Hochschule Rhein-Waal
Rhine-Waal University of Applied Sciences
Faculty of Communication and Environment

Degree Program Information Engineering and Computer Science, M. Sc.

Analysis of Wage Data from Mid-Atlantic region of the U.S.

Kultigin Bozdemir

24205

Assessment Paper

SS 2019

Module: “Data Analysis/Statistics”

This report is the result of my own work. Material from the published or unpublished work of others, which is referred to in the report, is credited to the author in the text.

Table of Contents

List of Abbreviations

List of Figures

Introduction

In this study, we are going to examine a number of factors that relate to wages for a group of males from the Atlantic region of the United States. We will particularly try to understand the association between an employee's age, educational attainment, as well as the calendar year with the wage. Besides that, we will have a look at the other factors as well, which are also provided in the data set.

The data provided by the lecturer is from US Census Bureau which covers the data of 3000 employees from Mid-Atlantic region of the U.S. The data set has 12 variables.

We will follow the structure of the course, first we will explore the data with histograms and plots, then analyze the associations between variables, finally develop a model for the data set. Through the R script, we will use different libraries and methods. Those are explained shortly in the relevant section of the paper and in the R script. Although different methods were used, only the important results have been written in the report to keep the study simple and understandable. It is recommended to read the study together with the R script.

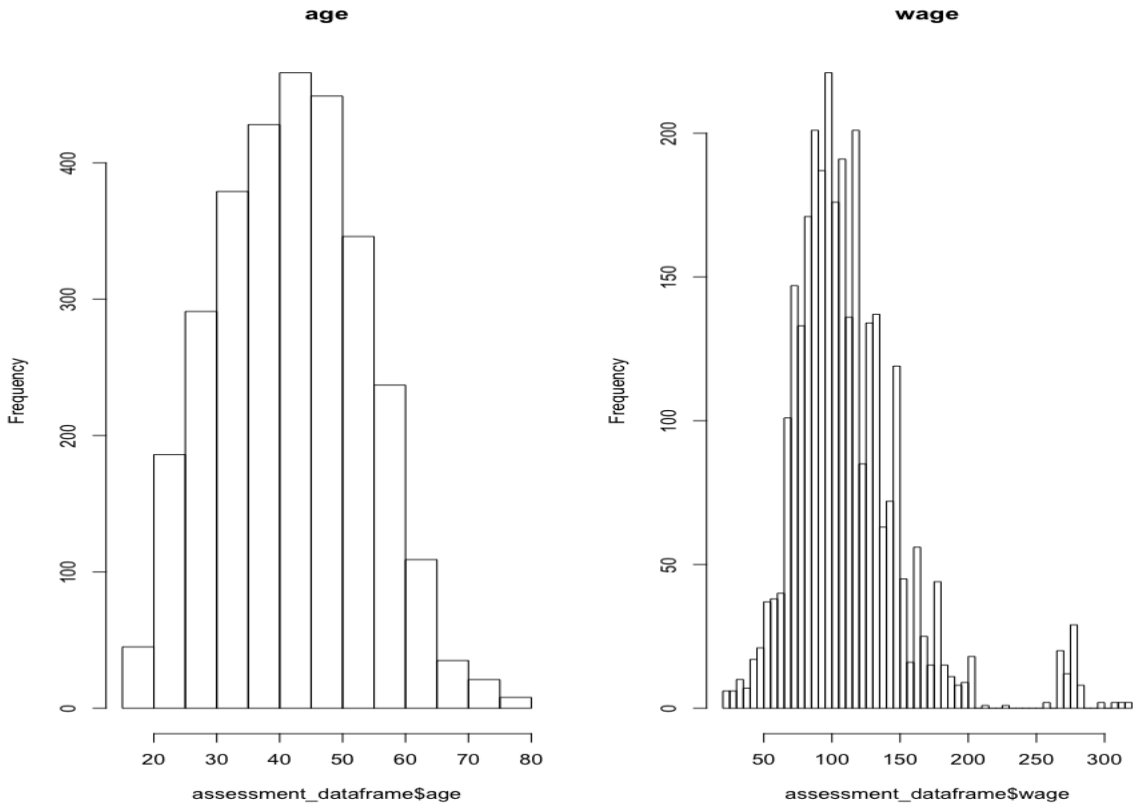
The research question is that there is a correlation between wage and other mentioned variables either negative or positive (non-directional). Thus, our null-hypothesis is no correlation between them.

1. Exploring the data

The list of variables is; year(year of the observations of 2003-2009), age, maritl (marital status, 5 categories), race (4 categories), education (5 categories), region (Mid-Atlantic is the sole one), jobclass (information or industrial), health(good or very good), health_ins (health insurance, yes or no), logwage and wage.

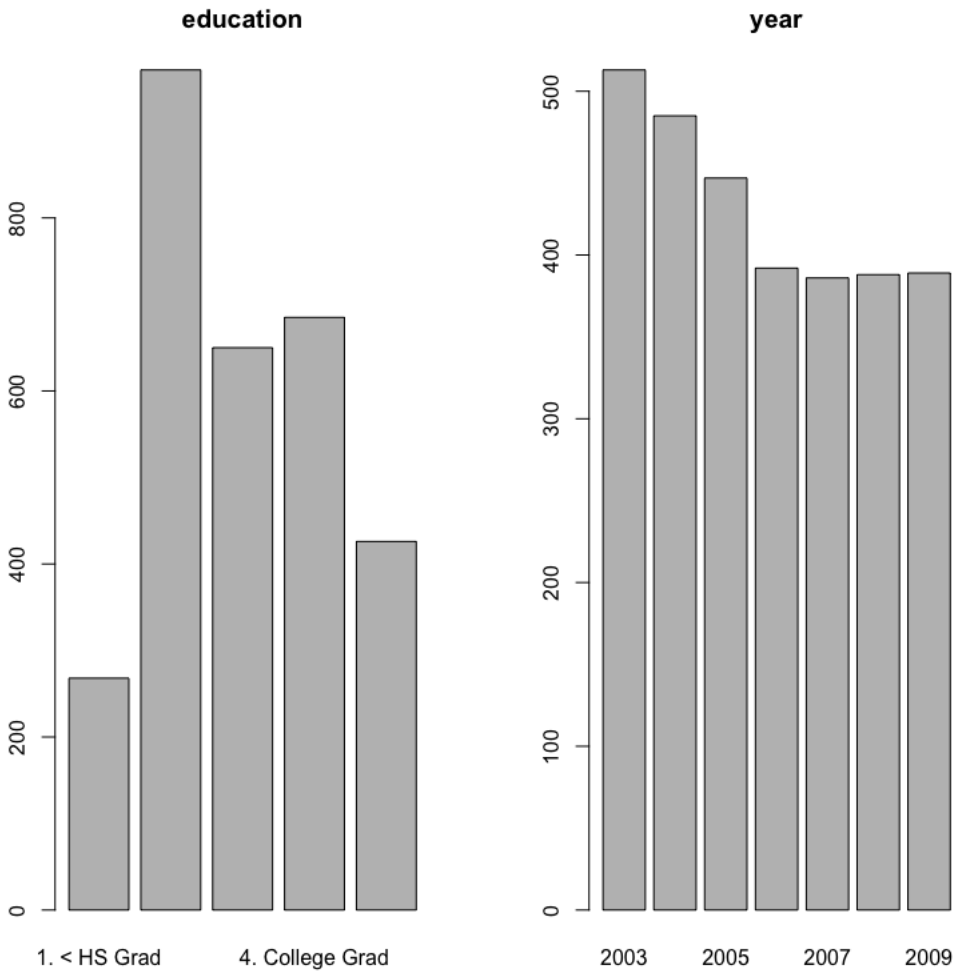
We plot histograms of age and wage. Age variable seems to have a normal distribution at the mean of around 42 and wage seems to have also a normal distribution if we exclude the outliers, which are higher than 250. Having no reason to exclude them, we will continue the full observation to explore the data.

Figure 1 Histograms of age and wage



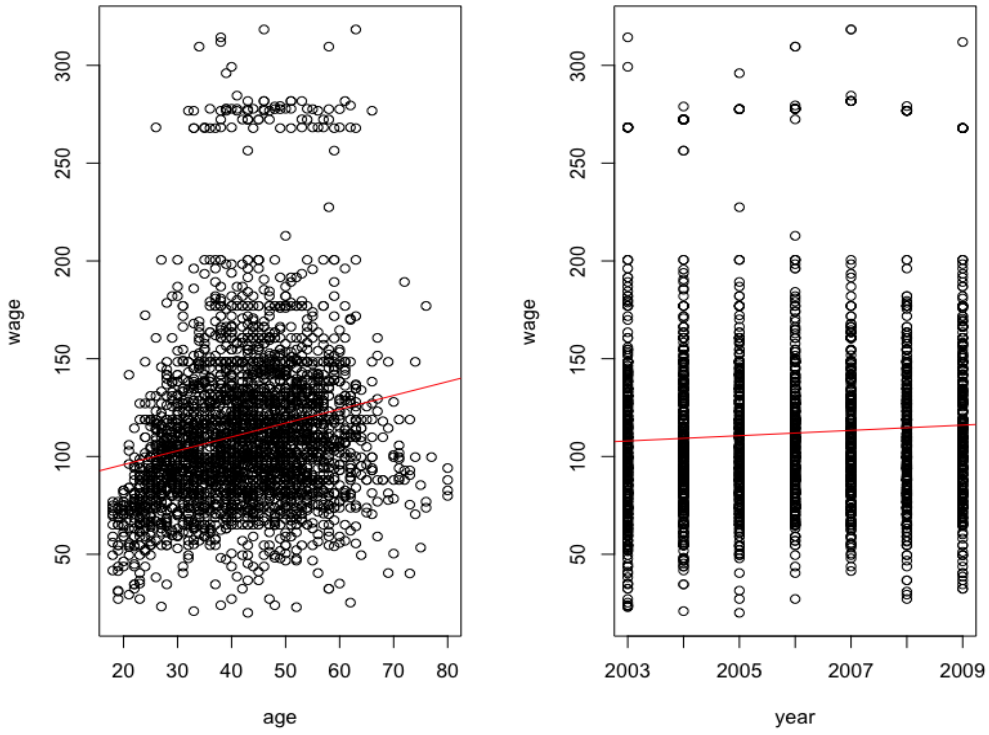
As seen on the histograms of education and year. The second “HS Grad” of education is obviously larger than categories in education. The issue of having different sample sizes for the categories in a variable has to be taken into account, since they might be important in terms of f-test and confidence intervals in the further steps.

Figure 2 Barplots of education and year



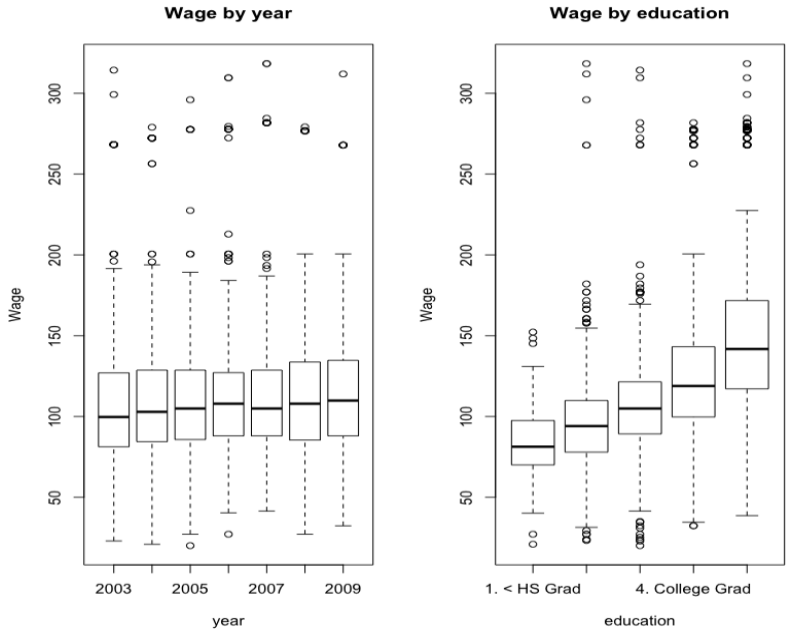
We plot age and year with wage subsequently. As seen on the scatter plots, we have created also a linear fit, although we are still in the exploration phase of this study. Nevertheless, they have a significant effect. Both p-values are less than 5%. The existence of outliers on the age plot is very clear which are above 250. Years have almost a uniform distribution as we will see soon more clearly.

Figure 3 Scatter plots of age and year to wage



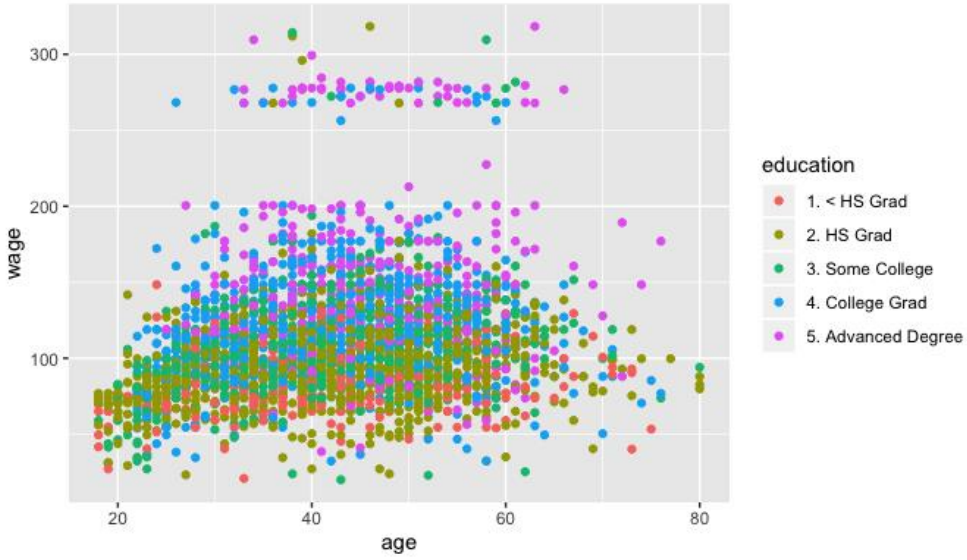
On the boxplots below, we see a slightly increasing wage until 2007, where we have a drop, as mentioned previously. If we do a pre-assessment, a linear fit would not be okay for this variable. On the second boxplot, it is clear that higher education pays off.

Figure 4 Boxplots of year, education, marital status and job class



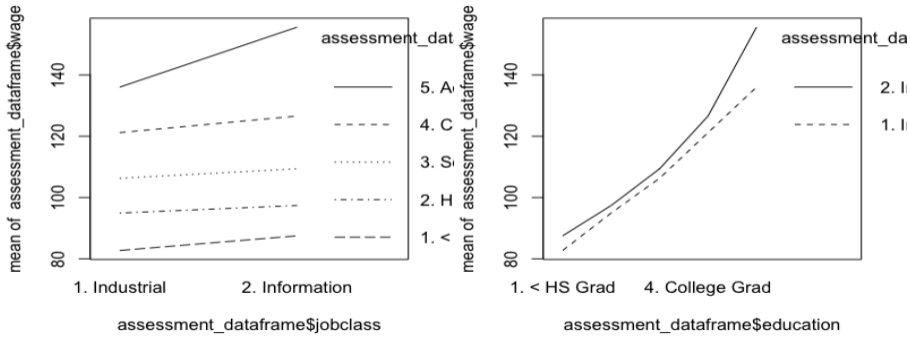
With the help of ggplot2 package, we plot the data in three dimensions. Better education yields better wages. By around age of 30, we see advanced degrees. Secondly, advanced degree dominates the highly paid outlier section (>250).

Figure 5 Plot of wage~age with education



It is very crucial to be sure if there is an relation between independent variables, because they might lead us to faulty judgements about their effects on the response variable, in our case on the "wage". As seen below, there is a significant interaction between high degree education and information category of job classes. Unparallel lines on the graphs below indicate the existence of an interaction. However, we will come back to this issue again, when we will finalize our model. In the next section, we will see also correlations between the independent variables.

Figure 6 Interaction plots



2. Correlations

We have already seen there are some correlations between the response (wage) and predictors and also among some of predictors. A correlation matrix is created below with the cor () function in R. the pair of education-wage is the highest one, as seen in the table. We would like noteworthily to state that, partial correlation drops down to 0.16 if we calculate its partial correlation caused by other predictors.

Table 1 Correlation Matrix

	wage	age	year	education_Code
wage	1			
age	0.19563720	1		
year	0.06554428	0.03842466	1	
education_Code	0.47577490	0.07078425	0.01411597	1

We have continued correlation analysis with cor.test () function. But before, we created dummy variables for education and marital status. Remember, their distributions are not normal, too, as we have seen above.

Since, the education is qualitative and not normal distribution, Spearman method should give us a more accurate correlation result. Not surprisingly, Spearman method gives us a higher correlation degree than results of Pearson method in these categorial variables.

In the cor.test, p-value tells us the rejection of the null hypothesis, because it is very smaller than 5%. In the same function, F-test is also very higher than 1, which justifies the associations are not coincidence.

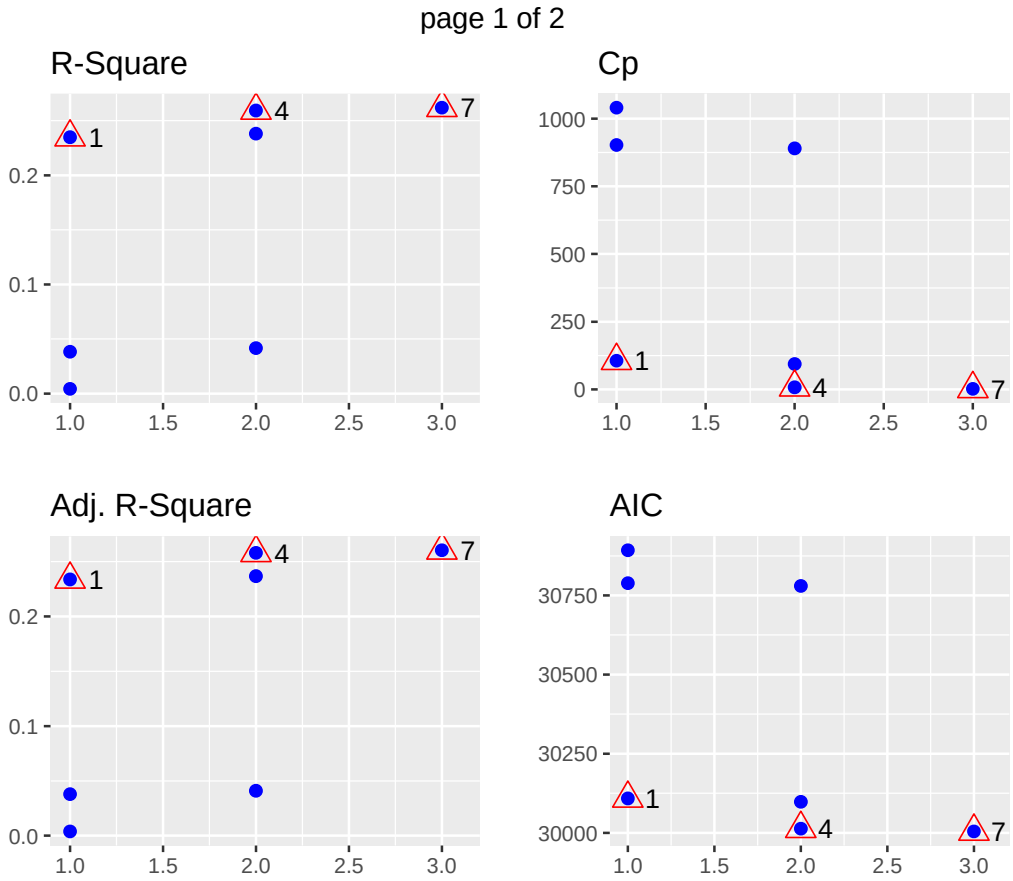
3. Predictors

We create a primitive model below following the previous analysis to decide on the number of predictors.

```
lm(wage~age + year  + education, data=assessment_dataframe)
```

We have used All Possible Regression analysis of the package of oslrr. It is seen that until 2nd predictor there is clear improvement. Those 2 predictors (4th model): are age and education. Nevertheless, we are not obliged to drop some predictors, since we have no problem such as computation complexity. As a second method, we used ISLR and leaps libraries to find the best predictors for our model. The second method justifies the first one completely.

Figure 7 Comparison of predictors



4. Modelling

4.1. Polynomial Regression

We will start the modelling analysis with one of quantitative variables, “age”. We introduced the polynomial models between wage~age up to the polynomial degree of 5. We compared these models with anova () function. In this analysis, we followed the discussions of James et al. in their book, chapter 7 (James, et al., 2013, pp. 265-301).

We can interpret the anova table in the following way. We test the null hypothesis that model 1 is sufficient to explain the data against Model 2.
P-value of which compares Model 1 to Model 2 is 2.2e-16. That says linear fit (Model 1) is insufficient. With the same token, p-value which compares Model 2 to Model 3 is also low; 0.001679. That says cubic model performs better than quadratic one. P-value which compares cubic model to 4-polynomial model is around 5%. Finally, p-value of model 4 to model 5 is 37%. This anova table says either model 3 or model-4 is better than other lower or higher polynomials.

Analysis of Variance Table

Model 1: wage ~ age

Model 2: wage ~ poly(age, 2)

Model 3: wage ~ poly(age, 3)

Model 4: wage ~ poly(age, 4)

Model 5: wage ~ poly(age, 5)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2998	5022216				
2	2997	4793430	1	228786	143.5931	< 2.2e-16 ***
3	2996	4777674	1	15756	9.8888	0.001679 **
4	2995	4771604	1	6070	3.8098	0.051046 .
5	2994	4770322	1	1283	0.8050	0.369682

Below, we plotted polynomial degree 3 function with its confidence intervals.

Figure 8 Polynomial degree 3 regression

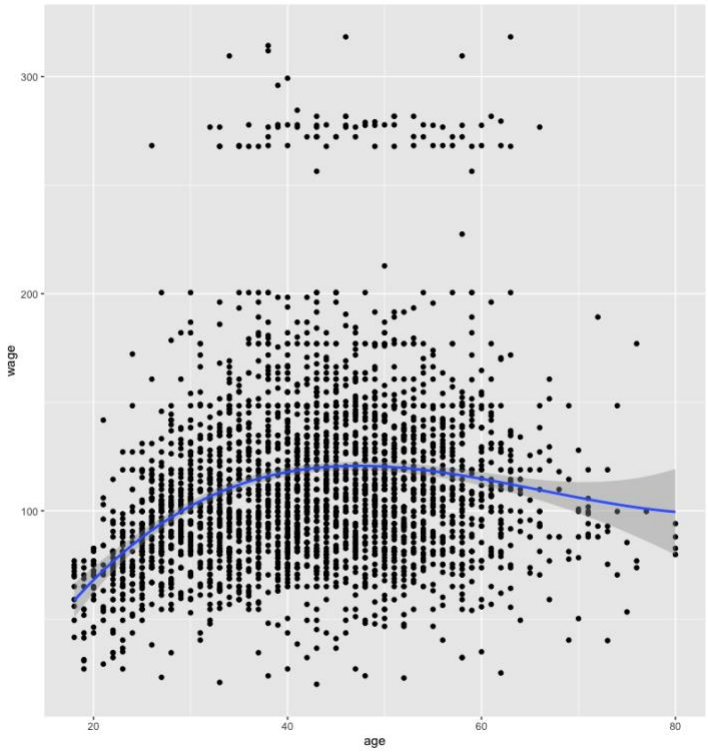
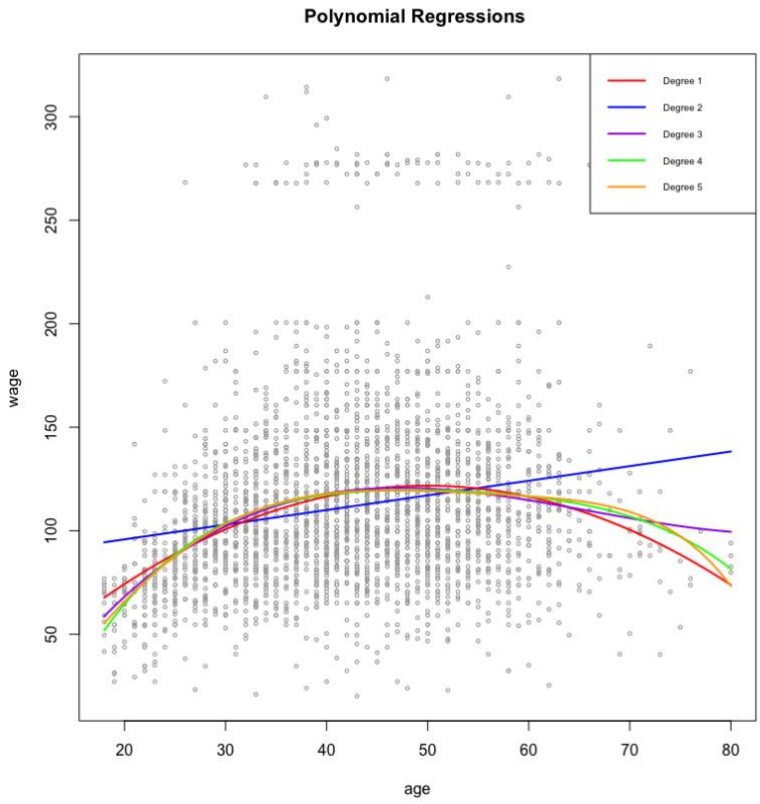


Figure 9 Polynomial regressions



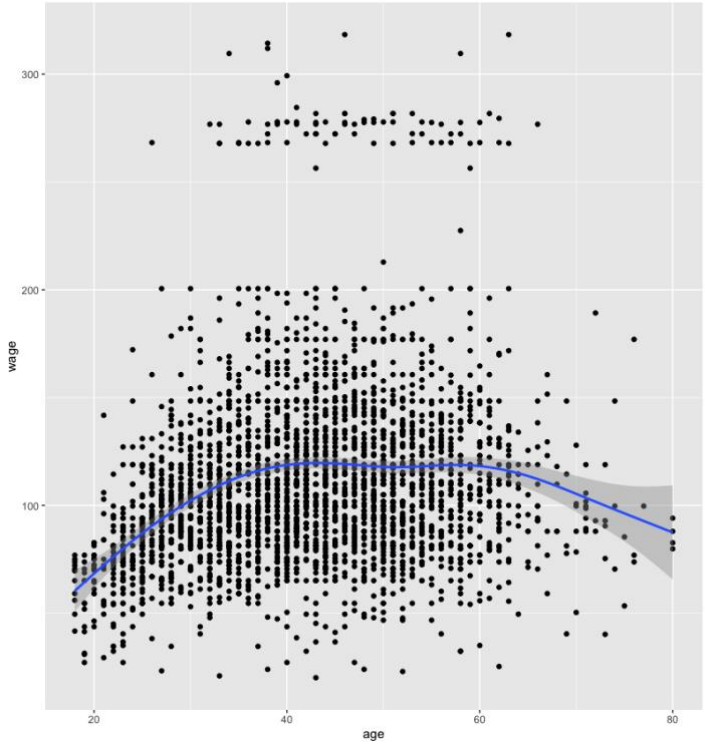
Above, all polynomial regressions are plotted together. As we discussed above, the graph shows that polynomial 2 to 5 degree lines are very close to each other, but they are obviously distinguished from the linear line.

4.2. Local Regression

Secondly, we analyzed wage~age pair with local regression method. It is a neighborhood clustering method to define the average of the function at that point by using kernel weighting (Hastie, et al., 2009, p. 194).

loess () function provides the local regression. Span parameter defines the range of neighborhood. Lower span factors give a wild fluctuating function, whereas higher ones will give smoother functions. Having no reason to have a wild function in our case, we would prefer to be around 0.5. Accordingly, we plotted the model with confidence level, which is illustrated with grey zones on both sides of the regression.

Figure 10 Local regressions

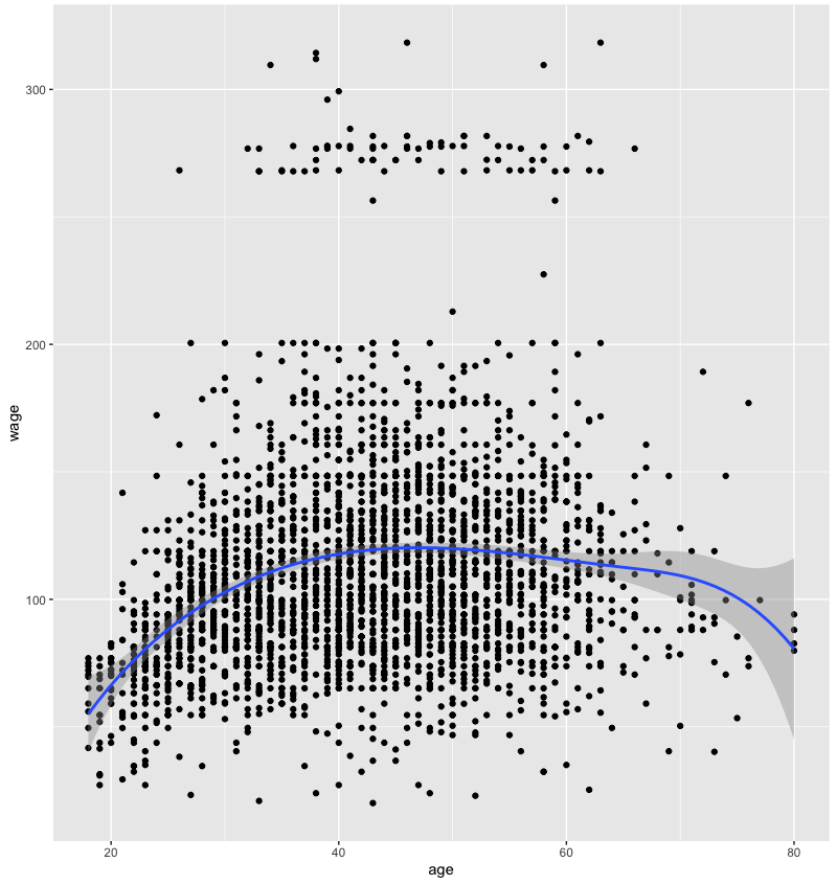


4.3. Splines Regression

With the help of splines library, we can use smoothing method for the fit.

The bs () function generates basis functions (by default cubic) at the segments which are divided by the defined knots, in our case 30 and 65. That means we have 6 degrees of freedom in total. Three coefficients come from three different segments which are divided by two knots (30 and 65). Since the default function is cubic, it has three coefficients. Now at hand, we have piecewise polynomial regressions. This approach provides us a continues function, because second derivative of the cubic function is not zero (James, et al., 2013, p. 273).

Figure 11 Spline regression

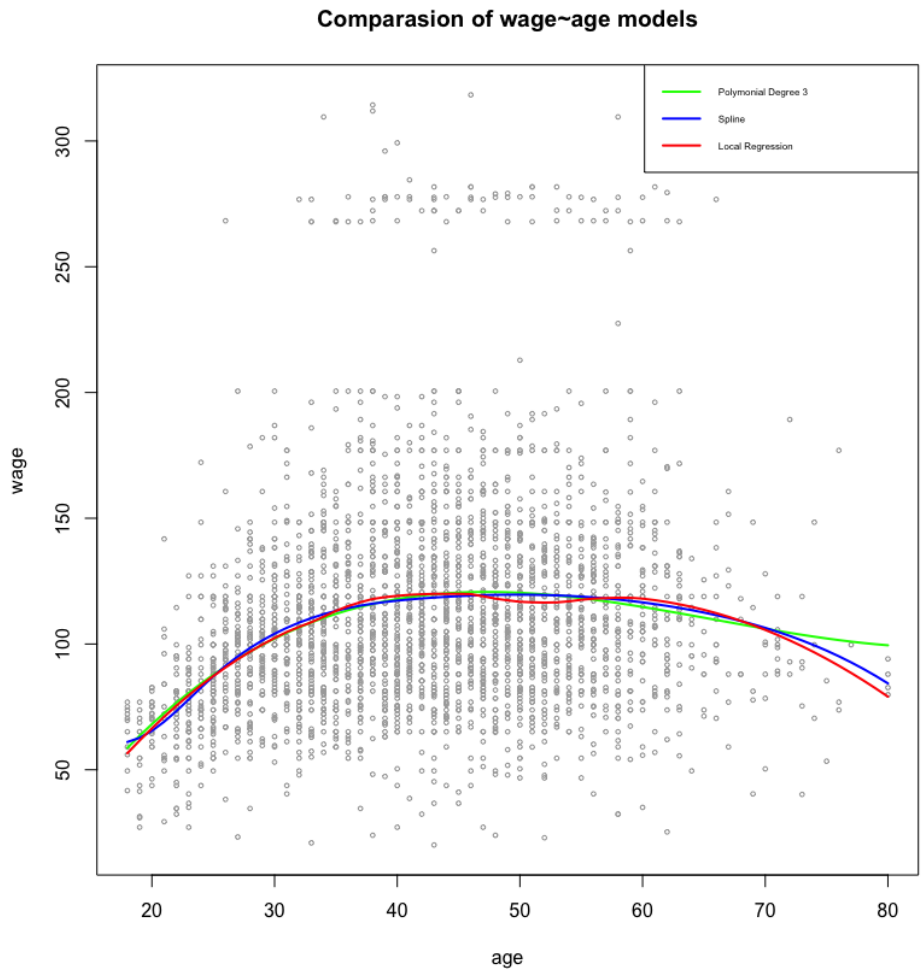


4.4. Comparasion of the models

Spline regression models have an upper hand over polynomial function. They provide a more stable function, since Spline models mainly use knots to have the degree of freedoms. On the other hand, polynomial models use polynomials which are more complicated in terms of degree of freedom. To have an accurate function, they have to use high polynomial degrees. Besides that, piecewise approach allows us to analyze the data easier such as giving idea for sliding the data, etc. Finally, polynomial functions might be very wild on the tails (James, et al., 2013, p. 273).

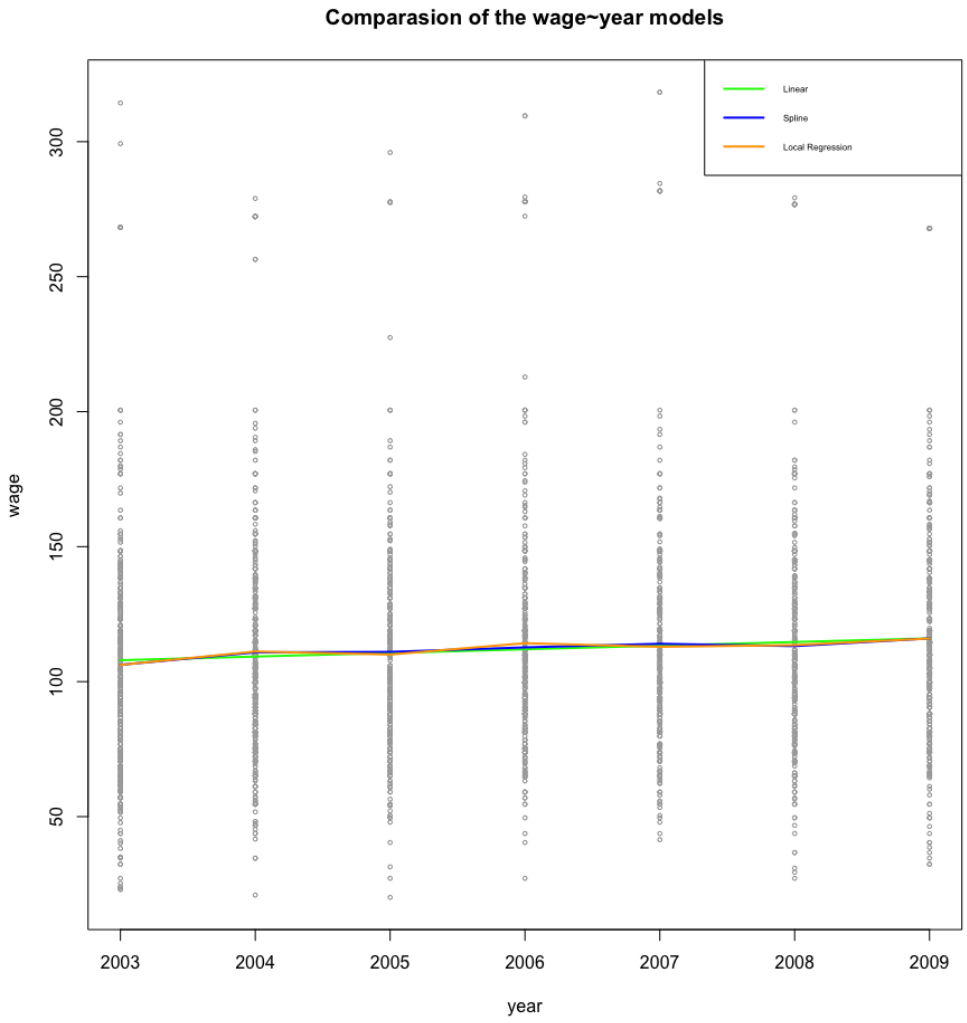
The regression lines of the models which we have analyzed up to now are shown on the graph below. Those methods are “polynomial”, “Spline” and “local regression”. As seen on the graph, the regression lines are similar in our case. Differences appear on the tails. We did anova analysis of the models. Since local regression is non-parametric method, it is not possible to analyze it with anova. Therefore, we continued the analysis with cross-plotting the models and cor.test () function. Those analysis justify the graph below, there is no any significant difference between them in terms of fitness.

Figure 12 Comparison of wage~age models



Up to this point, we have analyzed only wage-age pair. Now, we moved to wage-year pair, but this time, we will show only the regression lines, since we do not want to repeat same discussions again. We have plotted lines of linear, Spline and Local regression models. Surprisingly, nonlinear models do not perform better than linear one. Nevertheless, we will used Spline model, since we had already seen in the "exploring" section the fact that wage increase is not linear, there is a drop in 2007. We applied again anova, cross-plotting of the models and cor.test analysis, but that did not change our initial conclusion.

Figure 13 Comparison of wage~year models



4.5. Final Model

After discussions and comparisons, we have selected the model below. But we have used sign of "*" instead of "+" to see the interactions in the model with the help of anova () function, as we promised earlier.

fit0= lm(wage~bs(age,knots = c(25,40,65)) * bs(year, knots = c(2006, 2008)) * education, data=assessment_dataframe)

Analysis of Variance Table

Response: wage

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bs(age, knots = c(25, 40, 65))	6	450117	75020	62.2685	< 2.2e-16 ***
bs(year, knots = c(2006, 2008))	5	23269	4654	3.8628	0.001718 **
education	4	1049971	262493	217.8769	< 2.2e-16 ***
Residuals	2979	3589025	1205		

After looking anova table for the values of F-test and Pr, we have found again same significant interaction; education:age.

All in all, when we check F and Pr values in the anova table, the model proves the existence of the significant associations between wage and other defined variables.

Although we have not seen any important difference between nonlinear wage~age models, we have chosen Spline model for age, since they are more stable in general, as discussed earlier. We defined the knots manually, although there are some sophisticated spline functions which can choose the best knots (number of knots and their locations).

Because this approach allows us to do further analysis easier on the data, such as dividing, sliding or introducing constraints etc.

Nonlinear wage~year models have not given us much benefit in terms of the fitness of the models. Nevertheless, it was clear on the boxplot of the years that the increase of the wages is not linear. Therefore, we did not select the linear one, but again Spline model.

lm () function generates dummy variables for categorial variables such as education.

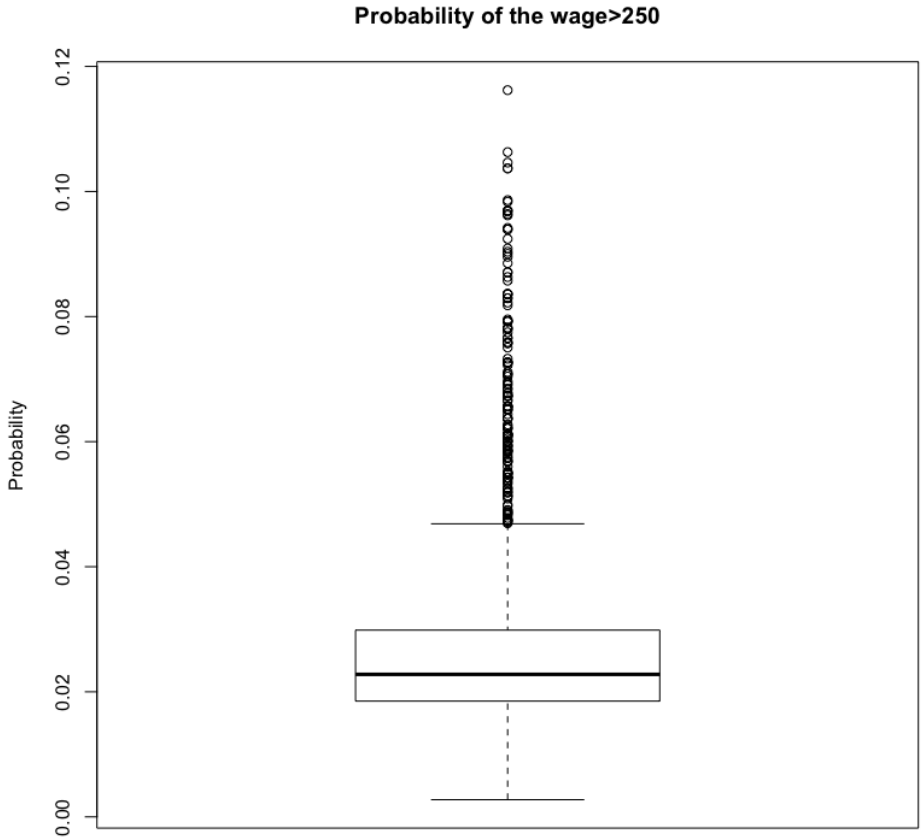
If we plot fitted.values and residuals, we can conclude that the model does not indicate a significant heteroscedasticity.

As seen in the summary () function, they are both statistically significant. Overall adjusted r-squared is increased as well. Adjusted r-squared is 0.28, which says that the fit explains 28% of the data.

5. Further Analysis

We have discovered some distinctions while we were exploring the data, such as a drop of wages in 2007 or outliers of >250 in the wages. Needless to say, we can slide the data at that points and continue analysis and introduce more accurate models. However, these are not the subjects of this study.

Figure 14 Probability of the wage>250



Nevertheless, we did logistic regression analysis to see how it is possible to predict these outliers who earn more than 250. We have used the glm () function, instead of lm (), which is allowing logistic regression. Surprisingly, given the values of the independent variables, the probability of predicting highly paid people is very low, around 0.02, as shown on the boxplot above (James, et al., 2013, p. 154).

Conclusion

In this study, we initially explored the data of "Wages of Mid-Atlantic region of the U.S.". It was clear in the histograms and scatter plots that linear models would not fit. Besides that, we have discovered some interesting distinctions such as outliers of people who are getting more

than 250 thousand. Additionally, we have seen the pattern of the independent variables, which might help us in the following phases when we are constructing our model.

Then, in a different chapter we analyzed the importance of the predictors for our draft model. The most important predictors appeared as education and age. However, we included year as well, since there is no problem with computation resources. In the light of this analysis, we started modelling with numeric predictors. We have compared linear, spline, local and polynomial functions. Finally, we have decided on Spline functions, mainly because of their stability. Thereafter, we completed our final model after generating dummy variables for categorial variables. Although we have chosen spline for nonlinear regression, we have to admit that other methods, namely polynomial and local regression perform almost identically in our data set at least.

The model rejects our initial null hypothesis. In other words, there is a significant correlation between wage and others.

Finally, we used logistic regression method to figure out if we can predict highly paid people. Actually, the probability of that was happened around 0.02

Needless to say, further analysis can be done by sliding the data into pieces or removing outliers. Nevertheless, these further analyses are left to other researchers.

Bibliography

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning.* 2nd Edition Hrsg. New York: Springer.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An Introduction to Statistical Learning with Applications in R.* New York: Springer.