Tampereen yliopisto

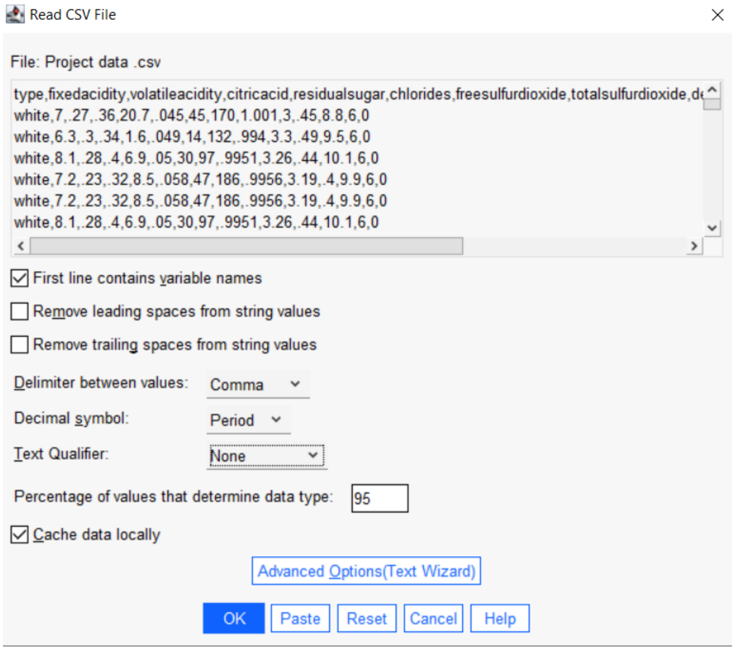
Enna Augustin

PROJECT WORK
Basics for Business Data Analytics

December 2022

1. DATA CHECKING AND PRE-PROCESSING

The data analysis was started by opening the provided data in SPSS. Since data was in csv-form, some settings had to be set before we could easily view the data. These settings were as shown in figure 1.



Read CSV File

File: Project data .csv

```
type,fixedacidity,volatileacidity,citricacid,residualsugar,chlorides,freesulfurdioxide,totalsulfurdioxide,de
white,7,.27,.36,20.7,.045,45,170,1.001,3,.45,8.8,6,0
white,6.3,.3,.34,1.6,.049,14,132,.994,3.3,.49,9.5,6,0
white,8.1,.28,.4,6.9,.05,30,97,.9951,3.26,.44,10.1,6,0
white,7.2,.23,.32,8.5,.058,47,186,.9956,3.19,.4,9.9,6,0
white,7.2,.23,.32,8.5,.058,47,186,.9956,3.19,.4,9.9,6,0
white,8.1,.28,.4,6.9,.05,30,97,.9951,3.26,.44,10.1,6,0
```

☑ First line contains variable names

☐ Remove leading spaces from string values

☐ Remove trailing spaces from string values

Delimiter between values: Comma

Decimal symbol: Period

Text Qualifier: None

Percentage of values that determine data type: 95

☑ Cache data locally

Figure 1. *Settings for the csv-data*

In the settings of figure 1 the delimiter between values was set as a comma, decimal symbol was set as a period and the test qualifier was set as “none”.

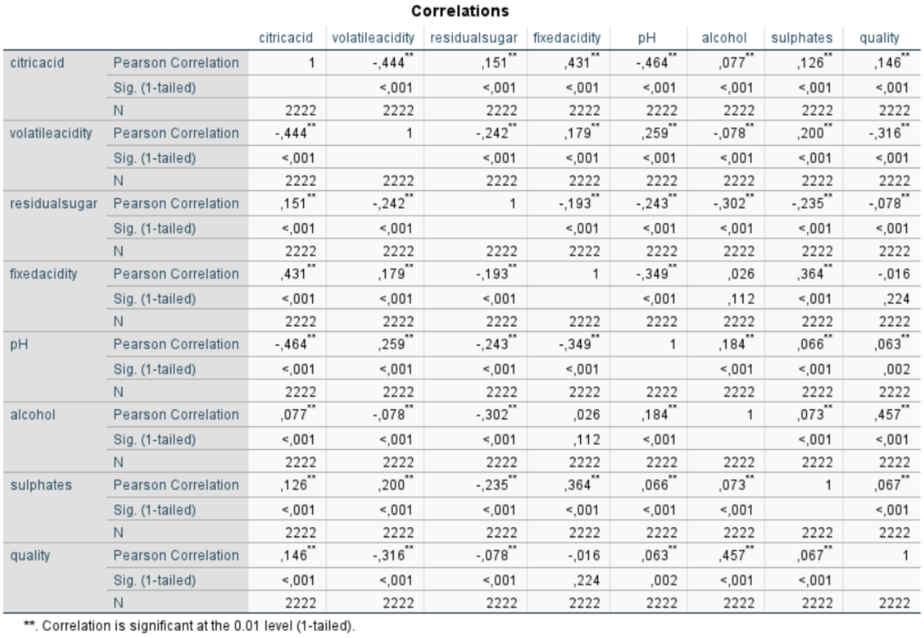
The quantity of the data is quite big. It has 2499 rows for each of the 14 columns. The data consists of string and numeric data. The data has both nominal and ordinal data, for example the “quality good or not” is nominal data and the “fixed acidity” is ordinal data. Considering the factors above, the amount of data should be ideal for the implementation of good quality and versatile analysis.

The quality of data is at first glance okay, but there are quite a few empty values in the set. The cleaning of the data was started by removing the rows that had these empty values. Next the outlying values were removed. For example, the variable “free sulfur dioxide” had two values that were over 100. Both the outliers and empty values were easy to detect when the values were sorted ascending. Thirdly the duplicate values, meaning rows having exactly same values in each of the columns. These extra rows were removed so that the results would not twist. After these operations there were 2222 rows in the dataset.

2. THE CORRELATIONS OF VARIABLES

In this section we explore the correlations between different variables. The explored correlations were between variables “citric acid”, “volatile acidity”, “residual acidity”, “fixed acidity”, “pH”, “alcohol”, “sulphates” and “quality”. I decided to compare these variables because based on the descriptions provided in the assignment document these variables seemed like they could have interesting correlations.

The correlation analysis was started by setting the variables and selecting settings. The selected correlation coefficient was Pearson, and the selected test of significance was one-tailed. Other settings were kept as default. When the correlation analysis was performed correlations were presented in a table. These correlations are presented in figure 2.



Correlations

		citricacid	volatileacidity	residualsugar	fixedacidity	pH	alcohol	sulphates	quality
citricacid	Pearson Correlation	1	-,444**	,151**	,431**	-,464**	,077**	,126**	,146**
	Sig. (1-tailed)		<,001	<,001	<,001	<,001	<,001	<,001	<,001
	N	2222	2222	2222	2222	2222	2222	2222	2222
volatileacidity	Pearson Correlation	-,444**	1	-,242**	,179**	,259**	-,078**	,200**	-,316**
	Sig. (1-tailed)	<,001		<,001	<,001	<,001	<,001	<,001	<,001
	N	2222	2222	2222	2222	2222	2222	2222	2222
residualsugar	Pearson Correlation	,151**	-,242**	1	-,193**	-,243**	-,302**	-,235**	-,078**
	Sig. (1-tailed)	<,001	<,001		<,001	<,001	<,001	<,001	<,001
	N	2222	2222	2222	2222	2222	2222	2222	2222
fixedacidity	Pearson Correlation	,431**	,179**	-,193**	1	-,349**	,026	,364**	-,016
	Sig. (1-tailed)	<,001	<,001	<,001		<,001	,112	<,001	,224
	N	2222	2222	2222	2222	2222	2222	2222	2222
pH	Pearson Correlation	-,464**	,259**	-,243**	-,349**	1	,184**	,066**	,063**
	Sig. (1-tailed)	<,001	<,001	<,001	<,001		<,001	<,001	,002
	N	2222	2222	2222	2222	2222	2222	2222	2222
alcohol	Pearson Correlation	,077**	-,078**	-,302**	,026	,184**	1	,073**	,457**
	Sig. (1-tailed)	<,001	<,001	<,001	,112	<,001		<,001	<,001
	N	2222	2222	2222	2222	2222	2222	2222	2222
sulphates	Pearson Correlation	,126**	,200**	-,235**	,364**	,066**	,073**	1	,067**
	Sig. (1-tailed)	<,001	<,001	<,001	<,001	<,001	<,001		<,001
	N	2222	2222	2222	2222	2222	2222	2222	2222
quality	Pearson Correlation	,146**	-,316**	-,078**	-,016	,063**	,457**	,067**	1
	Sig. (1-tailed)	<,001	<,001	<,001	,224	,002	<,001	<,001	
	N	2222	2222	2222	2222	2222	2222	2222	2222

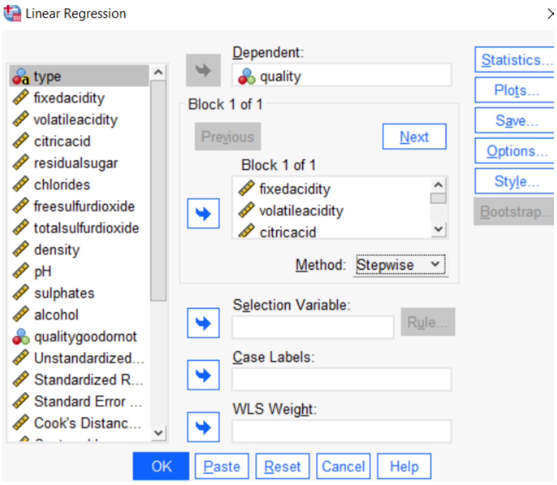
**. Correlation is significant at the 0.01 level (1-tailed).

Figure 2. *Correlations*

From figure can be seen that citric acid has highest correlation with fixed acidity, volatile acidity has highest correlation with pH, residual sugar has highest correlation with citric acid, fixed acidity has highest correlation with citric acid, pH has highest correlation with volatile acidity, alcohol has highest correlation with quality, sulphates have highest correlation with fixed acidity and quality has highest quality with alcohol.

3. STEPWISE REGRESSION ANALYSIS

In this section we perform a stepwise regression analysis. We performed a linear regression analysis where the dependent variable was quality, and the other explored variables were independent variables. For the linear regression to be stepwise regression the method had to be changed to “stepwise”. This can be seen in figure 3.



Dependent: quality

Block 1 of 1

Block 1 of 1: fixedacidity, volatileacidity, citricacid

Method: Stepwise

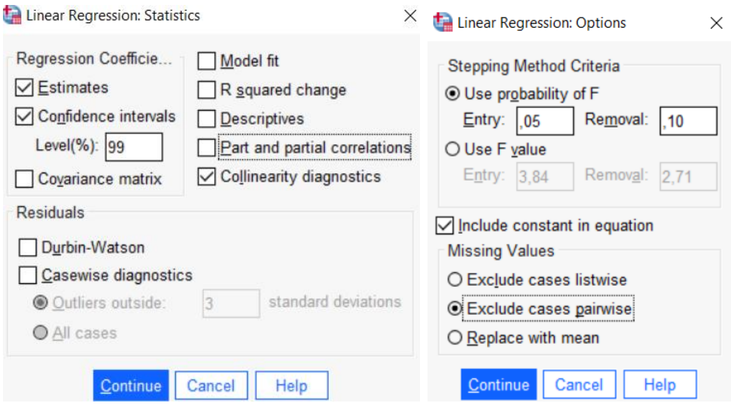
Selection Variable:

Case Labels:

WLS Weight:

Figure 3. *Changing the method and selecting variables*

The next step in the analysis was to select some settings from the menus “statistics”, “options” and “plots”. These settings are presented in figures 4 and 5.



Regression Coefficie...

☑ Estimates

☑ Confidence intervals

Level(%): 99

□ Covariance matrix

□ Model fit

□ R squared change

□ Descriptives

□ Part and partial correlations

☑ Collinearity diagnostics

Residuals

□ Durbin-Watson

□ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

Stepping Method Criteria

● Use probability of F

Entry: ,05 Removal: ,10

○ Use F value

Entry: 3,84 Removal: 2,71

☑ Include constant in equation

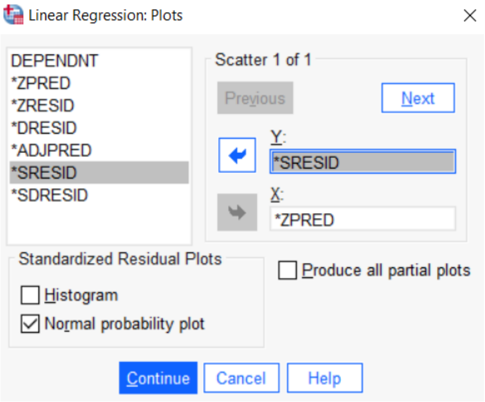
Missing Values

○ Exclude cases listwise

● Exclude cases pairwise

○ Replace with mean

Figure 4. *Statistics and options settings*



Linear Regression: Plots

DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID

Scatter 1 of 1

Y: *SRESID

X: *ZPRED

Standardized Residual Plots

□ Histogram

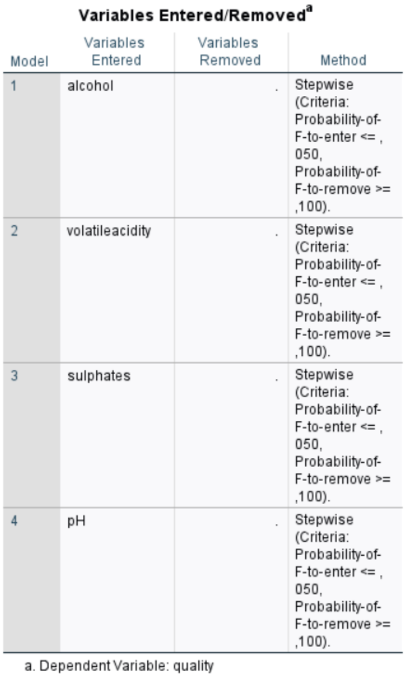
☑ Normal probability plot

□ Produce all partial plots

Figure 5. *Plots settings*

In the statistics settings the selected settings allow to generate several statistics. The most important in this context is "Collinearity diagnostics". The “confidence intervals level” was also increased to 99 %. In the options settings selecting “exclude cases pairwise” uses as many cases as possible for computing the correlations on which our regression is based. The plots settings draw the data into a plot and compares it to a linear regression.

After these settings the stepwise regression analysis was performed. Figure 6 shows the variables to build the models.



Variables Entered/Removed[a]

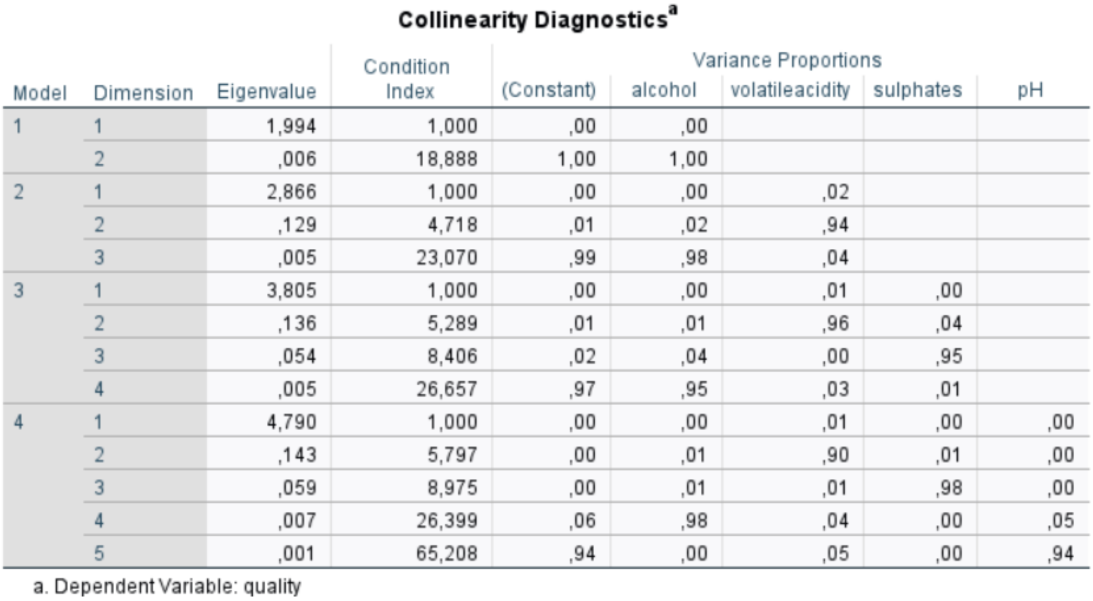
Model	Variables Entered	Variables Removed	Method
1	alcohol	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	volatileacidity	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
3	sulphates	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
4	pH	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: quality

Figure 6. *Variables used in the models*

In the table in figure 6 SPSS has been excluding some of the selected variables from the table. This is because SPSS uses these variables as a reference category to be able to interpret the coefficients in the regression analysis.

Figure 7 shows the "Collinearity Diagnostics" box.



Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	alcohol	volatileacidity	sulphates	pH
1	1	1,994	1,000	,00	,00			
	2	,006	18,888	1,00	1,00			
2	1	2,866	1,000	,00	,00	,02		
	2	,129	4,718	,01	,02	,94		
	3	,005	23,070	,99	,98	,04		
3	1	3,805	1,000	,00	,00	,01	,00	
	2	,136	5,289	,01	,01	,96	,04	
	3	,054	8,406	,02	,04	,00	,95	
	4	,005	26,657	,97	,95	,03	,01	
4	1	4,790	1,000	,00	,00	,01	,00	,00
	2	,143	5,797	,00	,01	,90	,01	,00
	3	,059	8,975	,00	,01	,01	,98	,00
	4	,007	26,399	,06	,98	,04	,00	,05
	5	,001	65,208	,94	,00	,05	,00	,94

a. Dependent Variable: quality

Figure 7. *Collinearity Diagnostics box*

Figure 7 shows how the variables vary with each other. When two or more of the independent variables are correlated, the condition index for each will be above one. From the figure we can see that alcohol correlates with volatile acidity.

Figure 8 presents the plot of the data.

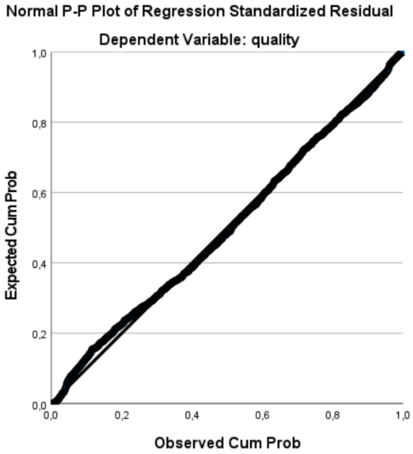
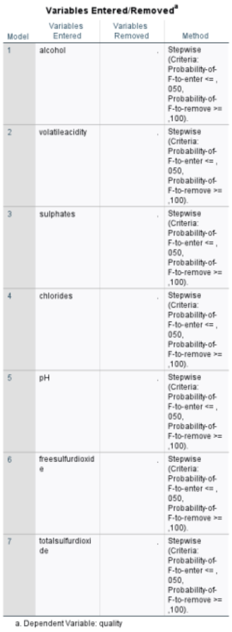


Figure 8. *Linear plot*

From figure 8 we can see how the data follows linearity almost perfectly.

4. ROBUSTNESS CHECK

The robustness of the analysis can be checked by reanalyzing the data using other analytical models. For the robustness check in this case, we selected to perform the same analysis as above but with some noise in the data. This means that we added rest of the variables in the dataset to the analysis. The results for the analysis are presented in figures 9, 10 and 11.

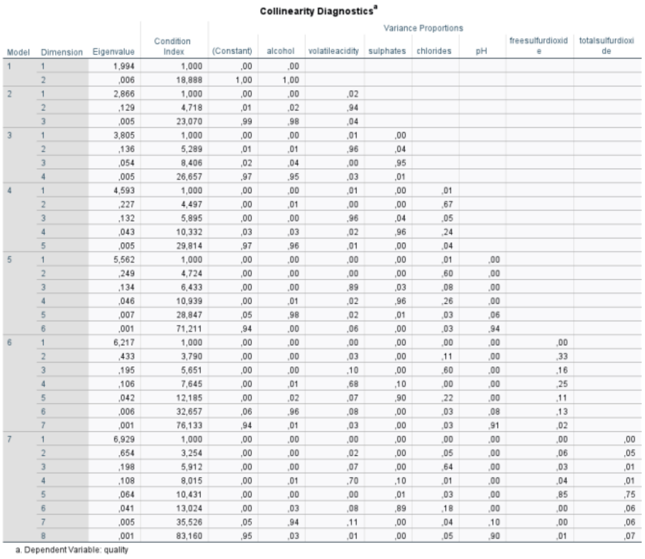


Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	alcohol	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	volatileacidity	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
3	sulphates	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
4	chlorides	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
5	pH	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
6	freesulfurdioxide	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
7	totalsulfurdioxide	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: quality

Figure 9. *Variables*



Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	alcohol	volatileacidity	sulphates	chlorides	pH	freesulfurdioxide	totalsulfurdioxide
1	1	1,994	1,000	,00	,00						
	2	,006	18,888	1,00	1,00						
2	1	2,866	1,000	,00	,00	,02					
	2	,129	4,718	,01	,02	,94					
	3	,005	23,070	,99	,98	,04					
3	1	3,805	1,000	,00	,00	,01	,00				
	2	,136	5,289	,01	,01	,96	,04				
	3	,054	8,406	,02	,04	,00	,95				
	4	,005	26,657	,97	,95	,03	,01				
4	1	4,593	1,000	,00	,00	,01	,00	,01			
	2	,227	4,497	,00	,01	,00	,00	,67			
	3	,132	5,895	,00	,00	,96	,04	,05			
	4	,043	10,332	,03	,03	,02	,96	,24			
	5	,005	29,814	,97	,96	,01	,00	,04			
5	1	5,562	1,000	,00	,00	,00	,00	,01	,00		
	2	,249	4,724	,00	,00	,00	,00	,60	,00		
	3	,134	6,433	,00	,00	,89	,03	,08	,00		
	4	,046	10,939	,00	,01	,02	,96	,26	,00		
	5	,007	28,847	,05	,98	,02	,01	,03	,06		
	6	,001	71,211	,94	,00	,06	,00	,03	,94		
6	1	6,217	1,000	,00	,00	,00	,00	,00	,00	,00	
	2	,433	3,790	,00	,00	,03	,00	,11	,00	,33	
	3	,195	5,651	,00	,00	,10	,00	,60	,00	,16	
	4	,106	7,645	,00	,01	,68	,10	,00	,00	,25	
	5	,042	12,185	,00	,02	,07	,90	,22	,00	,11	
	6	,006	32,657	,06	,96	,08	,00	,03	,08	,13	
	7	,001	76,133	,94	,01	,03	,00	,03	,91	,02	
7	1	6,929	1,000	,00	,00	,00	,00	,00	,00	,00	,00
	2	,654	3,254	,00	,00	,02	,00	,05	,00	,06	,05
	3	,198	5,912	,00	,00	,07	,00	,64	,00	,03	,01
	4	,108	8,015	,00	,01	,70	,10	,01	,00	,04	,01
	5	,064	10,431	,00	,00	,00	,01	,03	,00	,85	,75
	6	,041	13,024	,00	,03	,08	,89	,18	,00	,00	,06
	7	,005	35,526	,05	,94	,11	,00	,04	,10	,00	,06
	8	,001	83,160	,95	,03	,01	,00	,05	,90	,01	,07

a. Dependent Variable: quality

Figure 10. *Collinearity*

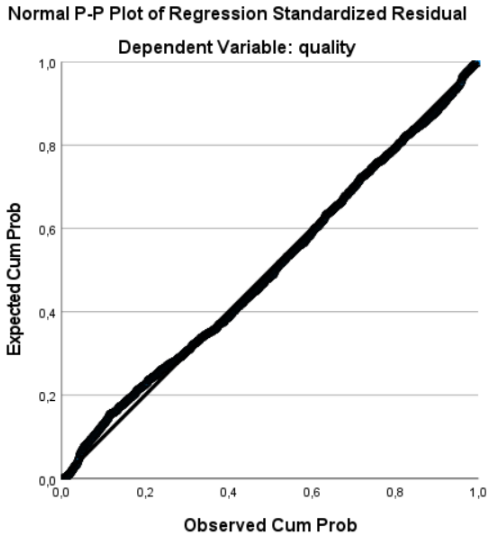


Figure 11. *Plot*

From the results we can say that both analysis match with each other, so the stepwise regression was successful.

5. REPORTING THE EFFECTS

In this section we report the effects of fixed acidity, volatile acidity, citric acid, residual sugar, alcohol, pH, and sulphates on the quality of wine.

When exploring the correlation analysis, we can say that the more there is alcohol in the wine, the better quality the wine is. Also, the higher the amount of citric acid, pH and sulphates have, the better the wine is. On the other hand, if the values of volatile acidity, fixed acidity and residual sugar are high the quality of the wine suffers.

The results of the stepwise regression are in line with the correlation analysis since SPSS removed citric acid, fixed acidity, and residual sugar to the comparison data. This is because they do not have much of a great impact on the wine. In this analysis the “places” between citric acid and volatile acidity are opposite compared to the correlation analysis but this can be explained by small data errors of different ways of calculating between these two analyses.

From the analysis we can say that the most important factors considering the quality of wine are alcohol, sulphates and pH. This is because all of the analyses performed value these variables and their amount seem to increase the quality of wine. On the other hand, the amounts of fixed acidity and residual sugar should be low for better wine quality since they correlate negatively with the quality. Citric acid and volatile acidity are quite neutral variables since their effects vary between different analyses.