions

In this paper we explored a subset of a survey of the U.S. wages. We saw how to detect violations of the linear regression assumptions and a possible way to deal with them. The violation that affected the most our model was heteroscedasticity even if Ordinary Least Squares regression is, in some way, consistent to it. Another nontrivial problem was due to outliers. In our case, we did not know whether influential points were outliers or points suggesting mis specified parameters of the model. Finally, we obtained a model that, for sure, is not perfect but it could certainly be used to reach our aim of pointing out the discriminatory patterns. We also tried Weighted Least Squares regression but, since we did not know the source of heteroscedasticity, we only applied a common formula of the weights, that is Weights = $1/x_i$. The resulting model showed a similar standardized variance between the bottom (0.69) and the top (0.94) half of the distribution, suggesting that we were working in the right direction. However, regressing the residuals leads to a strong distortion of the dataset and the interpretation of the model can be misleading.