Master 2 Financial Economics – Université Paris 1 Panthéon-Sorbonne

Strategic Asset Allocation
Scoring and Financial Macroeconometrics

Final project

Roland BOUILLOT
(Roland.Bouillot@etu.univ-paris1.fr)

Khalil JANBEK
(Khalil.Janbek@etu.univ-paris1.fr)

Mehdi LOUAFI
(Mehdi.Louafi@etu.univ-paris1.fr)

November 2021

TABLE OF CONTENTS

Question 1

What is the use of « label » in SAS and STATA?

The “label” command in SAS and STATA (and other statistical software) is used to assign an explanatory string to each variable, to give the user more information on the variables and a clearer understanding of what they are.

Question 2

Comment the tables of univariate statistics for the 15 variables of the database.

We firstly display the labels of each variable to understand, from the start of this application, what is the information they convey. In our sample, we sort the data by “ebita”.

Table 1

Labeled data

Contains data from **CleanDataEstimation.dta**

obs:	**91**
vars:	**15**
size:	**5,187**
	9 Nov 2021 16:35

variable name	storage type	display format	value label	variable label
yd	byte	%11.0g	ydl	**Indic Difficultés Financières**
tdta	float	%9.0g		**Dette/Actif**
reta	float	%9.0g		**Reserves et report ‡ nouveau/Actif**
opita	float	%9.0g		**Resultat/Actif**
ebita	float	%9.0g		**EBIT ou EBE/Actif**
lsls	float	%9.0g		**Log(ventes)**
lta	float	%9.0g		**Log(actif)**
gempl	float	%9.0g		**Croissance du nombre de salariés**
invsls	float	%9.0g		**Stocks/Ventes**
nwcta	float	%9.0g		**Capital circulant net/Actif**
cacl	float	%9.0g		**Actif circulant/Passif circulant**
qacl	float	%9.0g		**Disponibilites+VMP/Passif circulant**
fata	float	%9.0g		**Actifs Fixes/Actif**
ltdta	float	%9.0g		**Dette Long Terme/Actif**
mveltd	float	%9.0g		**Valeur de marche Actions/Dette long terme**

Sorted by: **yd ebita**

Now that we have in mind the information contained in each variable, we comment on their univariate statistics, in order to have a descriptive overview of each variable’s observations. The descriptive statistics are displayed in Table 2.1 and 2.2 below.

Table 2.1

Descriptive statistics

variable	N	mean	min	max	sd	variance	skewness	kurtosis
yd	91	.4725275	0	1	.5020106	.2520147	.1100564	1.012112
tdta	91	.5477133	.1455394	1.194813	.2161716	.0467302	.5027642	3.269872
reta	91	.2417644	-.6871598	.8352498	.2732161	.0746471	-.8536342	4.366948
opita	91	.1132649	-.3425781	.382154	.1165597	.0135862	-1.041327	6.07854
ebita	91	.0679348	-.427457	.310211	.1268709	.0160962	-1.361358	6.319612
lsls	91	5.670071	2.243384	11.37875	1.812881	3.286539	.5078379	3.193086
lta	91	5.273319	2.238314	11.01703	1.790933	3.20744	.5348511	3.02485
gempl	91	.0109068	-.2879793	.4564754	.1160379	.0134648	.7262452	5.727594
invsls	91	.1775606	.0299708	.485257	.0901908	.0081344	.7590834	4.170681
nwcta	91	.2863189	-.4294932	.7936063	.1906568	.03635	-.6777694	5.249776
cacl	91	2.371138	.478824	6.452848	1.187911	1.411132	1.409705	5.043047
qacl	91	1.33685	.136322	6.093447	.9650174	.9312585	2.290538	9.886408
fata	90	.195022	0	.9020172	.1597723	.0255272	1.345526	6.111276
ltdta	90	2.214532	.0741454	27.63183	4.323752	18.69483	4.738115	27.39406
mveltd	91	.3389339	.0272014	.6968414	.153292	.0234984	.3419301	2.463912

Table 2.2

Descriptive statistics

variable	N	mean	p5	p25	p50	p75	p95	iqr
yd	91	.4725275	0	0	0	1	1	1
tdta	91	.5477133	.2132748	.4113172	.5414321	.6435415	.9971208	.2322243
reta	91	.2417644	-.3502091	.0743997	.2812767	.4085758	.6287595	.3341761
opita	91	.1132649	-.0636567	.0619479	.1143884	.1852604	.280728	.1233125
ebita	91	.0679348	-.1705686	.0215174	.0841074	.1418997	.2490068	.1203823
lsls	91	5.670071	2.963405	4.328114	5.65417	6.753203	8.752879	2.425089
lta	91	5.273319	2.760355	3.827649	5.101276	6.614799	8.306961	2.78715
gempl	91	.0109068	-.1698624	-.0425729	.0112499	.0525194	.2055173	.0950923
invsls	91	.1775606	.0508379	.1080653	.1767557	.2379831	.3225927	.1299178
nwcta	91	.2863189	.0311957	.166754	.2956362	.4152768	.585332	.2485228
cacl	91	2.371138	1.082683	1.59745	2.082227	2.868885	5.027792	1.271436
qacl	91	1.33685	.2753191	.8212906	1.075795	1.604781	3.243222	.7834901
fata	90	.195022	.0002793	.0715278	.1762754	.2846393	.4394668	.2131115
ltdta	90	2.214532	.1358952	.463836	.9504595	2.20321	8.326773	1.739374
mveltd	91	.3389339	.1093956	.22982	.3185168	.4462523	.6085151	.2164323

The first thing we have to notice, is that we have 15 variables over 91 period observations (N). This gives us the depth of the sample and gives some information about the statistical relevance of the tests we will conduct ahead (a large number of observations N is preferred when performing statistical analysis). As most of the variables are ratios, the descriptive statistics of the variables range between 0 and 1, especially for the mean, the different percentile thresholds (5%, 25%, 50%, 75%, 95%), the minimum and maximum values, the standard deviation and the variance. Are also reported the skewness, the kurtosis and the interquartile range.

Question 3

Comment the histograms of total debt/total asset ($tdta$) for default versus healthy firms.

We plot the overall sample distribution in Figure 1. We notice that the overall distribution is centered around 0.5, which is consistent with the descriptive statistics presented in Table 2.1 and 2.2. The *kdensity* distribution represented by the blue line seems similar to the normal distribution represented by the red line, which is rather a good sign. We then plot the sample distribution sorted by the financial distress indicator yd in Figure 2. Again, we find some consistent results with regard to Table 3.1 and 3.2. The distribution for firms not in financial distress is skewed to the left, meaning that those firms have a low total debt/total assets ratio, which is consistent with its $mean = 0.44$ and $p50 = 0.47$. Conversely, the distribution of the firms experiencing financial distress is skewed to the right, implying that those firms do have a high total debt/total assets ratio, which is also consistent with its $mean = 0.66$ and $p50 = 0.60$.

Table 3.1

Descriptive statistics

Summary for variables: tdta
by categories of: yd (Indic Difficultés Financières)

yd	N	mean	min	max	sd	variance	skewness	kurtosis
No Fin.Dis.	48	.4421523	.1455394	.7997144	.1607716	.0258475	-.1104316	2.013993
Fin.Dist	43	.6655488	.2132748	1.194813	.2103823	.0442607	.490504	2.798976
Total	91	.5477133	.1455394	1.194813	.2161716	.0467302	.5027642	3.269872

Table 3.1

Descriptive statistics

Summary for variables: tdta
by categories of: yd (Indic Difficultés Financières)

yd	N	mean	p5	p25	p50	p75	p95	iqr
No Fin.Dis.	48	.4421523	.1969582	.29617	.4706952	.5597867	.6727151	.2636167
Fin.Dist	43	.6655488	.4113172	.4999897	.603791	.8245901	1.040991	.3246004
Total	91	.5477133	.2132748	.4113172	.5414321	.6435415	.9971208	.2322243

Figure 1

Plotted distribution of the Total debt over total assets variable

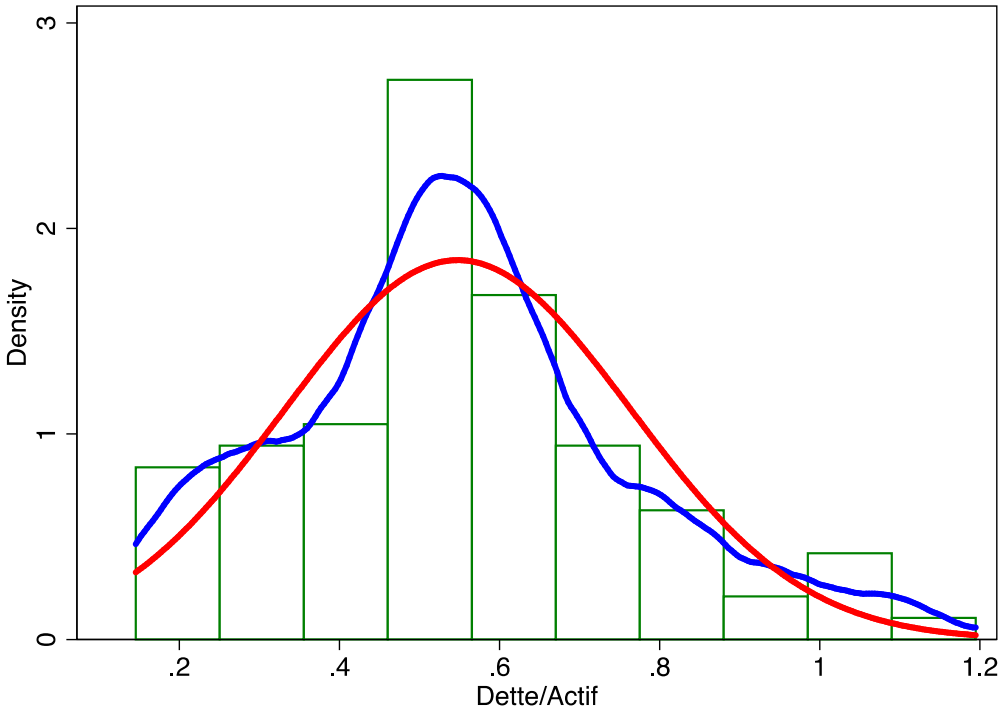
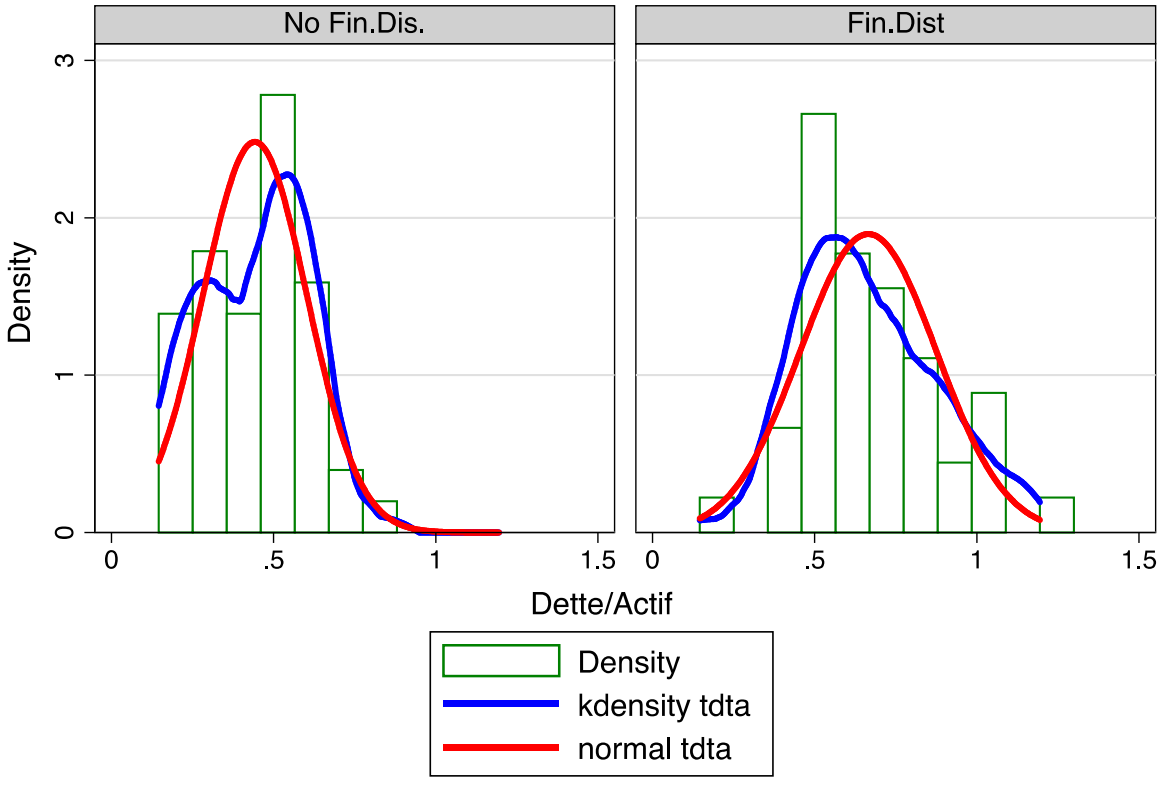


Figure 2

Plotted distribution of the Total debt over total assets variable sorted by financial distress



Question 4

What are the results of the normality tests for total debt/total assets for default versus healthy firms?

We first have a glance at the standardized normal probability plots (Figure 3) and at the inverse cumulative distribution functions (Figure 4). The idea behind those visual representations of the distributions is that the blue-plotted distribution of the series has to be as close as the black reference diagonal. We notice that the distribution functions are quite close to the reference diagonal, suggesting that the sorted sample is close to be normally distributed.

However, it is important to formalize those visual interpretations by performing some statistical tests. The first test we conduct is the skewness and kurtosis tests for normality and then we complement our normality analysis with a Shapiro-Wilk test. The results for the first tests are presented in Table 4 while results for the second test are displayed in Table 5. The results show that the distributions of the samples of the $tdta$ variable conditional of the state of yd (either 0 or 1) are normally distributed. Indeed, the skewness and kurtosis tests suggest that we should not reject the joint null hypothesis of normality as the $Prob > \mathrm{X}^2$ exceeds the 5% critical threshold for both $yd = 0$ and $yd = 1$. The results for the Shapiro-Wilk tests also suggest that we should not reject the null hypothesis of normality as the $Prob > z$ exceeds the 5% critical threshold for both $yd = 0$ and $yd = 1$.

Table 4

Skewness and Kurtosis tests for normality

			------ Joint ------		
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj_chi2(2)	Prob>chi2
tdta (yd=0)	48	0.727	0.030	4.810	0.090

Note: Not financial distressed firms

			------ Joint ------		
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj_chi2(2)	Prob>chi2
tdta (yd=1)	43	0.152	0.912	2.190	0.334

Note: Financial distressed firms

Table 5

Shapiro-Wilk normality tests

Variable	Obs	W	V	z	Prob>z
tdta (yd=0)	48	0.956	1.984	1.457	0.072

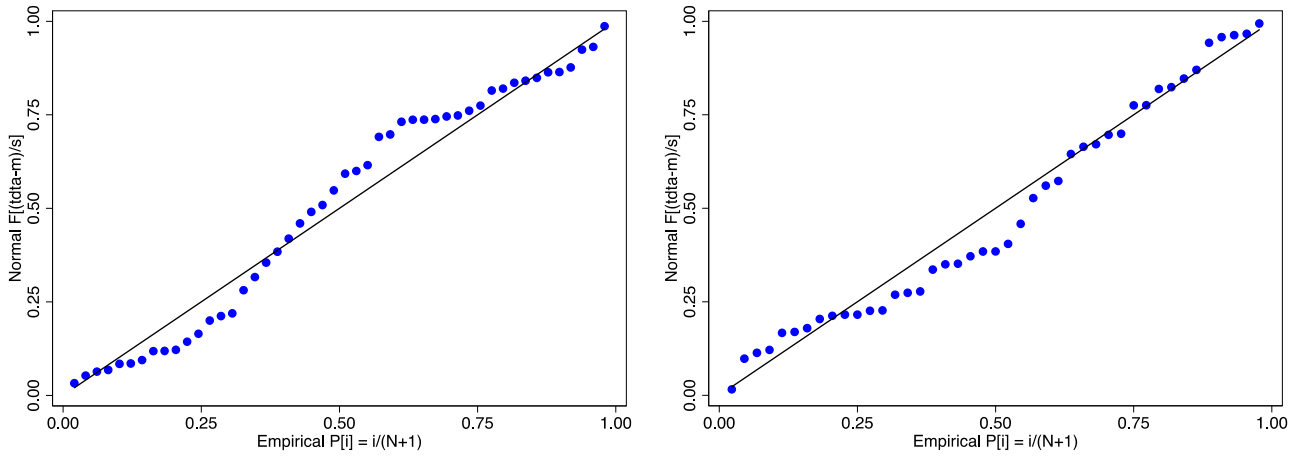
Note: Not financial distressed firms

Variable	Obs	W	V	z	Prob>z
tdta (yd=1)	43	0.964	1.515	0.878	0.190

Note: Financial distressed firms

Figure 3

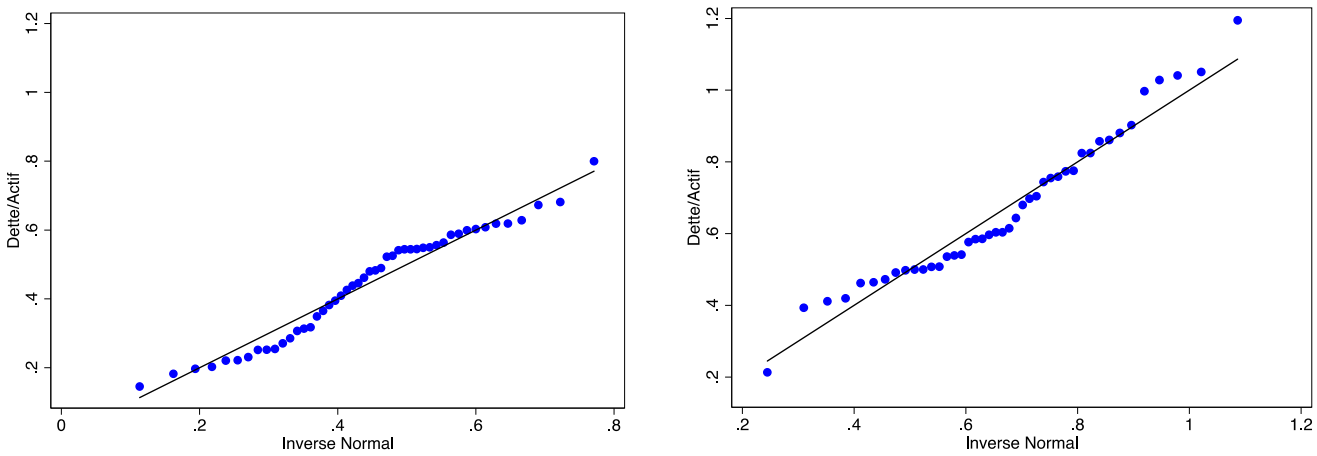
Standardized normal probability plot



Note: Not financial distressed firms (left graph), financial distressed firms (right graph)

Figure 4

Standardized normal probability plot



Note: Not financial distressed firms (left graph), financial distressed firms (right graph)

Question 5

What is the value of Students t-statistics related to the test of equality of means of $tdta$ between each groups? Check that it is the same for three other students test described in the course. Can you detail which are the four null hypothesis in each of these four cases?

In a first part, we present the results of the test of equality of means (for tdta). Then, in a second part, we present the three other student tests described in the course: (i) the analysis of variance (ANOVA), (ii) the linear probability model and the (iii) simple correlation test.

Table 6

Two-sample t-test of equality of means

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
No Fin.D	48	.4421523	.0232054	.1607716	.3954691	.4888355
Fin.Dist	43	.6655488	.032083	.2103823	.6008027	.7302949
combined	91	.5477133	.0226609	.2161716	.5026934	.5927332
diff		-.2233964	.0390218		-.3009319	-.145861

diff = mean(**No Fin.D**) - mean(**Fin.Dist**) t = **-5.7249**

Ho: diff = 0 degrees of freedom = **89**

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = **0.0000** Pr(|T| > |t|) = **0.0000** Pr(T > t) = **1.0000**

First, we perform a two-sample t-test of equality of means with the $tdta$ variable sorted by the yd indicator. The equality of means null hypothesis is tested against three alternative hypotheses:

$$H_0 : mean(tdta_{yd=0}) - mean(tdta_{yd=1}) = 0 \quad (1)$$

$$H_{A1} : diff < 0 \quad (2)$$

$$H_{A2} : diff \neq 0 \quad (3)$$

$$H_{A3} : diff > 0 \quad (4)$$

The results of the t-test are presented in Table 6. We find that alternative hypotheses H_{A1} and H_{A2} are not rejected but H_{A3} is strongly rejected. By construction, this implies that the null hypothesis is also rejected and that the means of both samples not only are different but more precisely that:

$$mean(tdta_{yd=0}) - mean(tdta_{yd=1}) < 0 \quad (5)$$

$$0.44 - 0.66 = -0.22 < 0 \quad (6)$$

Note that the $tstat = |5.72|$.

Secondly, we try to find this $tstat = -5.72$ value by other means. The first of three tests is the analysis of variance regression (ANOVA), which results are displayed in Table 7. The second test we run is a simple linear OLS regression, which results are displayed in Table 8. Finally, we conduct a pairwise correlation test between the dependent variable yd and the explanatory variable $tdta$. The results are shown in Table 9.

Table 7
Analysis of variance (ANOVA)
First and second stages

Number of obs = **91** R-squared = **0.2691**
Root MSE = **.185841** Adj R-squared = **0.2609**

Source	Partial SS	df	MS	F	Prob>F
Model	**1.1319332**	**1**	**1.1319332**	**32.77**	**0.0000**
yd	**1.1319332**	**1**	**1.1319332**	**32.77**	**0.0000**
Residual	**3.0737815**	**89**	**.03453687**		
Total	**4.2057147**	**90**	**.04673016**		

Note: first stage ANOVA

yd	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
tdta	1.205	.21	**5.72**	0	.787	1.623	***
Constant	-.187	.124	-1.51	.134	-.433	.059	
Mean dependent var		0.473	SD dependent var			0.502	
R-squared		0.269	Number of obs			91	
F-test		32.775	Prob > F			0.000	
Akaike crit. (AIC)		107.287	Bayesian crit. (BIC)			112.309	

*Note: second stage ANOVA, *** p<.01, ** p<.05, * p<.1*

Table 8
Linear OLS regression

yd	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
tdta	1.205	.21	**5.72**	0	.787	1.623	***
Constant	-.187	.124	-1.51	.134	-.433	.059	

Mean dependent var	0.473	SD dependent var	0.502
R-squared	0.269	Number of obs	91
F-test	32.775	Prob > F	0.000
Akaike crit. (AIC)	107.287	Bayesian crit. (BIC)	112.309

**** p<.01, ** p<.05, * p<.1*

Table 9
Pairwise correlation test

Variables	(1)	(2)
(1) yd	1.000	
(2) tdta	0.519***	1.000
	(0.000)	

**** p<0.01, ** p<0.05, * p<0.1*

In the analysis of variance test, we clearly identify the $tstat = |5.72|$ in the second stage of the ANOVA. In the simple OLS regression, $tstat = |5.72|$ is also clearly visible. Yet, in the pairwise correlation matrix, STATA does not display the t-stat used to compute the p-value of the correlation coefficient. However, we know that the t-stat is computed with the same formula as in the OLS regression:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{7}$$

where n is the number of observations (minus the degrees of freedom) and r the correlation coefficient. Thus, the correlation coefficient share the same the $tstat = |5.72|$ as the three other methods.

Question 6

Bivariate correlations with the dependent variable: Present a table sorting the 14 ratios by their correlation with the dependent variable (or equivalently by the t-statistics of the test of difference of means). Which are the explanatory variables most correlated with the dummy for default? Are there some of them with a simple correlation below 0.1 in absolute value?

We turn our focus to the correlations between the dependent variable *yd* and the 14 explanatory regressors. We firstly notice that 10 variables are statistically significant at the 5% level – only the variables "Inventory / Sales" ($Invsls$), "market value of equity / long term debt" ($mveltd$), the "log of sales" variable ($lsls$) and "log assets" (lta) are not significant at this level.

Most correlation signs appear in line with the economic intuition: for instance, the "Total debt / total assets" ($tdta$) variable coefficient is positive, meaning that the leverage positively contributes to default, similarly to the "Inventory to sales" variable ($invsls$), as rising inventory imply less cash inflows. Likewise, variables linked to operating performance or working capital logically have a negative sign. However, the positive sign in front of the coefficient for the "Log of assets" (lta) or the "Log of sales" ($lsls$) seems quite surprising, as there is no economic rationale that could explain why higher assets or higher sales would positively contribute to the firm's default. Such an odd result could certainly be attributed to some noise in our estimation sample.

The "Total debt / total assets" ($tdta$) variable is the regressor with the highest correlation coefficient with the "Financial distress" variable (yd), displaying a correlation coefficient of 0.5188. The variables with the most important correlations with the dummy for default, in absolute value, are then "retained earnings" ($reta$), "operating income / total assets" ($opita$), "EBIT / total assets" (ebita), "Employee growth" ($gempl$), "Net working capital / Total assets" ($nwcta$), "Current assets / current liabilities" ($cacl$), "Fixed assets" / Total assets" ($fata$) and "Quick assets / current liabilities" ($qacl$).

The remaining regressors have correlation coefficients, in absolute value, close to 1 – such as the "log of sales" variable ($lsls$) or the "log of assets (lta) – or below 1 – as in the case of "Market value of equity / Long term debt" ($mveltd$). We can therefore reasonably conclude that *mveltd* can be excluded from our analysis, due to its low impact on our variable of interest yd. Moreover, given the relatively low correlation of *lsls* and *lta* with yd as well as the impossibility to give a meaningful interpretation to their correlation coefficient, we can reasonably assume that we can leave those two variables out of our analysis.

Table 10

Bivariate correlations with the default dummy variable (yd)

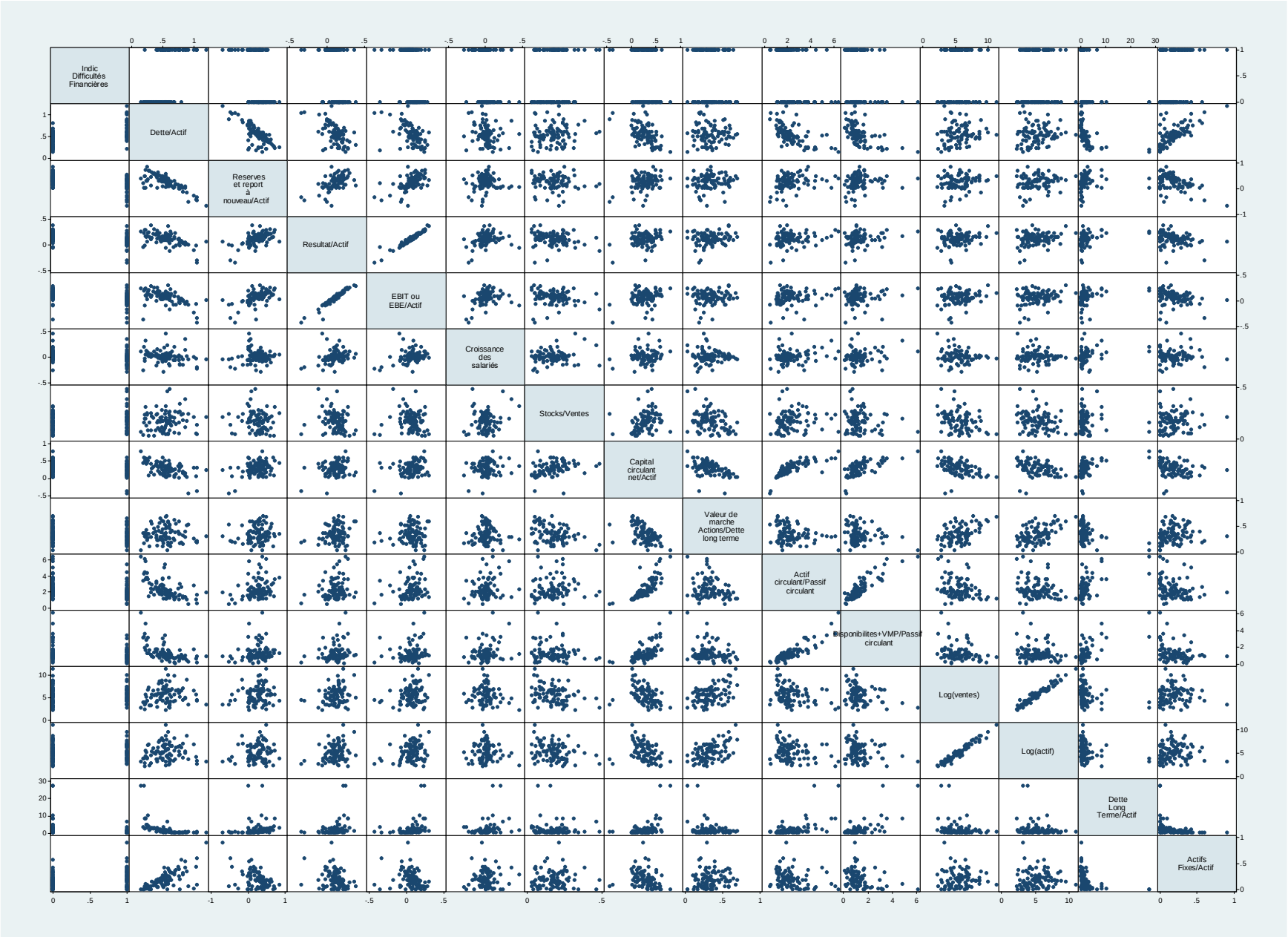
	yd	tdta	reta	opita	ebita	gempl	invsls
yd	1.0000						
tdta	0.5188*	1.0000					
	0.0000						
reta	-0.4515*	-0.7657*	1.0000				
	0.0000	0.0000					
opita	-0.4031*	-0.5721*	0.5721*	1.0000			
	0.0001	0.0000	0.0000				
ebita	-0.3635*	-0.5224*	0.5606*	0.9634*	1.0000		
	0.0004	0.0000	0.0000	0.0000			
gempl	-0.2766*	-0.1984	0.0465	0.2713*	0.3148*	1.0000	
	0.0080	0.0594	0.6616	0.0093	0.0024		
invsls	0.1149	0.0927	-0.0960	-0.1899	-0.1012	0.1717	1.0000
	0.2780	0.3821	0.3656	0.0714	0.3400	0.1037	
nwcta	-0.2658*	-0.5735*	0.3543*	0.2479*	0.2792*	0.1546	0.3209*
	0.0109	0.0000	0.0006	0.0178	0.0074	0.1434	0.0019
mveltd	-0.0109	-0.0071	0.1353	0.1837	0.0856	-0.0309	-0.3644*
	0.9182	0.9464	0.2010	0.0813	0.4195	0.7713	0.0004
cacl	-0.2823*	-0.6029*	0.2879*	0.2121*	0.2061*	0.2069*	0.1156
	0.0067	0.0000	0.0057	0.0435	0.0500	0.0491	0.2751
qacl	-0.2269*	-0.5592*	0.1664	0.2163*	0.1750	0.1978	-0.1438
	0.0305	0.0000	0.1148	0.0394	0.0971	0.0602	0.1739
lsls	0.0130	0.1222	0.1869	0.2334*	0.2059	-0.1133	-0.3373*
	0.9027	0.2485	0.0761	0.0260	0.0503	0.2849	0.0011
lta	0.0158	0.0840	0.1773	0.2488*	0.2141*	-0.0446	-0.2607*
	0.8817	0.4287	0.0927	0.0174	0.0415	0.6745	0.0126
ltdta	-0.3032*	-0.5120*	0.1195	0.3027*	0.3023*	0.2860*	-0.0630
	0.0037	0.0000	0.2618	0.0037	0.0038	0.0063	0.5552
fata	0.2883*	0.6676*	-0.4904*	-0.2898*	-0.2440*	0.0500	0.0352
	0.0059	0.0000	0.0000	0.0056	0.0205	0.6399	0.7419

Note: "" indicates statistical significance at the 5% level.*

We supplement the above results with a cloud point representation, to get a visual insight of bivariate correlations between our "Financial distress" variable *yd* and its 14 explanatory regressors, into which the above table gave us a first glimpse.

Figure 5

Bivariate correlations for all variables



Question 7

Compare and comment briefly the box plots of each of the two groups of firms for the 14 financial ratios. The larger the difference of means, the more discriminant is the ratio.

In line with our expectation, the median of "Total debt / Total assets" (Dette/Actif) for financially distressed firms is markedly higher than for non-financially distressed firms. This significant different in medians highlights the importance of this ratio, and its highly discriminant power. The dispersion of financially distressed firms within the "Total debt / total assets ratio" is also significantly greater than for non-financially distressed firms.

Likewise, non-financially distressed firms have markedly higher "Retained earnings / Assets" ratios, which makes this ratio powerfully discriminant as well. This is an expected result, as firms with higher retained earnings have more cash to face their financial obligations.

The "Operating income / Assets" (Résultat / Actif) ratio, "EBIT / Assets" (EBIT / Actifs) ratio and "Employee growth" variables are slightly higher for non-financially distressed firms. While this is expected, the difference in the medians of those ratios between the two groups is not as strong as for "Total debt / Total assets" or "Retained earnings / Assets".

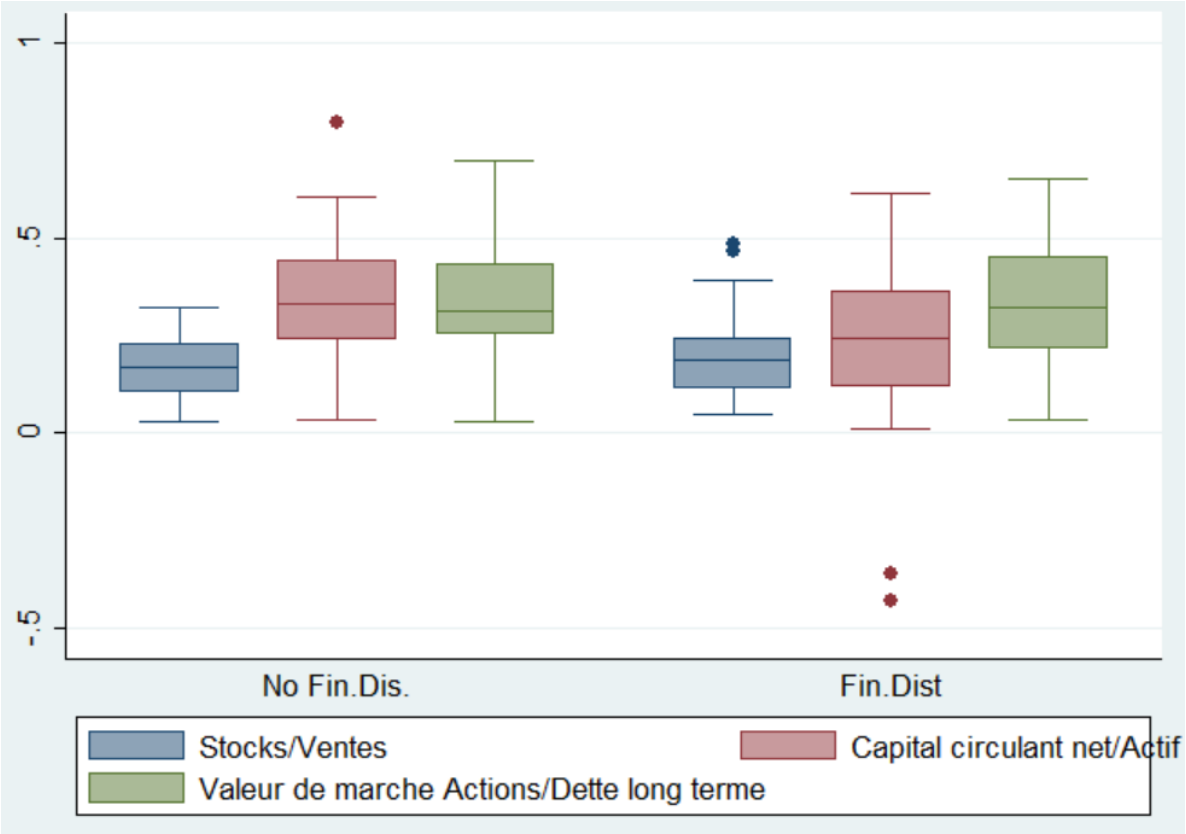
We notice that there are more outliers for those ratios (except for "Total debt / Total assets") for financially distressed firms. Those are represented by the point observations that exceed the quartile 1 and 4 limits (i.e. that exceed the upper and lower whiskers). The sample of non-financially distressed firms thus appears slightly more homogeneous in the light of those metrics.

Figure 6

Boxplot of selected explanatory variables sorted by the financial distress indicator

Figure 7

Boxplot of selected explanatory variables sorted by the financial distress indicator

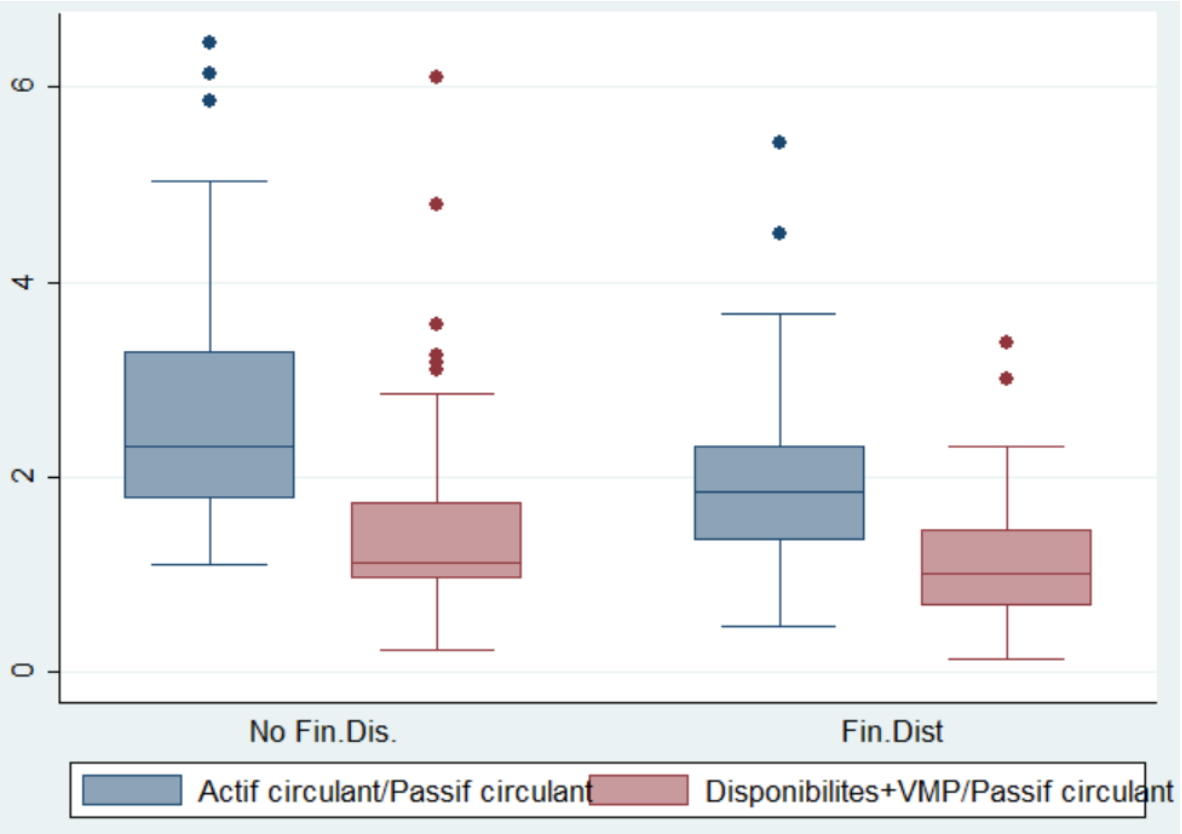


The discriminant power of the inventory/sales ratio (Stocks/Ventes) is not as stark as that of the ratios of the previous chart, as the median for financially distressed and non-financially distressed firms is very close. The distribution of firms for this ratio is actually quite similar between the two groups, judging by the size of the "box" (i.e. the upper and lower quartile range).

The median for the "Net working capital / total assets" ratio (Capital circulant net / Actif) is higher for non-financially distressed firms, and the distribution is markedly less dispersed than for financially distressed firms – as the size of the "box" is larger for financially distressed firms. The same applies for the "Market value of equity" ratio. However, the medians for this ratio appear to be nearly the same between the two groups.

Figure 8

Boxplot of selected explanatory variables sorted by the financial distress indicator

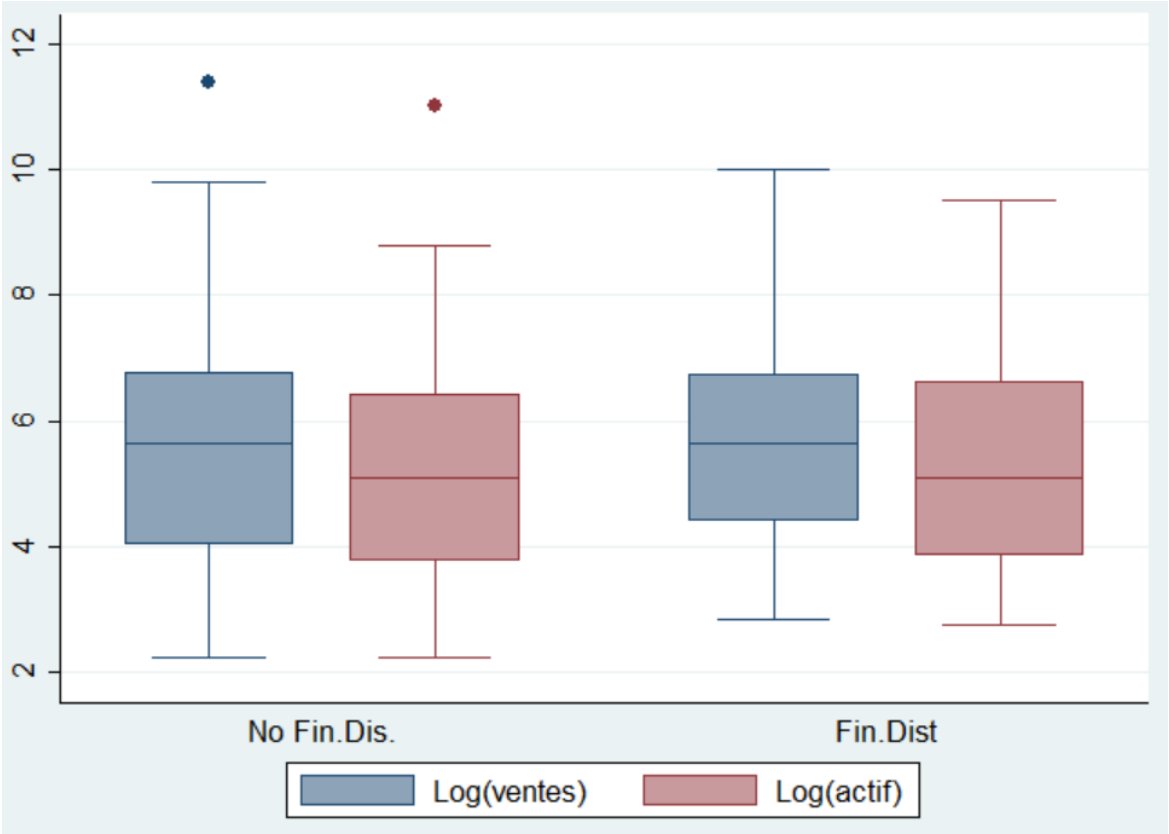


The "Current assets / current liabilities" ratio is markedly higher for non-financially distressed firms than for financially distressed firms. The median is not only notably higher, but also the upper quartile and upper whiskers are significantly higher for non-financially distressed firms. This shows how powerful this ratio is for discriminating between the two groups of observations.

The distribution of observations for the ratio of "Quick assets / current liabilities" is on the other hand quite similar between the two groups. The difference lies in an expectedly higher upper quartile and higher upper whisker for non-financially distressed firms, and in more outliers – beyond the 4th quartile – on the non-financially distressed firms' side.

Figure 9

Boxplot of selected explanatory variables sorted by the financial distress indicator

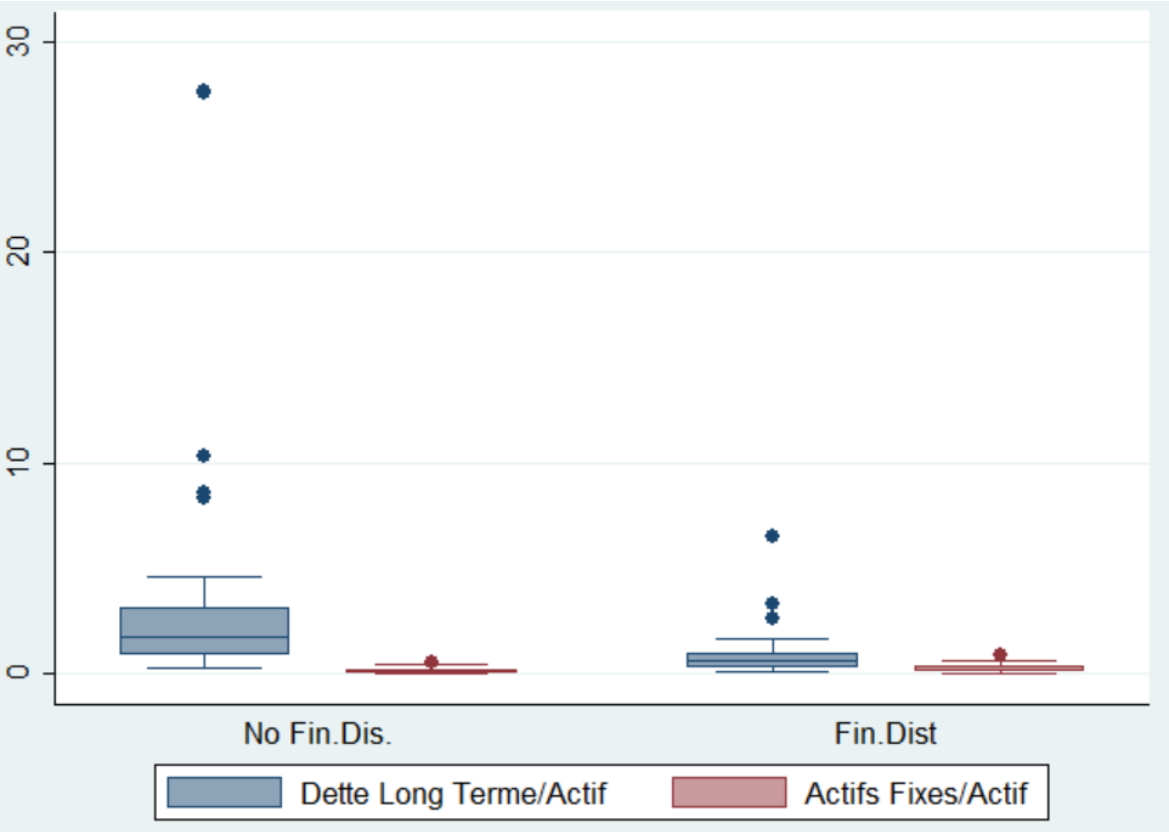


The log of assets and log of sales ratios are quite similar across the two groups, and therefore have little discriminant power – as we can expect, given the low correlation coefficient of those ratios with the "Default" dummy variable. As we noted earlier when noticing a positive correlation between those ratios and the "Default" dummy variable, the lower whiskers for both ratios are surprisingly higher for financially distressed firms.

This result confirms that those ratios should not be taken into account in the analysis, given their low discriminant power and not meaningful economic interpretation.

Figure 10

Boxplot of selected explanatory variables sorted by the financial distress indicator



The result of this chart is quite surprising as well, as we notice that the median of the Long-term debt / Assets ratio is higher for non-financially distressed firms, as well as the upper quartile and upper whisker. Non-financially distressed firms also count many more outliers, exceeding the upper whisker, for this ratio. This result comes as a surprise, as we would have expected financially distressed firms to have more long-term debt. However, the difference between the two groups remains tenuous.

This latter remark holds true for the "Fixed assets / total assets" ratio, as nearly no difference between the two groups can be noticed for this ratio.

In conclusion, the variables that matter the most in driving corporate default, according to this section, are the "Total / Assets" ratio, "Operating income / Total assets" ratio, "EBIT / Total assets" ratio, "Retained Earnings / Total assets" ratio and the "Current Assets / Current Liabilities" ratio.

Question 8

Bivariate correlation between explanatory variables (regressors): Are there explanatory variables with simple correlation coefficient with other explanatory variables larger than 0.8? Is it expected and explained by accounting properties? Will you include all of them in your regressions?

Our concern in this section is to enquire as to potential multicollinearity among our regressors. Multicollinearity could lead to unstable estimates of our model, as it would increase the variance of our estimation of the parameter and lead to incorrect hypothesis tests. We therefore turn our focus this time to bivariate correlations among the independent variables.

The results are displayed in Table 11. We firstly notice that many of the bivariate correlations among regressors are statistically significant at the 5% level. In line with our expectations, we notice some very strong simple correlation coefficients (ranging from 0.76 to 0.97). In particular:

- the "Retained earnings / Total assets" ratio has a correlation coefficient of -0.76 with the "Total debt / Total assets" ratio, very close to -0.8 (significant at the 5% level). This result makes sense economically: as retained earnings usually make up a significant part of the equity section of the balance sheet, and as assets can either be financed through debt or through equity, it is logical that a lower debt to assets ratio mechanically implies higher equity/assets ratio.
- The "EBIT / Total assets" ratio has a simple correlation coefficient ratio of 0.96 (significant at the 5% level) with "Operating Income / Total assets". This result is logical, as EBIT is an operating performance metric that includes operating income, along with income from non-operating activities from the company (i.e. from any activity which is not in the company's core activities).
- The "Current assets / current liabilities" ($cacl$) ratio has a simple correlation coefficient ratio of 0.77 (significant at the 5% level) with the "Net working capital / Total assets" ($nwcta$) ratio. Their very strong link is logical, as the former is the ratio of current assets and current liabilities while the latter is simply the difference between those two metrics.
- The "Quick assets / current liabilities" ($qacl$) ratio has a simple correlation coefficient ratio of 0.88 (significant at the 5% level) with the "Current assets / current liabilities" ($cacl$) ratio. Once again this result makes perfect sense from an accounting point of view, as quick assets – which comprise cash, securities that the firm can sell and accounts receivables – are simply equal to current assets, to which we remove inventories.
- The "Log Assets" (lta) ratio has a simple correlation coefficient ratio of 0.97 (significant at the 5% level) with the "Log Sales" ($lsls$) ratio. This result is more difficult to verify from an accounting perspective. From an economic point of view, we could argue that higher assets facilitate sales. However, this near perfect correlation between those two variables would more probably be explained by the log-specification of both variables.

In addition to the presented correlation coefficients, we notice many other statistically significant bivariate correlations at the 5% level, with coefficients ranging from 0.3 to 0.6 in absolute value. We supplement the above results with a cloud point representation to get a visual insight of bivariate correlations among our 14 regressors, which Table 11 gives us a first glimpse.

Table 11

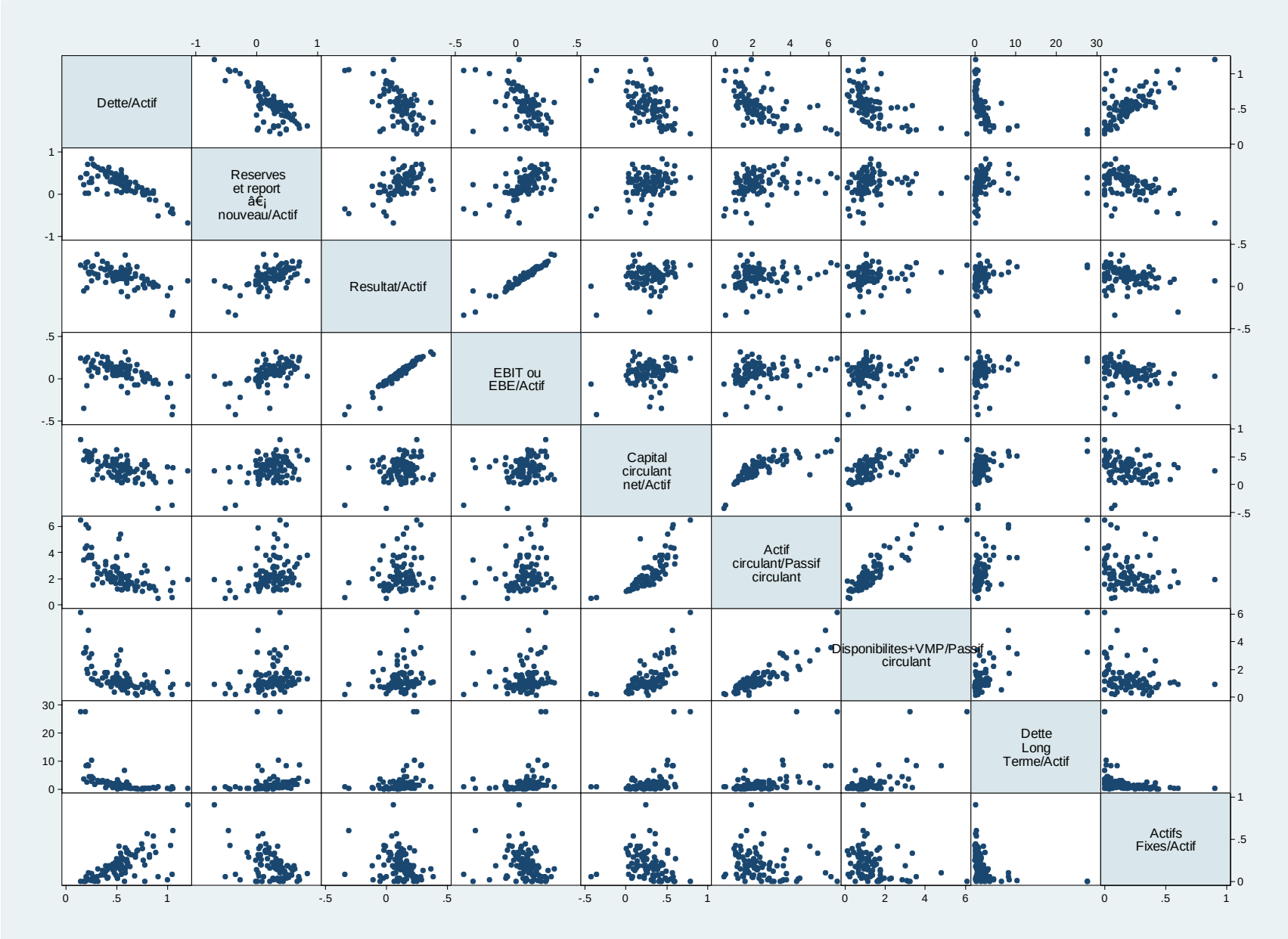
Bivariate correlations among the regressors

	tdta	reta	opita	ebita	gempl	invsls	nwcta	mveltd	cacl	qacl	lsls	lta	ltdta	fata
tdta	1.0000													
reta	-0.7657*	1.0000												
	0.0000													
opita	-0.5721*	0.5721*	1.0000											
	0.0000	0.0000												
ebita	-0.5224*	0.5606*	0.9634*	1.0000										
	0.0000	0.0000	0.0000											
gempl	-0.1984	0.0465	0.2713*	0.3148*	1.0000									
	0.0594	0.6616	0.0093	0.0024										
invsls	0.0927	-0.0960	-0.1899	-0.1012	0.1717	1.0000								
	0.3821	0.3656	0.0714	0.3400	0.1037									
nwcta	-0.5735*	0.3543*	0.2479*	0.2792*	0.1546	0.3209*	1.0000							
	0.0000	0.0006	0.0178	0.0074	0.1434	0.0019								
mveltd	-0.0071	0.1353	0.1837	0.0856	-0.0309	-0.3644*	-0.5074*	1.0000						
	0.9464	0.2010	0.0813	0.4195	0.7713	0.0004	0.0000							
cacl	-0.6029*	0.2879*	0.2121*	0.2061*	0.2069*	0.1156	0.7774*	-0.2624*	1.0000					
	0.0000	0.0057	0.0435	0.0500	0.0491	0.2751	0.0000	0.0120						
qacl	-0.5592*	0.1664	0.2163*	0.1750	0.1978	-0.1438	0.6387*	-0.2209*	0.8824*	1.0000				
	0.0000	0.1148	0.0394	0.0971	0.0602	0.1739	0.0000	0.0354	0.0000					
lsls	0.1222	0.1869	0.2334*	0.2059	-0.1133	-0.3373*	-0.3631*	0.4652*	-0.3113*	-0.2763*	1.0000			
	0.2485	0.0761	0.0260	0.0503	0.2849	0.0011	0.0004	0.0000	0.0027	0.0080				
lta	0.0840	0.1773	0.2488*	0.2141*	-0.0446	-0.2607*	-0.3221*	0.4917*	-0.2321*	-0.1711	0.9706*	1.0000		
	0.4287	0.0927	0.0174	0.0415	0.6745	0.0126	0.0018	0.0000	0.0268	0.1048	0.0000			
ltdta	-0.5120*	0.1195	0.3027*	0.3023*	0.2860*	-0.0630	0.4528*	-0.2723*	0.5652*	0.6804*	-0.2490*	-0.1934	1.0000	
	0.0000	0.2618	0.0037	0.0038	0.0063	0.5552	0.0000	0.0094	0.0000	0.0000	0.0179	0.0678		
fata	0.6676*	-0.4904*	-0.2898*	-0.2440*	0.0500	0.0352	-0.2447*	0.1049	-0.2470*	-0.2942*	0.0149	0.0074	-0.3685*	1.0000
	0.0000	0.0000	0.0056	0.0205	0.6399	0.7419	0.0201	0.3253	0.0189	0.0049	0.8895	0.9450	0.0004	

"" indicates statistical significance at the 5% level.*

Figure 11

Bivariate correlations among the regressors



Question 9

Comment the table of the bivariate clouds of points for six variables: yd, total debt and its highly correlated companion ratio, operating income / total assets and its highly correlated companion ratio, and growth of employees. Does the shape of the clouds matches with the simple correlation coefficients?

We turn our attention in this section to the dependent variable yd and a selected sample of regressors, who happen to be correlated among themselves.

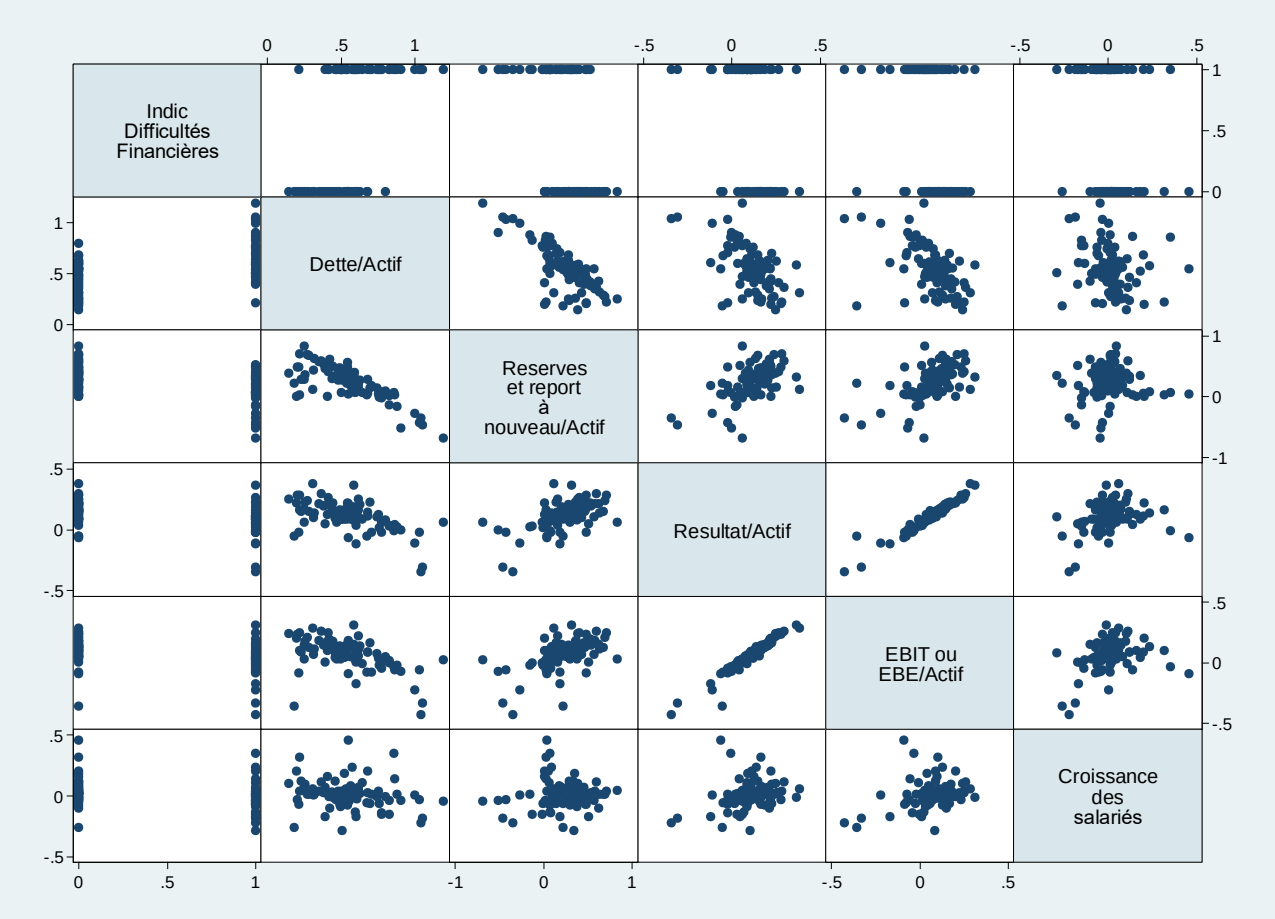
We notice that the shapes of clouds match with the simple correlation coefficients of the Table 11. We can for example easily notice the near-perfect correlation of 0.96 between $opita$ (Résultat / actif) and $ebita$ (EBIT / actif) in the below plot, and the very strong negative relationship (-0.76) between $tdta$ (Dette / actif) and $reta$ (Réserves / actif).

The plot also confirms the relatively weak correlation coefficients of the employee growth variable $gempl$, as the cloud points are relatively dispersed. This visual result matches the fact that $gempl$ is the variable, in Table 12, for which the p-value is the highest. Those results make perfect sense from an accounting point of view, as we developed in Question 8.

Finally, this plot confirms the result of Table 12 that all those regressors are negatively correlated with the "default" dummy variable yd, except for $tdta$ – for the very logical reason developed in question 6, namely that higher leverage likely implies higher probability of default.

Figure 12

Bivariate cloud of points for selected explanatory variables



We include below the table of simple correlation coefficients for reference.

Table 12

Bivariate correlations between yd and a chosen set of regressors

	yd	tdta	reta	opita	ebita	gempl
yd	1.0000					
	91					
tdta	0.5188*	1.0000				
	0.0000					
	91	91				
reta	-0.4515*	-0.7657*	1.0000			
	0.0000	0.0000				
	91	91	91			
opita	-0.4031*	-0.5721*	0.5721*	1.0000		
	0.0001	0.0000	0.0000			
	91	91	91	91		
ebita	-0.3635*	-0.5224*	0.5606*	0.9634*	1.0000	
	0.0004	0.0000	0.0000	0.0000		
	91	91	91	91	91	
gempl	-0.2766*	-0.1984	0.0465	0.2713*	0.3148*	1.0000
	0.0080	0.0594	0.6616	0.0093	0.0024	
	91	91	91	91	91	91

Question 10

Comment the results of the linear probability model (LPM) with only total debt/total assets ($tdta$) as an explanatory variable.

We perform a first regression, in the form of a univariate linear probability model (LPM). The only explanatory variable included is the "Total debt / Total Assets" ratio ($tdta$).

A linear probability model is a binary OLS-estimated regression model, where (i) the dependent variable is a binary variable, i.e. which has only two outcomes and (ii) one or more explanatory variables are used to predict this outcome. The aim of such a model is to model the probability of a binary outcome ($yd = 0$ or 1) given the explanatory variable ($tdta$).

This a particular case of a binary regression models, where coefficients can be interpreted as the change in the probability that the dependent variable takes on the value of 1.

Analyzing at the results of the linear probability model in the Table 13, we notice firstly notice that the coefficient for $tdta$ (of 1.20) has a positive sign, which matches the economic intuition: a 1% percentage point increase in the "Total debt / total assets" ratio increases the probability of default by 1.2% holding other coefficients – the intercept in this case – constant.

Table 13

Linear Probability Model (LPM) results

Source	SS	df	MS		
Model	6.10448886	1	6.10448886	Number of obs =	91
Residual	16.5768298	89	.186256515	F(1, 89) =	32.77
				Prob > F =	0.0000
Total	22.6813187	90	.252014652	R-squared =	0.2691
				Adj R-squared =	0.2609
				Root MSE =	.43157

yd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tdta	1.204772	.2104437	5.72	0.000	.7866246	1.622919
_cons	-.187342	.1238236	-1.51	0.134	-.4333768	.0586929

We also notice that the coefficient for $tdta$ has a t-statistic of 5.72, exceeding the critical value at the 1% significance level – and similarly displays a p-value of 0 – indicating that this variable is very significant. The coefficient for the constant however appears not to be significant.

The R^2 of the model seems rather high, for a model that includes only one regressor (26.9%). However, we should not give too much importance to the R^2 which cannot be meaningfully interpretated: as the dependent variable yd is binary and the variable $tdta$ is continuous, it is impossible for the regression line to precisely fit the data (as shown by the below plot). Therefore, this model calculates the predicted probability of default as:

$$yd = -0.187 + 1.20 \times tdta \tag{8}$$

Figure 13

Scatterplot of fitted vs *observed "Financial distress" data*

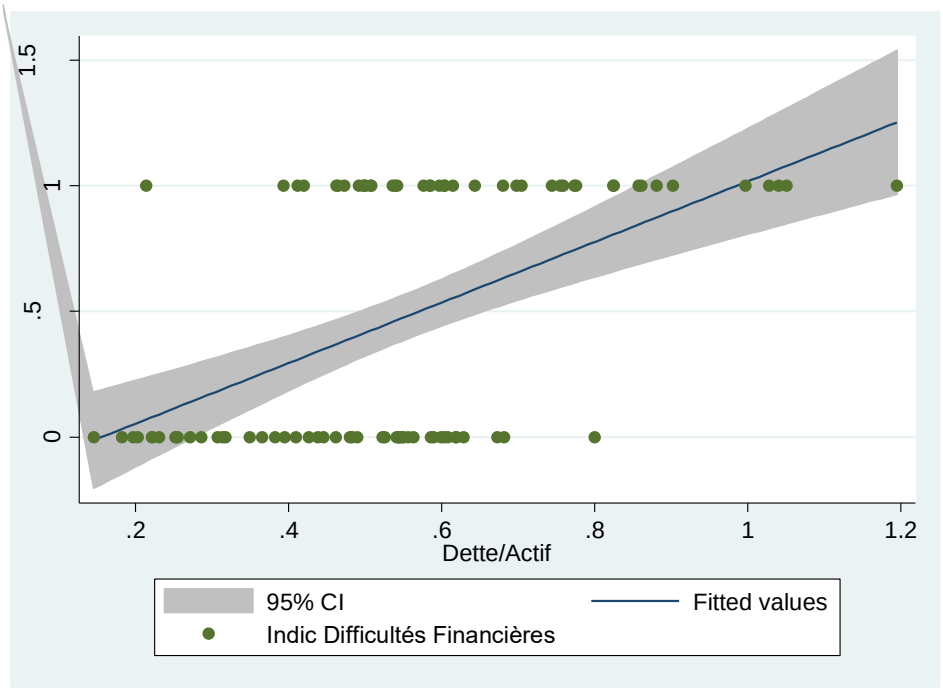


Figure 13 visually confirms the positive relationship between $tdta$ and yd, and confirms the impossibility for the regression line to exactly fit the observed data.

Question 11

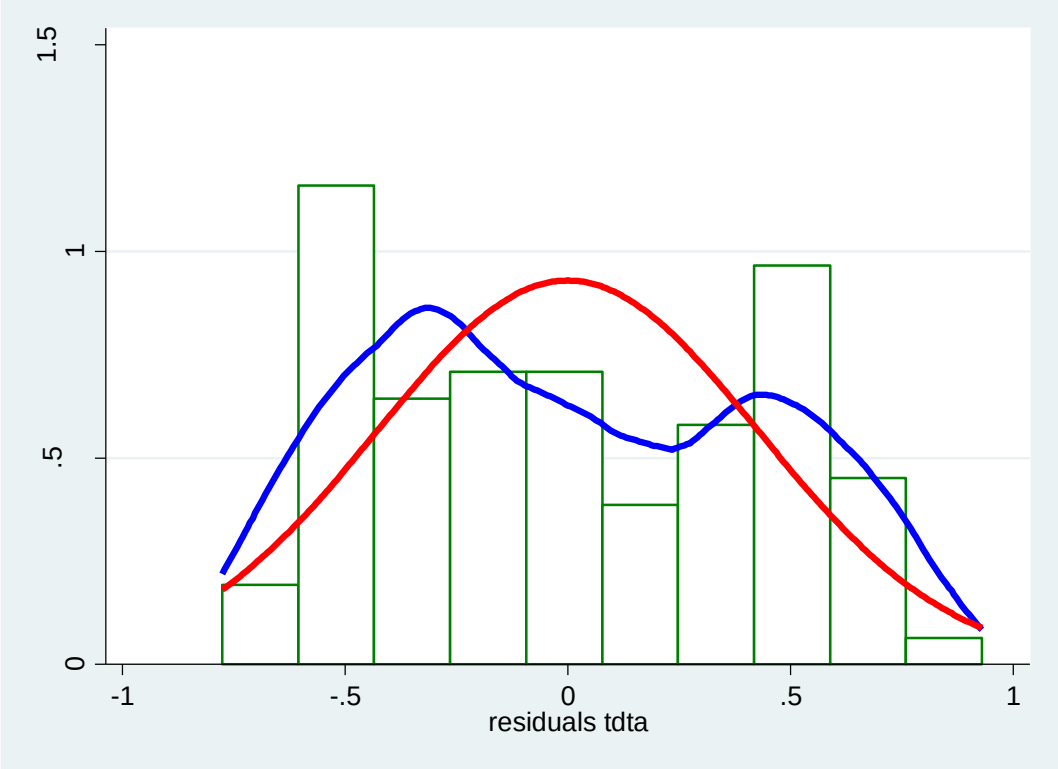
Comment on the plot of the distribution of residuals of the linear probability model. Does it correspond to normal law?

Having estimated the linear probability model, we analyze its residuals to detect any evidence of non-normality, which would falsify hypothesis tests for our estimated parameters.

Figure 14 plots the distribution of the model's residuals (the blue curve) and compares it to the normal distribution (the red curve). It is quite stark that the distribution of residuals is not a symmetrical bell shape centered around 0, as the red curve.

Figure 14

Linear probability model residuals distribution



We verify this visual insight – that the distribution of our model's residuals is not normal – with the following skewness test:

Table 14

Linear probability model's residuals skewness and kurtosis tests

Skewness/Kurtosis tests for Normality

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
etdta	91	0.3856	0.0000	25.82	0.0000

This test firstly indicates that we can reject the hypothesis that the model's residuals are normally distributed[1].

[1] Our interpretation relies on the STATA documentation for this test

It firstly indicates that we cannot reject that the distribution of our model's residuals has a skewness of 0 – as it is the case for normal distributions – given the p-value of 0.3856: the distribution's skewness is not significantly different to that of a normal distribution. However, the model rejects that the distribution has a kurtosis of 3 – as in the case of normal distributions – as the p-value for this test parameter is 0.00: the kurtosis of our distribution is therefore significantly different to that of a normal distribution, which enables us to conclude that this test rejects that our model's residuals are normally distributed.

Calling the summary statistics function in STATA can help us verify that those distribution parameters are not equal to those of a normal distribution:

Table 15

Summary statistics of the linear probability model's residuals

variable	N	mean	p50	skewness	kurtosis
etdta	91	1.60e-09	-.0785999	.2100312	1.769024

We finally confirm those results with a Shapiro-Wilk test for normality.

Table 16

Shapiro-Wilk test for normality

Shapiro-Wilk W test for normal data

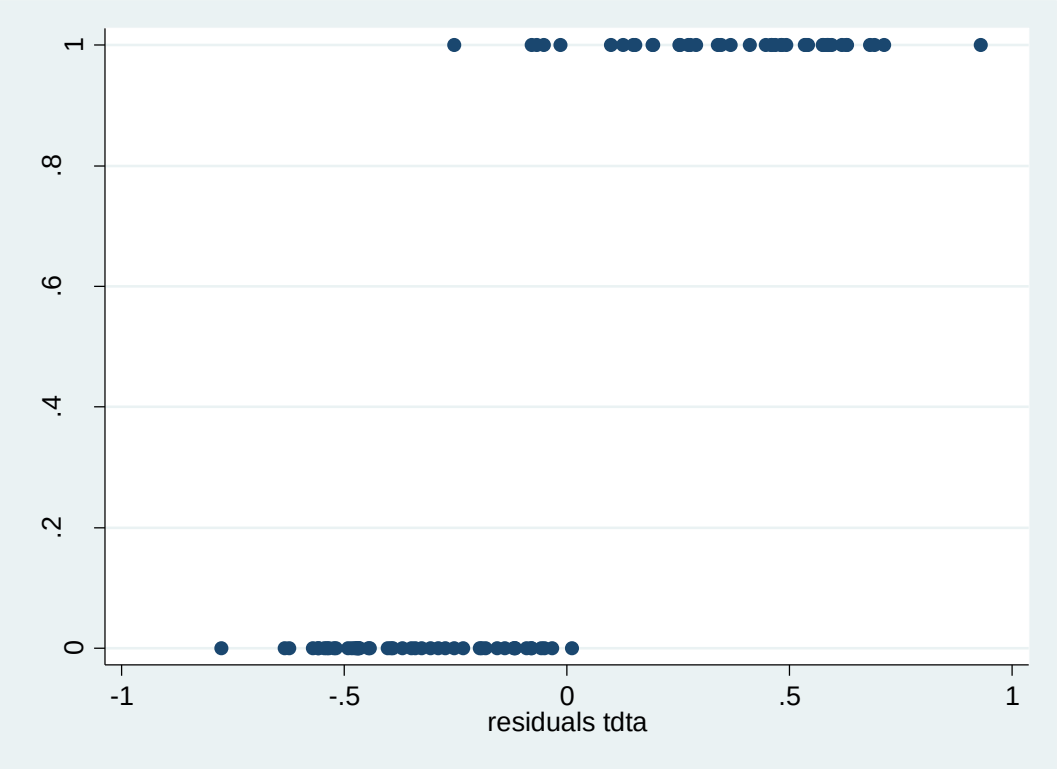
Variable	Obs	W	V	z	Prob>z
etdta	91	0.94035	4.553	3.345	0.00041

This Shapiro-Wilk test output indicates that the model's residuals are not normally distributed, as shown by the very high value of the W test-statistic. This test-statistic is statistically significant, as proven by the very low p-value.

In conclusion, we have demonstrated that the residuals of our linear probability model are not normally distributed. This is in fact expected, given the binary nature of the independent variable: the error term will consequently have two possible values as well, for a given value of the regressor. It is consequently impossible for the residuals to be normally distributed. Rather, its residuals are more likely to follow a binomial distribution. This is confirmed by the below plot of residuals.

Figure 15

Linear probability model residuals distribution



Question 12

Comment the results of the logit model with only total debt/total assets as an explanatory variable. Comment on the plot of the distribution of residuals of the logit model. Does it fit a normal law?

Now that our study has explored the linear regression models, we pivot towards binomial models, starting with the logit regression model.

The logit model we use is fitted to predict the probability of firms to default given the impact of the firm's total debt/total assets ratio. To do so, our dependent variable is the binary firms' distress indicator yd and the explanatory total debt over total assets $tdta$ variable. The probability of defaulting is p. The model can then be expressed as:

$$log\frac{p}{1-p} = \beta_0 + \beta_1 \times tdta + \varepsilon \quad (9)$$

Table 16 displays the results of the logit regression. Replacing the newly estimated betas in equation 9 yields:

$$log\frac{p}{1-p} = -3.93 + 7.00 \times tdta + \varepsilon \quad (10)$$

The **estimated coefficients of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$** are the values for the constant and the $tdta$ variable. We are mainly interested in the value of β_1 that is associated with the explanatory variable $tdta$. With $\beta_1 = 7.00$ the estimated coefficient suggests that the relationship between the probability of a firm defaulting and an increase of the firm's total debt/total assets ratio is positive, which concordant with the expected underlying economic theory (the more a firm has debt and/or less assets, the higher the probability of being in a financial distress situation). More precisely, finding a $\beta_1 = 7.00$ means that for a one unit increase of $tdta$, the log-odds to default increase by 7. Although the estimated value of the constant isn't as useful as the β_1 estimated parameter to understand the relationship between our selected variables, note that $\beta_0 = -3.93$. However, those results would mean little to nothing if their associated **p-value** (z-value to be exact) were not statistically significant, meaning that estimated parameters are inaccurate. In the regressed logit model, both the value corresponding to the constant (β_0) and the $tdta$ variable (β_1) are statistically significant at the 1% threshold. Indeed, their associated z-value is equal to $|4.16|$, which is deemed statistically different from 0 given the 1% critical value. Finally, the **95% confidence interval** helps to grasp the range within which the estimated parameter evolves.

Now that we have commented the upper section of the table, let us dive into the lower one. In addition to this lower section, Table 17 brings a more detailed analysis of the measures of fit of the regressed model.

Our **number of observations** is, as expected, 91. One measure of the goodness of fit of the model is the **$\mathbf{X^2}$ test statistic**. Here, taking into account one degree of freedom (as we only have one predictor), our $X^2 = 29.03$. We test our X^2 t-stat against a null hypothesis that is the probability of obtaining this same X^2 if there is no effect of the independent variables on the dependent variable. The associated p-value compared to a critical threshold determines if the overall model is statistically significant. In our case, the **Prob > $\mathbf{X^2}$** is 0.000 suggesting that we

can reject the null hypothesis and that the model is statistically significant at a 1% threshold. As logit regressions do not have a proper R^2 measure (unlike linear regression models), Stata computes and displays the **pseudo R^2** = 0.23 (among an extensive list of R^2, as displayed in Table 17). Yet beware, as the logit's pseudo R^2 has not the same interpretation as the traditional R^2 found in linear regression models. Indeed, the obtained pseudo R^2 has little meaning by itself and should be better used when compared to another pseudo R^2 obtained through the same kind of econometrical analysis (same data, same predicted outcome) in order to determine which of the tested models has the best predicting power. Although not presented in the tables, the list of the **log likelihoods at each iteration** suggest that the model converges. Indeed, the addition of new predictors at each iteration increases the maximum likelihood yet the difference between iterations becomes smaller each time and the iteration process stops when the difference is small enough, indicating the model has converged. The retained log likelihood of our final model is -48.42. However, again, such value has no particular meaning in itself but can be used to compare models. Finally, **Akaike and Bayesian information criteria** (AIC and BIC respectively) use the previously commented log likelihood result to compute a score that helps selecting a model. As for the pseudo R^2 and the log likelihood previously, the information criteria are useful when comparing models that use the same econometrical framework.

Table 16
Logit regression

yd	Coef.	St. Err.	t-value	p-value	[95% Conf	Interval]	Sig
tdta	**7.008**	**1.684**	**4.16**	**0**	**3.707**	**10.31**	*******
Constant	**-3.933**	**.946**	**-4.16**	**0**	**-5.788**	**-2.079**	*******

Mean dependent var	0.473	SD dependent var	0.502
Pseudo r-squared	0.231	Number of obs	91
Chi-square	29.033	Prob > chi2	0.000
Akaike crit. (AIC)	100.845	Bayesian crit. (BIC)	105.866

**** p<.01, ** p<.05, * p<.1*

Table 17

Measures of fit – logit regression

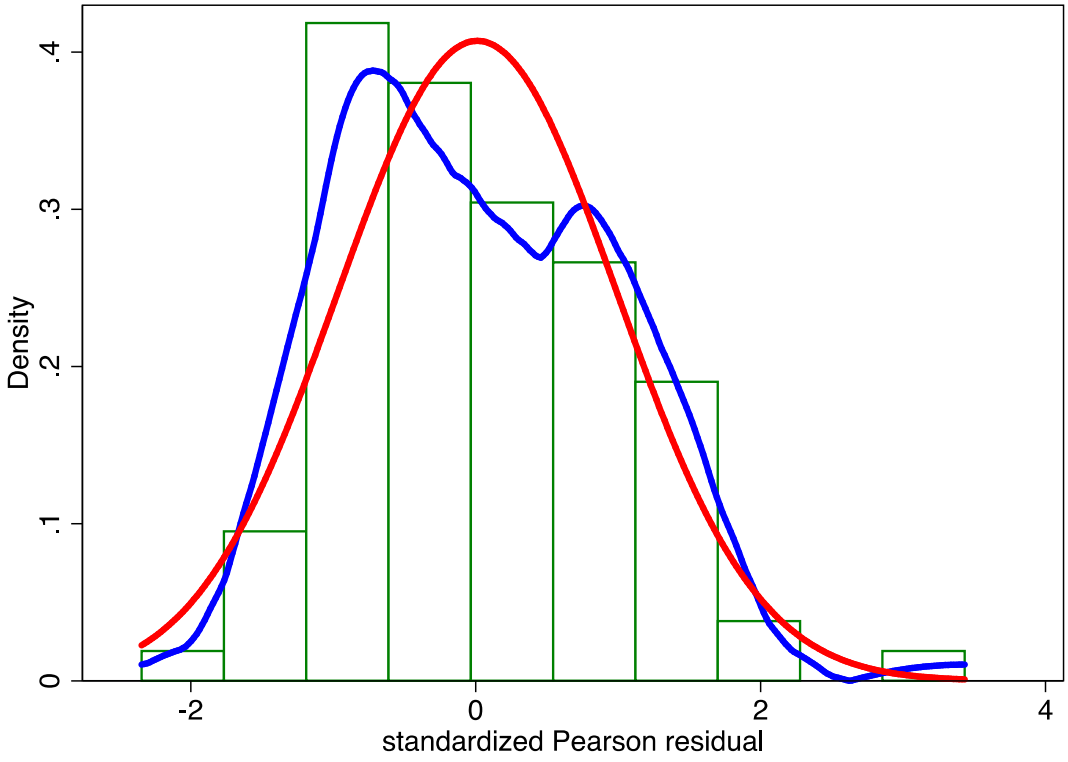
Measures of Fit for **logit** of **yd**

Log-Lik Intercept Only:	**-62.939**	Log-Lik Full Model:	**-48.422**
D(89):	**96.845**	LR(1):	**29.033**
		Prob > LR:	**0.000**
McFadden's R2:	**0.231**	McFadden's Adj R2:	**0.199**
ML (Cox-Snell) R2:	**0.273**	Cragg-Uhler(Nagelkerke) R2:	**0.365**
McKelvey & Zavoina's R2:	**0.411**	Efron's R2:	**0.266**
Variance of y*:	**5.585**	Variance of error:	**3.290**
Count R2:	**0.692**	Adj Count R2:	**0.349**
AIC:	**1.108**	AIC*n:	**100.845**
BIC:	**-304.622**	BIC':	**-24.522**
BIC used by Stata:	**105.866**	AIC used by Stata:	**100.845**

Once the logit regression model has been estimated, it is possible for us to extract the residuals ε and conduct some residual analysis on them. First, we have to determine which kind of residuals we want to extract: Pearson, deviance, working, partial or quantile residuals. We chose to work with standardized Pearson residuals, as their analysis is further down suggested, in question 17. The plotted distribution of the residuals is displayed in Figure 16 and the associated boxplots in Figure 17.

Figure 16

Distribution of standardized Pearson residuals – logit regression



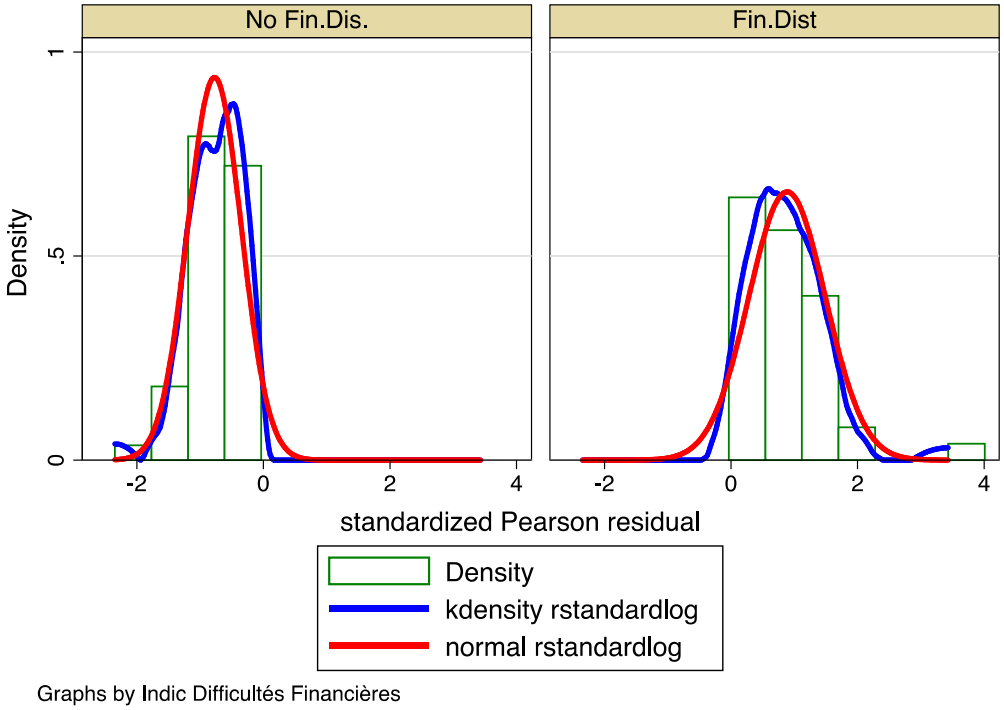
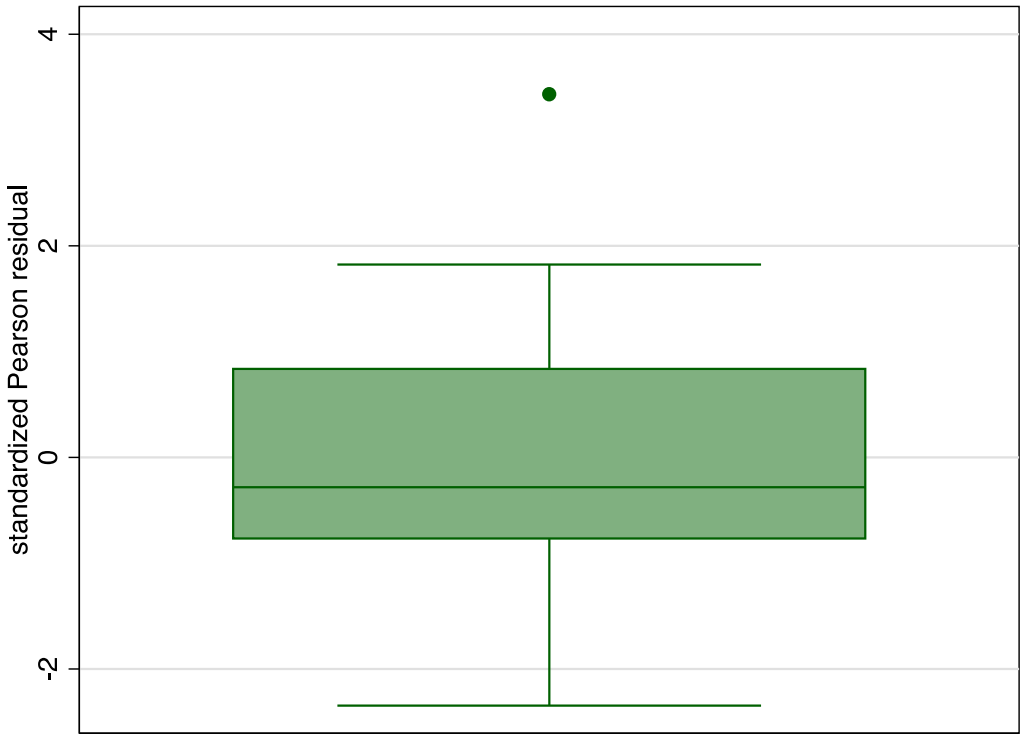
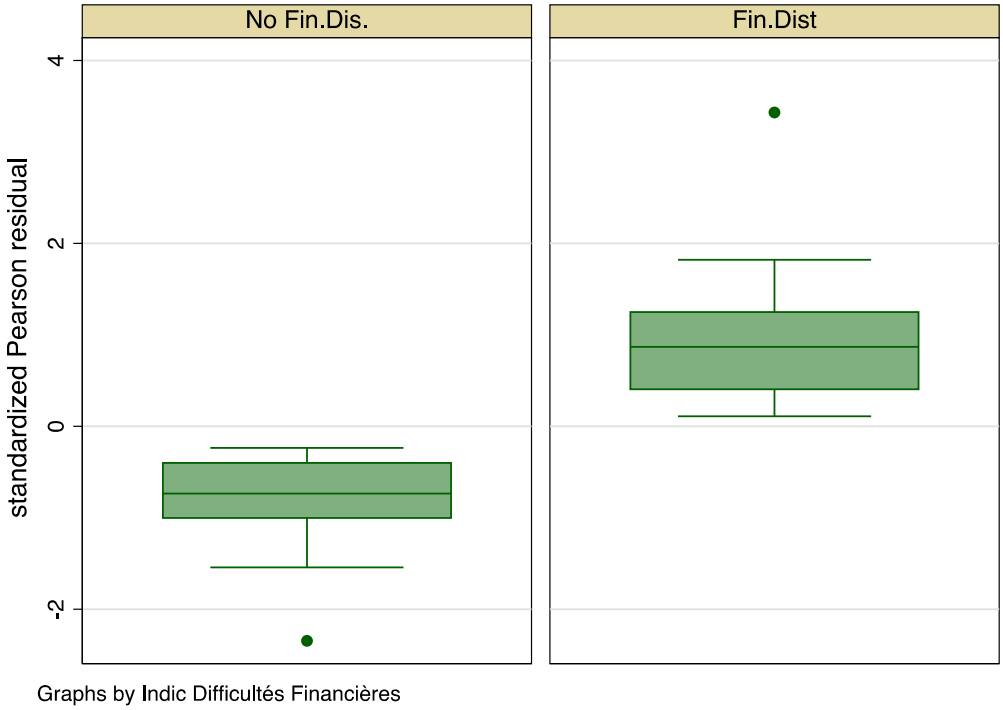


Figure 17

Boxplot distribution of standardized Pearson residuals – logit regression





From the plotted distribution and histogram of Figure 16, we can visually distinguish two different non-zero means by class from the *kdensity's* blue line, thus suggesting non-normality of residuals of the logit regression by class. However, the total sample distribution (all classes combined) seems to indicate that the distribution is normal, as illustrated by the normal distribution's red line. We might have a case of a mixture of two normal distributions.

To formally identify this hypothesis, we run two normality tests: first a skewness/kurtosis test and then a Shapiro-Wilk test. The results of such tests can be found in Table 18 and Table 19 respectively.

Table 18

Skewness and Kurtosis tests for Normality

			--------- Joint ---------		
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	Adj chi2(2)	Prob > chi2
rstandardlog	91	0.068	0.399	4.180	0.124

Table 19

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob > z
rstandardlog	91	0.965	2.637	2.140	0.016

Remember that a normal distribution has a skewness of zero and a kurtosis of three. The first test is then based on the difference between the residuals estimation's skewness and zero as well as its kurtosis and three. The test rejects the hypothesis of normality when the Prob $> X^2$ is less or equal to 0.05. In our case, the Prob $> X^2$ is equal to 0.124 suggesting that the skewness and kurtosis are sufficiently close to their respective normal values to not reject the hypothesis of normality.

In addition to this test, we perform a Shapiro-Wilk test based on the same general normality hypothesis. The test rejects the hypothesis of normality when the Prob > z is less or equal to 0.05. In our case, the Prob > z is equal to 0.016 implying that we reject the null hypothesis of normality.

Question 13

Compare the numerical values of the total debt/total assets ratio when it is the only explanatory variable for the linear probability model, the logit model and the probit model. Explain the differences?

Reminder: the full results for the linear and the logit models can be found in Table 8 and Table 16 respectively.

As we have already estimated the linear and logit models, we first estimate the third and last model: the probit. The results of the probit regression can be found in Table 20 below. Although very similar, the probit model differs from the logit one by the way it defines its predictors (that is the $f(*)$ in the predictor $\hat{Y} = f(\alpha + \beta x)$). The logit model uses a cumulative distribution function of the logistic distribution whereas the probit model uses a cumulative distribution of a standard normal distribution. However, both methods yield similar inference.

We can apply the same extensive methodology as in question 12 to interpret the results of the probit model but we will focus on the coefficient of interest $\beta_1 = 4.22$ which is statistically significant at the 1% threshold.

Table 20
Probit regression

yd	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
tdta	4.222	.936	4.51	0.000	2.386	6.057	***
Constant	-2.364	.524	-4.51	0.000	-3.391	-1.337	***

Mean dependent var	0.473	SD dependent var	0.502
Pseudo r-squared	0.233	Number of obs	91
Chi-square	29.373	Prob > chi2	0.000
Akaike crit. (AIC)	100.505	Bayesian crit. (BIC)	105.527

**** p<.01, ** p<.05, * p<.1*

It is now possible to gather all the beta estimates from the OLS, logit and probit regressions and compare them (Table 21).

Table 21
Beta estimates of OLS, Logit and Probit regressions
(including robust estimates)

Variable	bols	bolsr	bprobit	bprob~r	blogit	blogitr
_						
tdta	**1.205*****	**1.205*****				
	5.72	9.08				
_cons	**-0.187**	**-0.187****				
	-1.51	-2.11				
yd						
tdta			**4.222*****	**4.222*****	**7.008*****	**7.008*****
			4.51	5.06	4.16	4.79
_cons			**-2.364*****	**-2.364*****	**-3.933*****	**-3.933*****
			-4.51	-4.88	-4.16	-4.63
Statistics						
N	91	91	91	91	91	91
LL	-51.64	-51.64	-48.25	-48.25	-48.42	-48.42

*Note: b/t, *** p<.01, ** p<.05, * p<.1*

In this comparison, we focus mainly on the β coefficients. First, it is important to remember that the interpretation of the OLS β estimates differs from the ones obtained through logit and probit models. However, they share some concordant information. Indeed, the sign of all estimates is positive, which suggests that, whichever the model, the economic interpretation is the same: a higher total debt/total assets ratio increases the probability of being in a financial distress situation. We find another similarity in the fact that all β estimates are statistically significant at the 1% threshold, suggesting that the total debt/total assets ratio is a good predictor for the probability of being in a financial distress situation. The differences come mainly when interpreting the numerical values of the β estimates: for the OLS, the $\beta = 1.205$ is the slope of the linear relationship between yd and $tdta$ while for the logit and probit models, $\beta = 7.01$ and $\beta = 4.22$ respectively are the log-odds of being in a financial distress situation (yd) if the total debt/total assets ratio ($tdta$) increases by one unit. Also note that the log likelihoods are reported in the table, under the statistics section, supporting the idea that the nonlinear logit and probit models better fit our data than the linear OLS model and thus are more efficient to estimate the $\beta \times tdta$ coefficient ($LL_{OLS} = -51.64$ against $LL_{logit} = -48.42$ and $LL_{probit} = -48.25$).

Question 14

Explain how you obtain the percentage of concordant pairs when the dependent variable is binary. Is there a link of the AUC with the number of concordant pairs (define what is a concordant pair)?

The notion of concordant, discordant and tied pairs is related to both the cut-off value and the area under the curve (AUC) value. Indeed, after estimating the probabilities in our binary logit regression model, the data is divided into two datasets. The first one contains x observations having the actual value of dependent variable 1 (default event) and other contains y observations having the actual value of dependent variable 0 (non-default event). Of course, each pair has its own probability of happening. Each predicted value in the first dataset is then compared to the predicted value in the second dataset. We then have $x \times y$ pairs to compare. The pairings are defined as follows:

i. A pair is **concordant** if an observation x (event) has a higher predicted probability than y (non-event). The coexistence of both cases is possible since the probability of the event is higher than the one of the non-event. The cut-off value will be circumscribed by the probability of both events.

ii. A pair is **discordant** if an observation y (non-event) has a higher predicted probability than x (event). The coexistence of both cases is impossible since the non-event cannot have a higher probability of happening than the event itself.

iii. A pair is **tied** if an observation x (event) has the same predicted probability than y (non-event). The coexistence of both cases is possible only if the cut-off value is equal to the shared probability of happening of both the event and non-event.

It is now possible to compute the ratios for each category of pairs. The concordance ratio is the number of concordant pairs over the total number of pairs. The same computation applies to the discordance and tied ratios. By using the concordance and tied pairs' ratios, it is possible to compute the AUC (or C-stat):

$$AUC = Concordant\ pairs\ ratio + 0{,}5 \times Tied\ pairs\ ratio \qquad (11)$$

The AUC is therefore a measure of separability of the model, telling us how much the model is able to distinguish between classes. The higher the number of concordant pairs, the higher the value of the AUC and thus the better the model is at predicting 1 events as 1 and 0 non-events as 0.

Question 15

What is the value of the AUC, area under the ROC curve when you used only total debt/total assets as an explanatory variable?

As presented in the previous question 14, the AUC tells us the separability power of our model, in distinguishing between event and non-event categories. Yet, we know that our model is not perfect and has two overlapping distributions, introducing both type 1 and type 2 errors. We can however minimize them depending on the cut-off value we choose. In order to find the optimal cut-off value, some tests have been developed but might not be optimal in practice since they are based on the predicted estimates of sensitivity and specificity, which are generated by the prediction model and thus are positively biased. Having this warning in mind, we can however proceed to the test to understand how the AUC curve works. The first test is the empirical estimation of cut-off value (Clayton, 2013) displayed in Table 22. It is also possible to graphically represent the sensitivity, specificity and probability cut-off, as displayed in Figure 18.

Table 22

Empirical estimation of the cut-off value

Method	Liu
Reference variable	*yd* (0=neg, 1=pos)
Classification variable	*tdta*
Empirical optimal cutpoint	**0.56340277**
Sensitivity at cutpoint	0.63
Specificity at cutpoint	0.77
Area under ROC curve at cutpoint	**0.70**

Figure 18

Sensitivity, Specificity and probability cut-off

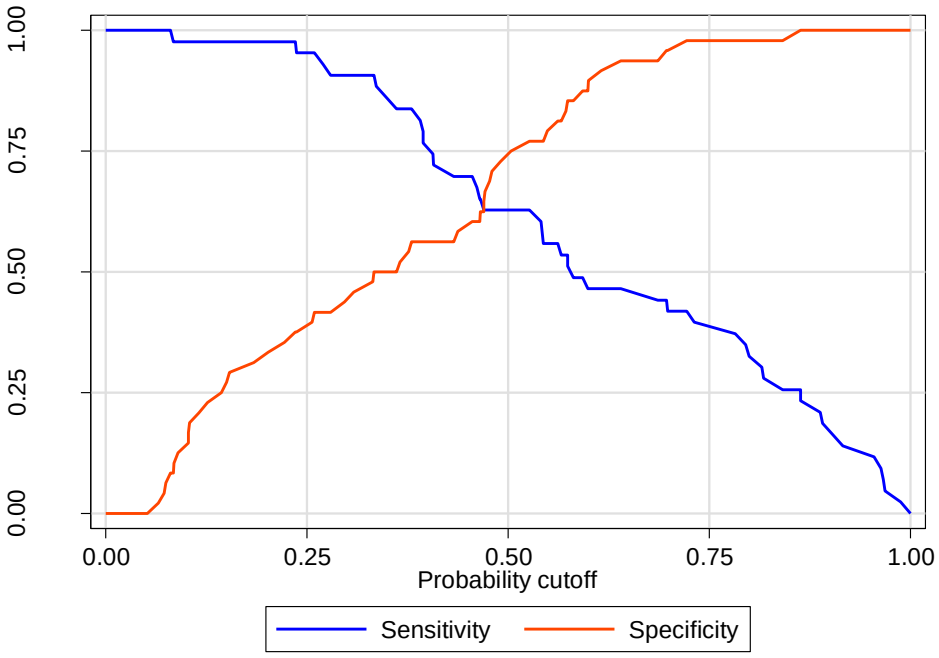


Table 23
Classification of the Logit regression model
(cut-off value: 0.56)

Logistic model for yd

Classified	True D	True ~D	Total
+	24	9	33
-	19	39	58
Total	43	48	91

Classified + if predicted Pr(D) >= .56
True D defined as yd != 0

Sensitivity	Pr(+| D)	55.81%
Specificity	Pr(-|~D)	81.25%
Positive predictive value	Pr(D| +)	72.73%
Negative predictive value	Pr(~D| -)	67.24%
False + rate for true ~D	Pr(+|~D)	18.75%
False - rate for true D	Pr(-| D)	44.19%
False + rate for classified +	Pr(~D| +)	27.27%
False - rate for classified -	Pr(D| -)	32.76%
Correctly classified		69.23%

Table 24
Classification of the Logit regression model
(cut-off value: 0.7)

Logistic model for yd

Classified	True D	True ~D	Total
+	18	1	19
-	25	47	72
Total	43	48	91

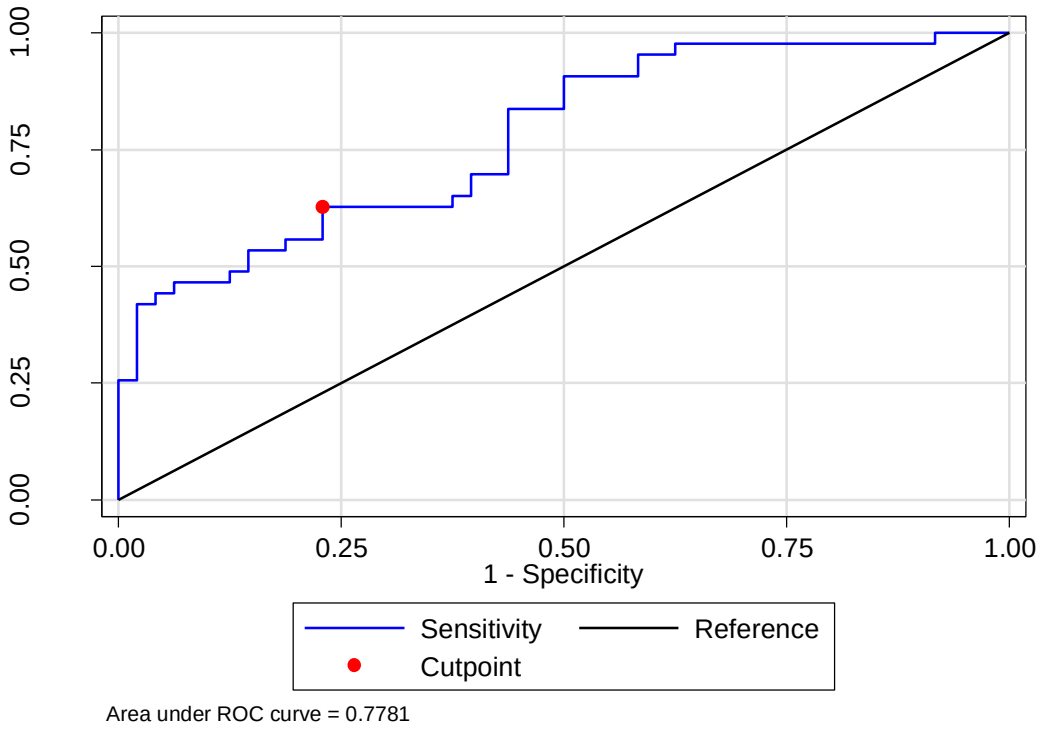
Classified + if predicted Pr(D) >= .7
True D defined as yd != 0

Sensitivity	Pr(+| D)	41.86%
Specificity	Pr(-|~D)	97.92%
Positive predictive value	Pr(D| +)	94.74%
Negative predictive value	Pr(~D| -)	65.28%
False + rate for true ~D	Pr(+|~D)	2.08%
False - rate for true D	Pr(-| D)	58.14%
False + rate for classified +	Pr(~D| +)	5.26%
False - rate for classified -	Pr(D| -)	34.72%
Correctly classified		71.43%

The suggested empirical optimal cut-off value is 0.56 which we use as a parameter in our classification for the logit model displayed in Table 23. The percentage of correctly classified events is 69.23%. However, as the econometrician can tweak the cut-off value, it is possible to find higher classification percentages, but at the detriment of the balance between positive and negative predicted values as the comparison between Table 23 (cut-off value: 0.56) and Table 24 (cut-off value: 0.7) show.

Finally, we can compute the AUROC curve. The plotted result can be found in Figure 19 (the presented cut-off value is 0.56). The area under the ROC curve is 0.7781, meaning that the model is good at predicting the observations that will experience the event and those who will not, with a 77.81% probability of having the good classification.

Figure 19
AUC and ROC curves



Question 16

Include your preferred list of explanatory variables out of the 14 financial ratios. Estimate the model with logit. Comment the results. Comment the value of the area under the ROC curve.

Adding new explanatory variables should help the model to increase its separability power and thus improve the overall AUROC value. Note that ROC curves are quite useful for comparing the predictive capability of more than one explanatory variable at once for the same event (here, the default probability). Of course, by choosing or adding different explanatory variables, one can improve the predictability power of the model even further. Our first model takes into account the variables that have the higher correlation with yd. Our second model takes a set of diversified variables to test if variables belonging to different categories of financial indicators improve the model. Finally, our last model includes all the fourteen explanatory variables.

i. Model 1: Explanatory variables with the highest correlation rank with yd

Following the correlation matrix presented in question 10, we decide to add the following explanatory variables: $reta, opita\ and\ ebita$. Table 25 presents the results of the logit regression estimation.

Table 25
Estimation of the Logit regression model
Model 1: Correlation set

yd	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
tdta	4.759	2.152	2.21	0.027	.542	8.976	**
reta	-1.477	1.759	-0.84	0.401	-4.925	1.971	
opita	-10.353	10.239	-1.01	0.312	-30.422	9.715	
ebita	5.388	9.465	0.57	0.569	-13.163	23.939	
Constant	-1.454	1.639	-0.89	0.375	-4.666	1.758	

Mean dependent var	0.473	SD dependent var	0.502
Pseudo r-squared	0.264	Number of obs	91
Chi-square	33.171	Prob > chi2	0.000
Akaike crit. (AIC)	102.707	Bayesian crit. (BIC)	115.261

**** p<.01, ** p<.05, * p<.1*

Adding three new explanatory variables brings some disturbance in the individual β estimates of the model as all variables are not statistically significant at the 5% threshold (except $tdta$). However, even if it might seem scary at first glance because of our "OLS output bias" (the newly added pertinent variables should be also significant), it is not the case for logit models. Indeed, adding new variables does not improve the individual β estimation precision but rather the overall model fitting. This is confirmed by the Prob > X^2 = 0.000 suggesting the overall model fits well the input data. Now, let us turn to the presentation of the classification and AUROC curves.

The cut-off value has been set at 0.5 for all models. Table 26 shows that introducing some new explanatory variables has slightly improved the correct classification percentage to 70.33% (against 69.23% for the $tdta$ variable alone). Figure 21 reflects this improvement as the area under the ROC curve also improves to 0.8246 (against 0.7791). We can conclude stating that this first augmented model has improved the predictability power of our previous model.

Table 26

Classification of the Logit regression model

Model 1: Correlation set

Logistic model for yd

Classified	True D	True ~D	Total
+	26	10	36
-	17	38	55
Total	43	48	91

Classified + if predicted Pr(D) >= .5
True D defined as yd != 0

Sensitivity	Pr(+| D)	60.47%
Specificity	Pr(-|~D)	79.17%
Positive predictive value	Pr(D| +)	72.22%
Negative predictive value	Pr(~D| -)	69.09%
False + rate for true ~D	Pr(+|~D)	20.83%
False - rate for true D	Pr(-| D)	39.53%
False + rate for classified +	Pr(~D| +)	27.78%
False - rate for classified -	Pr(D| -)	30.91%
Correctly classified		70.33%

Figure 20

Sensitivity, Specificity and probability cut-off

Model 1: Correlation set

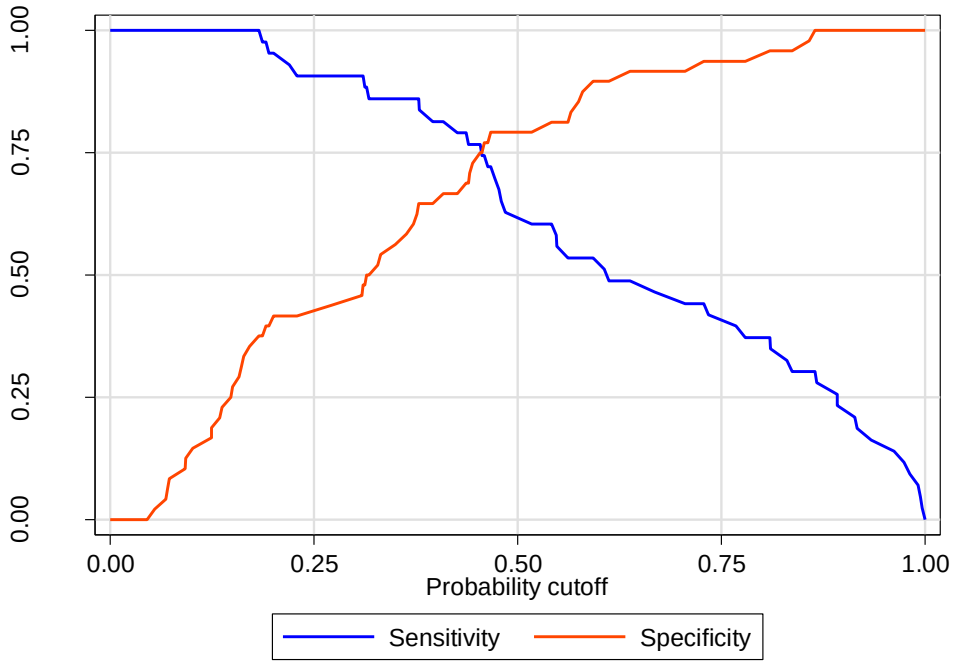
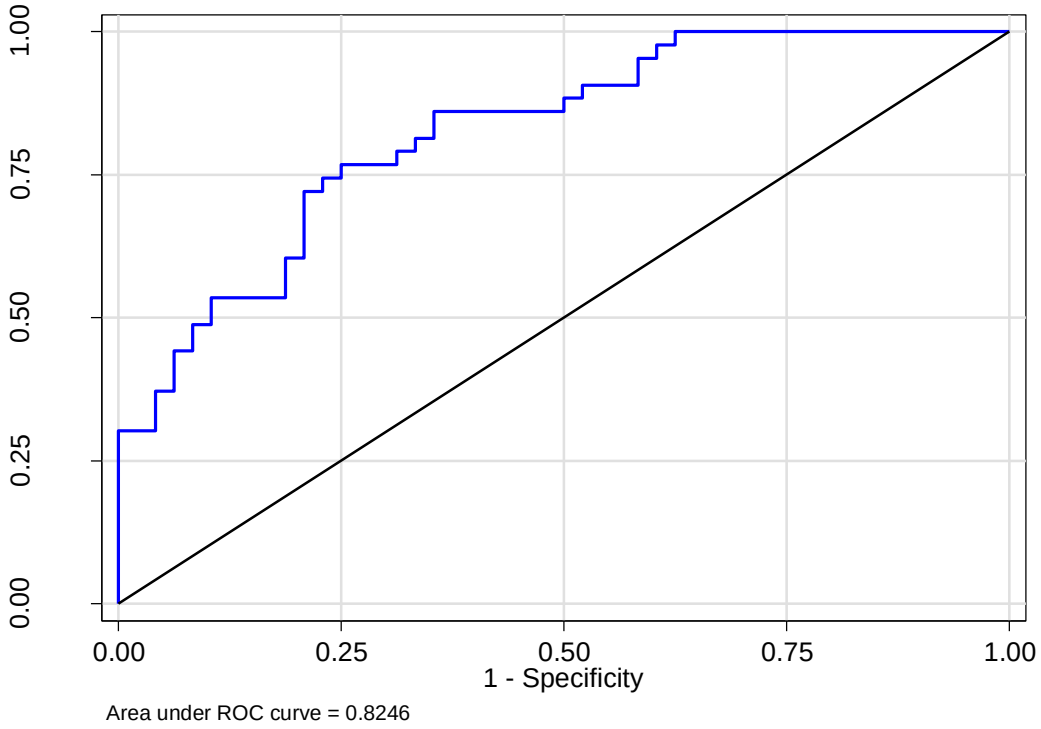


Figure 21
AUC and ROC curves
Model 1: Correlation set



ii. Model 2: Diversified set of financial explanatory variables

The idea behind this second model is to avoid some possible redundancies using financial indicators that belong to the same group (such as the $reta$, $opita$ and $ebita$ variables as in model 1 which fall into the "earnings" indicators category) and might bias the estimated model with multicollinearity issues and thus affect the predictability power of our model. This model is based on a diversified set of indicators: retained earnings, log of sales, the growth of number of employees, the net working capital/total assets, fixed assets/total assets and the market value equity/long term debt. The added explanatory variables are: $tdta$, $reta$, $lsls$, $gempl$, $nwcta$, $mveltd$ and $fata$.

As for model 1, Table 27 shows that most of the individual β estimates are not significant (except for $gempl$) yet the Prob > X^2 = 0.000 suggests that the overall model fits well the input data again.

Remember that the cut-off value has been set at 0.5. Table 28 shows that introducing some diversified explanatory variables has improved the correct classification percentage to 74.44% (against 69.23% for the $tdta$ variable alone and 70.33% for model 1). Figure 23 reflects this improvement as the area under the ROC curve also improves to 0.8461 (against 0.7791 and 0.8246 respectively). We can conclude stating that this second augmented model has improved even further the predictability power of our two previous models.

Table 27

Estimation of the Logit regression model

Model 2: Diversified set

yd	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
tdta	8.131	4.378	1.86	0.063	-.45	16.713	*
reta	-3.052	2.064	-1.48	0.139	-7.097	.994	
lsls	-.104	.189	-0.55	0.584	-.475	.267	
gempl	-5.684	2.618	-2.17	0.030	-10.816	-.552	**
nwcta	1.523	3.508	0.43	0.664	-5.354	8.399	
mveltd	1.28	3.008	0.43	0.670	-4.616	7.177	
fata	-3.245	3.956	-0.82	0.412	-10.998	4.509	
Constant	-3.175	3.434	-0.92	0.355	-9.905	3.556	

Mean dependent var	0.478	SD dependent var	0.502
Pseudo r-squared	0.312	Number of obs	90
Chi-square	38.810	Prob > chi2	0.000
Akaike crit. (AIC)	101.779	Bayesian crit. (BIC)	121.778

**** p<.01, ** p<.05, * p<.1*

Table 28

Classification of the Logit regression model

Model 2: Diversified set

Logistic model for yd

	True		
Classified	D	~D	Total
+	29	9	38
-	14	38	52
Total	43	47	90

Classified + if predicted Pr(D) >= .5
True D defined as yd != 0

Sensitivity	Pr(+| D)	67.44%
Specificity	Pr(-|~D)	80.85%
Positive predictive value	Pr(D| +)	76.32%
Negative predictive value	Pr(~D| -)	73.08%
False + rate for true ~D	Pr(+|~D)	19.15%
False - rate for true D	Pr(-| D)	32.56%
False + rate for classified +	Pr(~D| +)	23.68%
False - rate for classified -	Pr(D| -)	26.92%
Correctly classified		74.44%

Figure 22
Sensitivity, Specificity and probability cut-off
Model 2: Diversified set

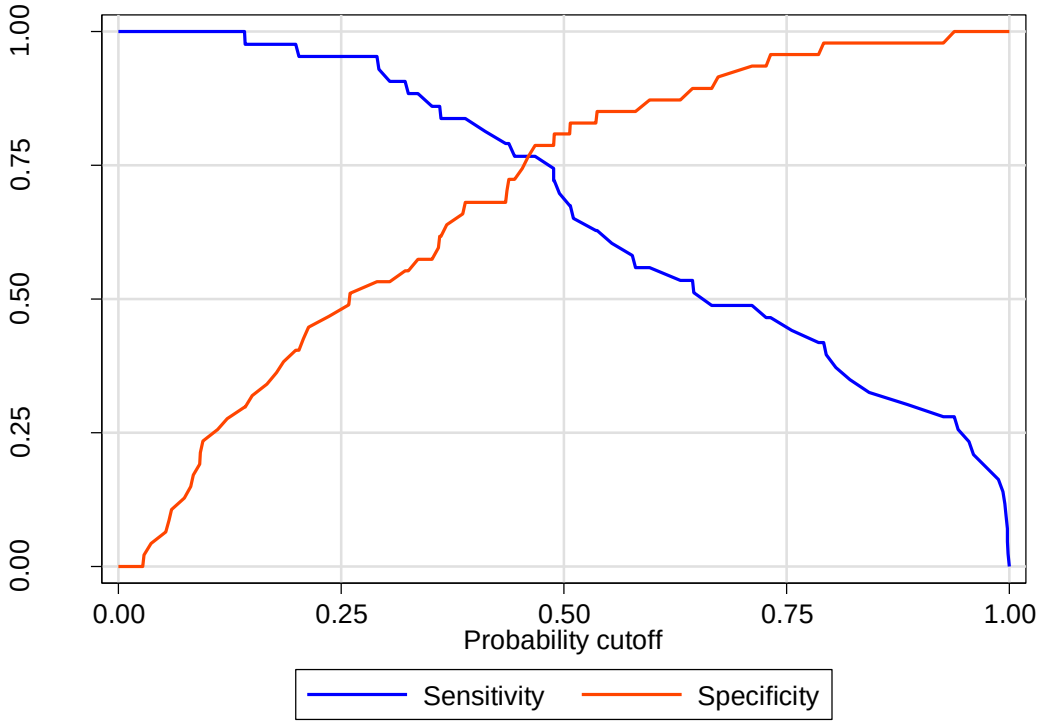
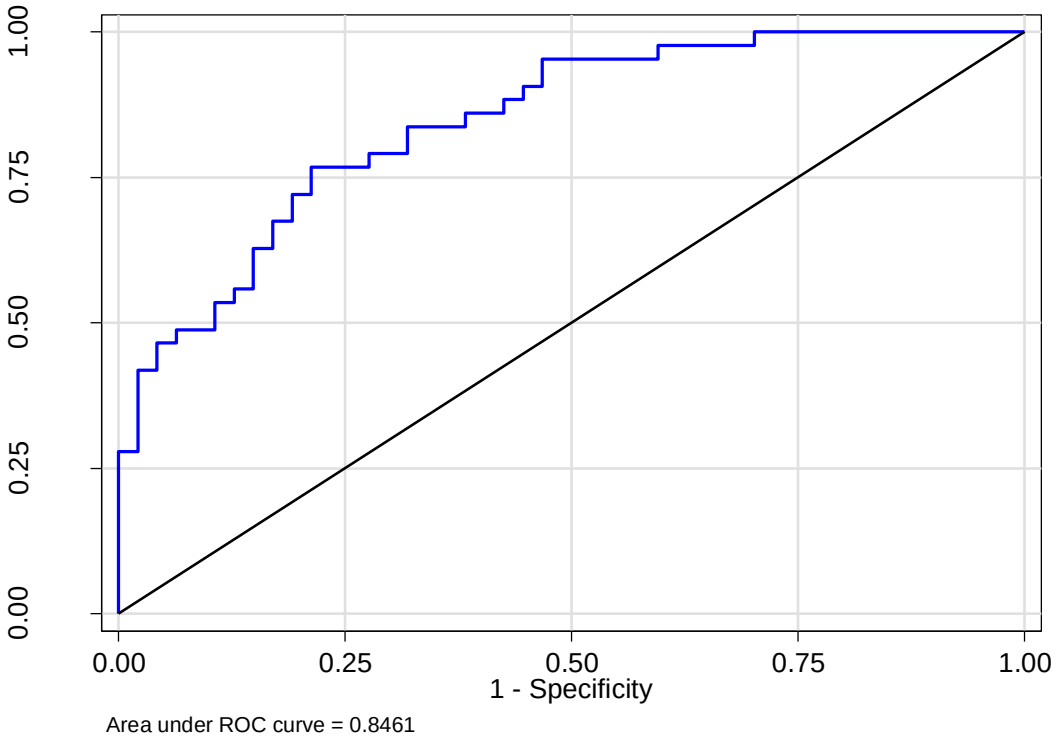


Figure 23
AUC and ROC curves
Model 2: Diversified set



iii. Model 3: The whole set of financial indicators

Finally, we test the logit model with the whole set of financial indicators. The idea is to understand how adding an increasing number of variables improves (or not) the baseline model.

As for model 1 and 2, Table 29 shows that most of the individual β estimates are not significant (except for $gempl$ and $ltdta$) yet the Prob > X^2 = 0.000 suggests that the overall model fits well the input data again.

Remember that the cut-off value has been set at 0.5. Table 30 shows that introducing the whole set of explanatory variables has improved the correct classification percentage to 77.53% (against 69.23% for the $tdta$ variable alone, 70.33% for model 1 and 75.44% for model 2). Figure 25 reflects this improvement as the area under the ROC curve also improves to 0.8857 (against 0.7791; 0.8246 and 0.8461 respectively). We can conclude stating that this third and final augmented model has improved even further and beyond the predictability power of our three previous models.

Table 29
Estimation of the Logit regression model
Model 3: Whole set

yd	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
tdta	-2.103	5.952	-0.35	0.724	-13.768	9.562	
reta	-5.795	3.06	-1.89	0.058	-11.793	.203	*
opita	-16.757	12.659	-1.32	0.186	-41.568	8.055	
ebita	17.222	11.007	1.56	0.118	-4.352	38.797	
lsls	1.596	1.392	1.15	0.252	-1.132	4.324	
lta	-1.674	1.401	-1.19	0.232	-4.421	1.072	
gempl	-8.027	3.528	-2.28	0.023	-14.941	-1.112	**
invsls	17.705	9.867	1.79	0.073	-1.634	37.043	*
nwcta	-9.26	7.021	-1.32	0.187	-23.022	4.502	
cacl	-1.186	1.228	-0.97	0.334	-3.593	1.221	
qacl	3.322	1.816	1.83	0.067	-.238	6.882	*
fata	-1.207	4.499	-0.27	0.788	-10.025	7.61	
ltdta	-.761	.388	-1.96	0.050	-1.522	0	**
mveltd	.114	3.459	0.03	0.974	-6.666	6.893	
Constant	2.683	4.881	0.55	0.583	-6.884	12.249	

Mean dependent var	0.483	SD dependent var	0.503
Pseudo r-squared	0.395	Number of obs	89
Chi-square	48.740	Prob > chi2	0.000
Akaike crit. (AIC)	104.540	Bayesian crit. (BIC)	141.869

**** p<.01, ** p<.05, * p<.1*

Table 30

Classification of the Logit regression model

Model 3: Whole set

Logistic model for yd

	True		
Classified	D	~D	Total
+	33	10	43
-	10	36	46
Total	43	46	89

Classified + if predicted Pr(D) >= .5
True D defined as yd != 0

Sensitivity	Pr(+| D)	76.74%
Specificity	Pr(-|~D)	78.26%
Positive predictive value	Pr(D| +)	76.74%
Negative predictive value	Pr(~D| -)	78.26%
False + rate for true ~D	Pr(+|~D)	21.74%
False - rate for true D	Pr(-| D)	23.26%
False + rate for classified +	Pr(~D| +)	23.26%
False - rate for classified -	Pr(D| -)	21.74%
Correctly classified		77.53%

Figure 24

Sensitivity, Specificity and probability cut-off

Model 3: Whole set

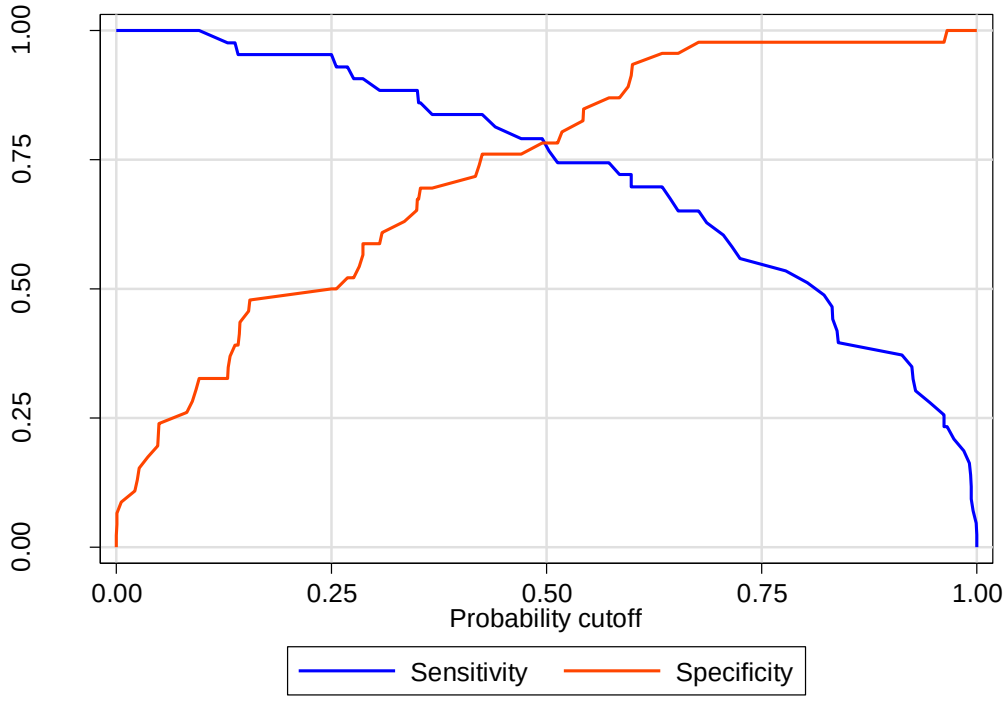
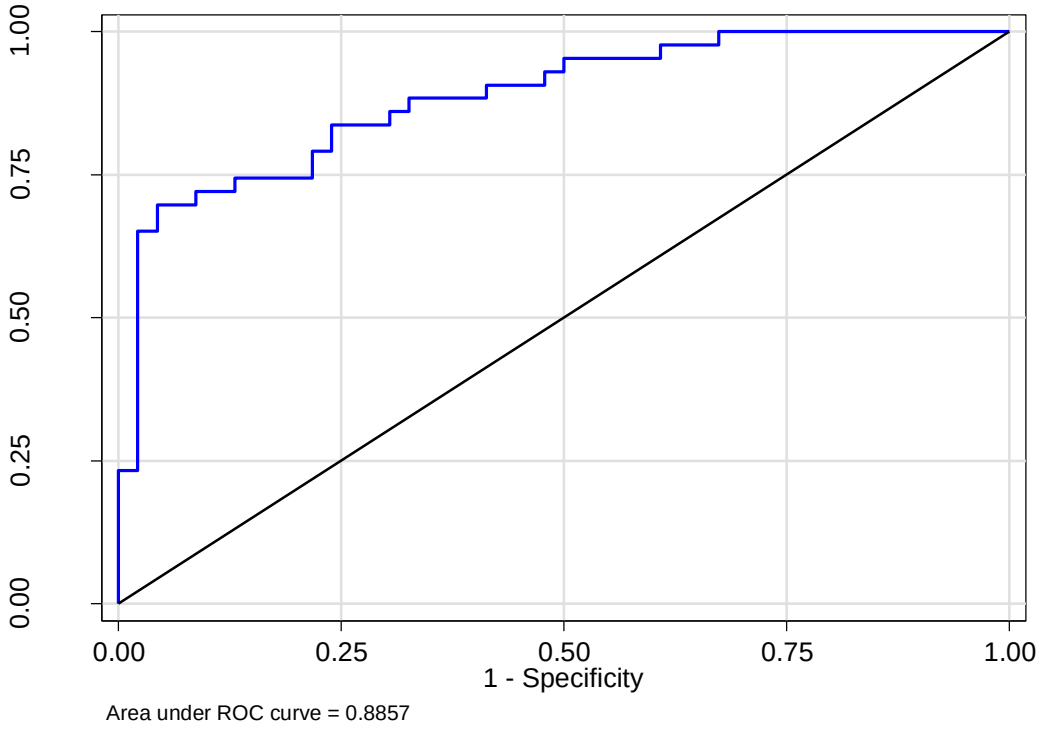


Figure 25
AUC and ROC curves
Model 3: Whole set



Question 17
Compute studentized residuals of your model. Do you find outliers?

We come back to our baseline model with only the $tdta$ explanatory variable. We compute the standardized Pearson residuals for the $tdta$ variable but also for the predicted $tdta$ distress indicator extracted from the logit regression. The plotted Pearson standardized residuals for both variables can be found in Figures 26 ($tdta$) and 27 (predicted $tdta$). Given $|2|$ bounds for outliers identification (which is the non-normality threshold) we notice some outliers of type I and type II. Type I outliers correspond to financially healthy firms (low total debt/total asset ratio in our case) but that defaulted. On the contrary, type II outliers are associated to unhealthy firms (high $tdta$ ratio) but that did not default. In our sample, we seem to have one type I and one type II outliers.

Figure 26
Outliers of Pearson standardized residuals
(tdta variable)

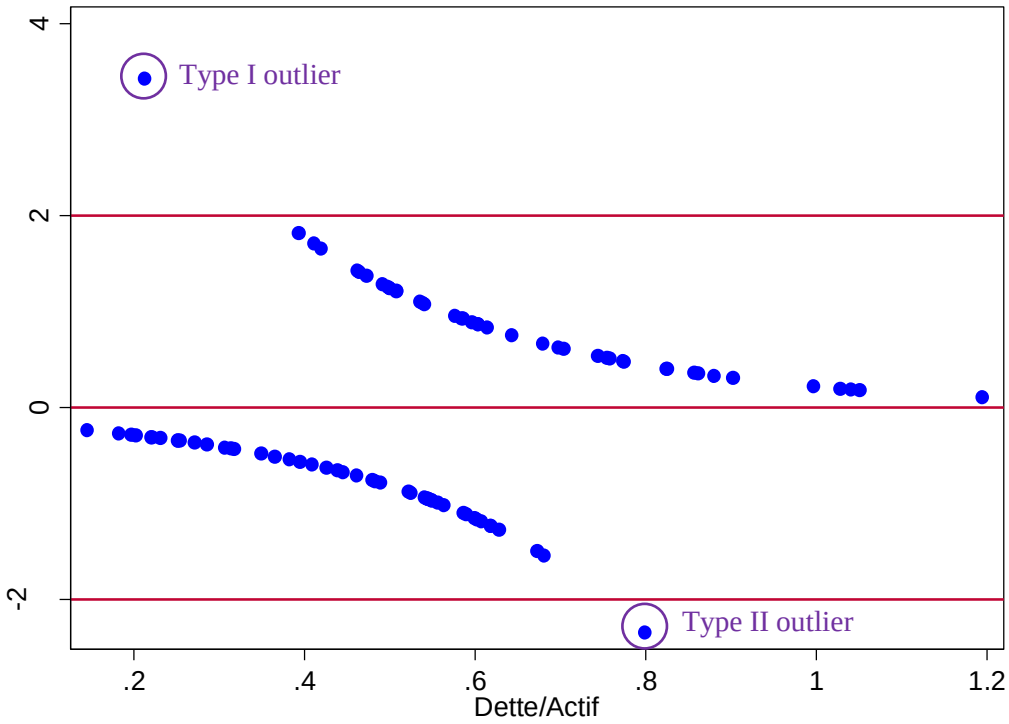
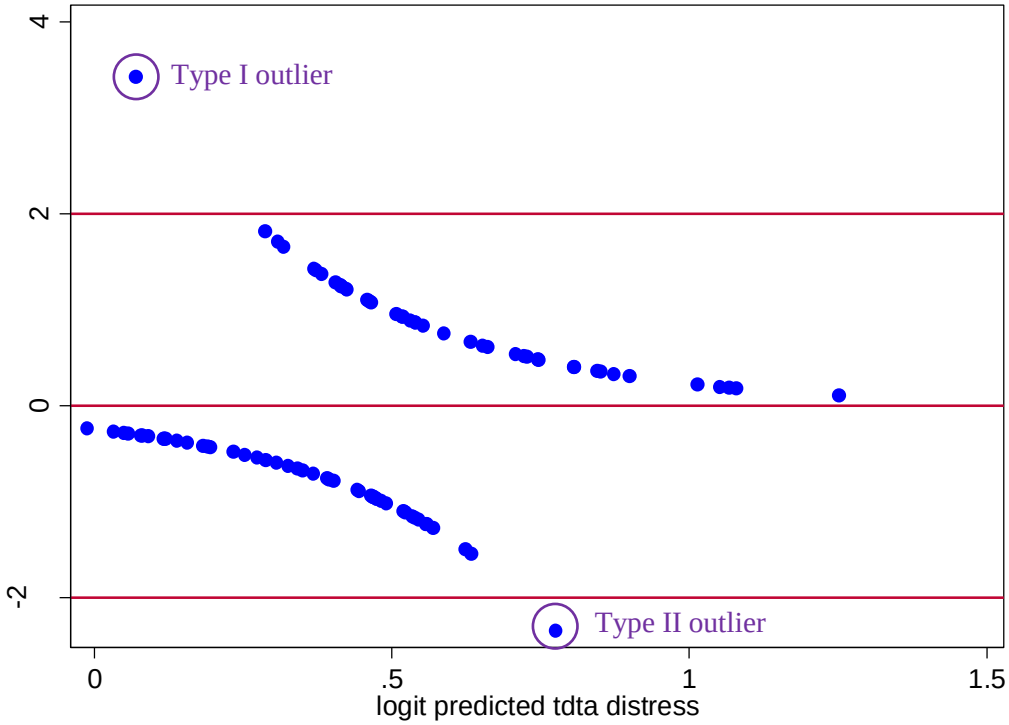


Figure 27
AUC and ROC curves
(predicted tdta)



Question 18

For a loss function of the type 1 and type 2 errors, what would be relative weight for a private banker? How one can take into account this loss function for selecting between scoring models taking into account the ROC curve of each model, in particular if the ROC curve intersect (locally one is over the other and conversely).

In an ideal world, the private banker would choose the threshold s^* so that the indicator variables always exceed the critical thresholds ahead of the event (which is our case is the firm defaulting) but never during non-events (the firm not defaulting). However, in practice, bankers' models are confronted with type I (healthy firms defaulting) and type II (unhealthy firms not defaulting) errors. Of course, the banker is trying to maximize its gains and thus is interested in minimizing its own loss function. The banker has then to arbitrage the weights of each kind of errors allowed in his model, by tweaking the s threshold. By lowering the s threshold, he allows his model to predict a higher percentage of potentially defaulting firms (as more positive signals are issued) which reduces the amount of type I errors but at the detriment of the precision to predict type II error, which is more defaulting firms that in the end do not. The banker will then choose its optimal threshold taking into account the relative costs of type I against type II errors. We can formally present it such as:

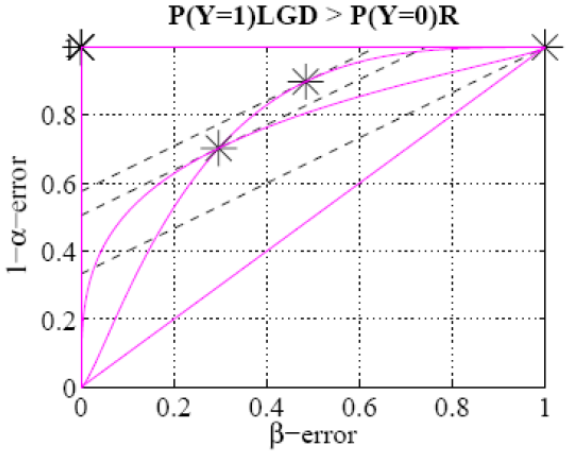
$$\min_s[LGD] = \min_s[\alpha T_1 + (1-\alpha)T_2] \qquad (12)$$

where the banker minimizes the loss given default (LGD) function by arbitraging the type I (T_1) and type II (T_2) errors weighted by α given the s threshold. In the case of the banker, he will be keen on increasing the weight on the type I errors in order to avoid them (a lower threshold s implies more potentially defaulting firms in the period to come) since the relative cost of type I errors is higher than the type II. Indeed, type I errors induce that lending money to a firm which then defaults makes the banker lose both the principal (the amount of money lent) and the interests on that principal. Type II errors are less costly as the banker misses the opportunity to lend to a firm that will not default, thus only losing the benefits linked to the interests.

If the banker knows the ROC curves of his models (which can locally overlap), he can then compute the iso-lines that define the optimal threshold s^* taking into account his own loss function and determine which model suits his activity the most. The higher the iso-line, the better the minimization of his losses.

Figure 28

ROC curves and iso-lines



Question 19

Compute the AUC, area under ROC curve, for the sample of VALIDATION (add any other statistics on the sample of validation which you think they are interesting).

It is time to swap our samples and use the validation observations. We follow the same procedure as in question 15 for our first model (the baseline model) and then reuse the methodology presented in question 16 for our second model (diversified set of variables). By presenting both models, we try to demonstrate –with the validation sample- that the model improves its predictability power when a higher number of pertinent variables are added.

i. Model 1: Baseline model

In the baseline model, the only explanatory variable is $tdta$. The results of the optimal cut-off test is reported in Table 31, the plotted sensitivity and specificity are shown in Figure 29, the correctly classified classes are presented in Table 32 and finally the AUC and ROC curves are displayed in Figure 30.

Table 31

Empirical estimation of the cut-off value

(Baseline model)

Method	Liu
Reference variable	*yd* (0=neg, 1=pos)
Classification variable	*tdta*
Empirical optimal cutpoint	**0. 50446898**
Sensitivity at cutpoint	0.72
Specificity at cutpoint	0.53
Area under ROC curve at cutpoint	**0.63**

Figure 29

Sensitivity, Specificity and probability cut-off

(Baseline model)

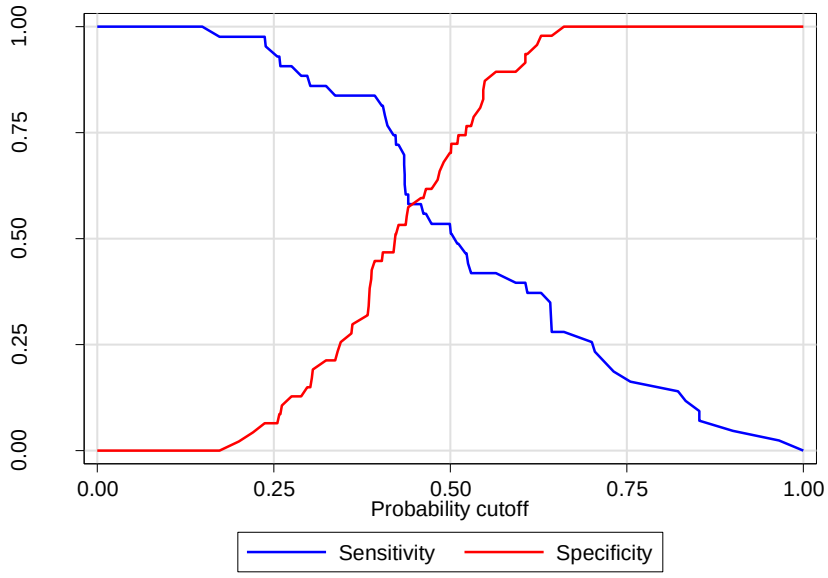


Table 32
Classification of the Logit regression model
(Baseline model)

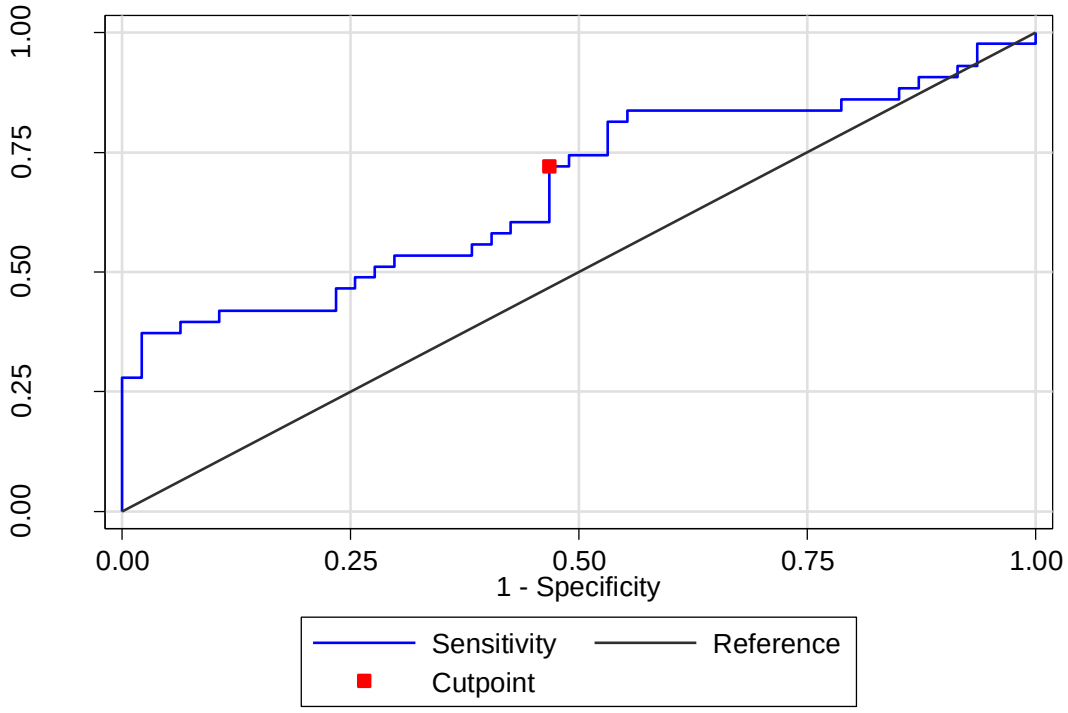
Logistic model for yd

	True		
Classified	D	~D	Total
+	22	14	36
-	21	33	54
Total	43	47	90

Classified + if predicted Pr(D) >= .5
True D defined as yd != 0

Sensitivity	Pr(+| D)	51.16%
Specificity	Pr(-|~D)	70.21%
Positive predictive value	Pr(D| +)	61.11%
Negative predictive value	Pr(~D| -)	61.11%
False + rate for true ~D	Pr(+|~D)	29.79%
False - rate for true D	Pr(-| D)	48.84%
False + rate for classified +	Pr(~D| +)	38.89%
False - rate for classified -	Pr(D| -)	38.89%
Correctly classified		61.11%

Figure 30
AUC and ROC curves
(Baseline model)



Area under ROC curve = 0.6734

The area under the ROC curve using the validation sample is lower than for the estimation sample: 0.6734 against 0.7791. The model is also less efficient at categorizing firms (61.11% correctly classified firms against 69.23%) and thus at predicting the defaulting firms (67.34%). This result, although worse off than in the previous sample, is however quite in line with the overall results obtained during this empirical application: the total debt/total assets helps predicting the defaulting firms.

ii. Model 2: Diversified set of financial explanatory variables

This model is based on a diversified set of indicators: retained earnings, log of sales, the growth of number of employees, the net working capital/total assets, fixed assets/total assets and the market value equity/long term debt. The model has the following explanatory variables: $tdta$, $reta$, $lsls$, $gempl$, $nwcta$, $mveltd$ and $fata$.

The results of this second model are an encouraging improvement of the previous baseline model. Indeed, 74.16% of the firms are correctly classified (which is quite similar to the 74.44% of the estimation sample) while the AUROC value has improved to 0.8504 (even better than the one in the estimation sample, at 0.8461). This model is strong at predicting defaulting firms given the diversified set of financial explanatory variables we used as input data.

In conclusion, we showed that the estimation and validation samples yield similar results and that introducing new relevant variables do improve the prediction power of our model.

Figure 31
Sensitivity, Specificity and probability cut-off
(Diversified set of variables model)

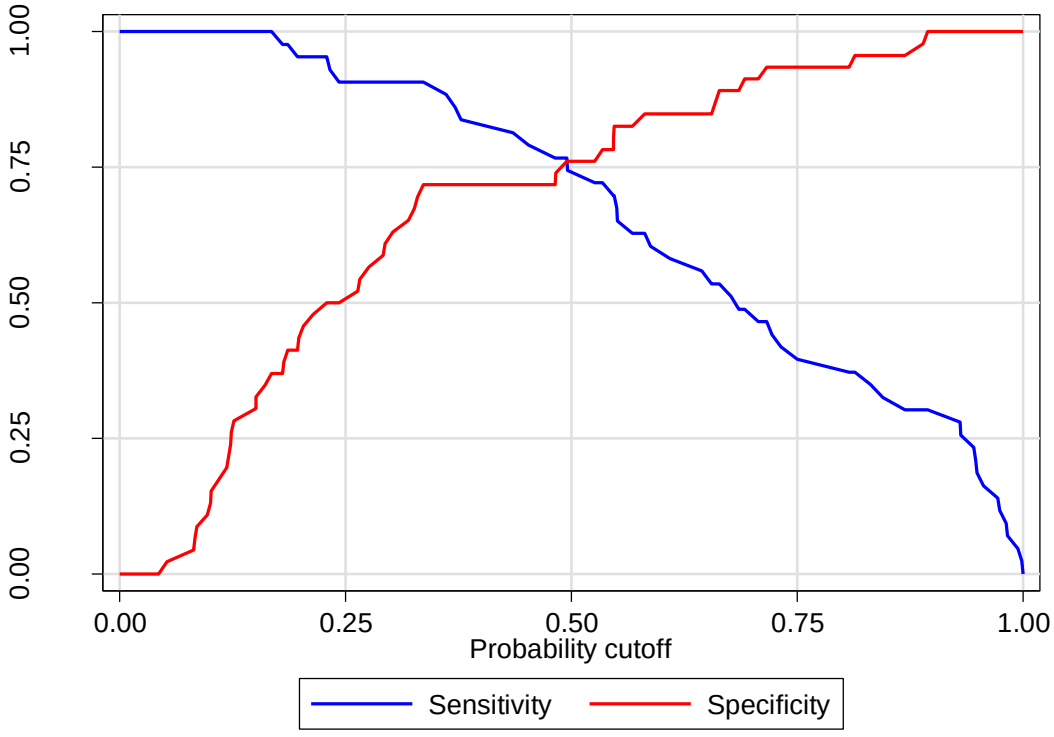


Table 33
Classification of the Logit regression model
(Diversified set of variables model)

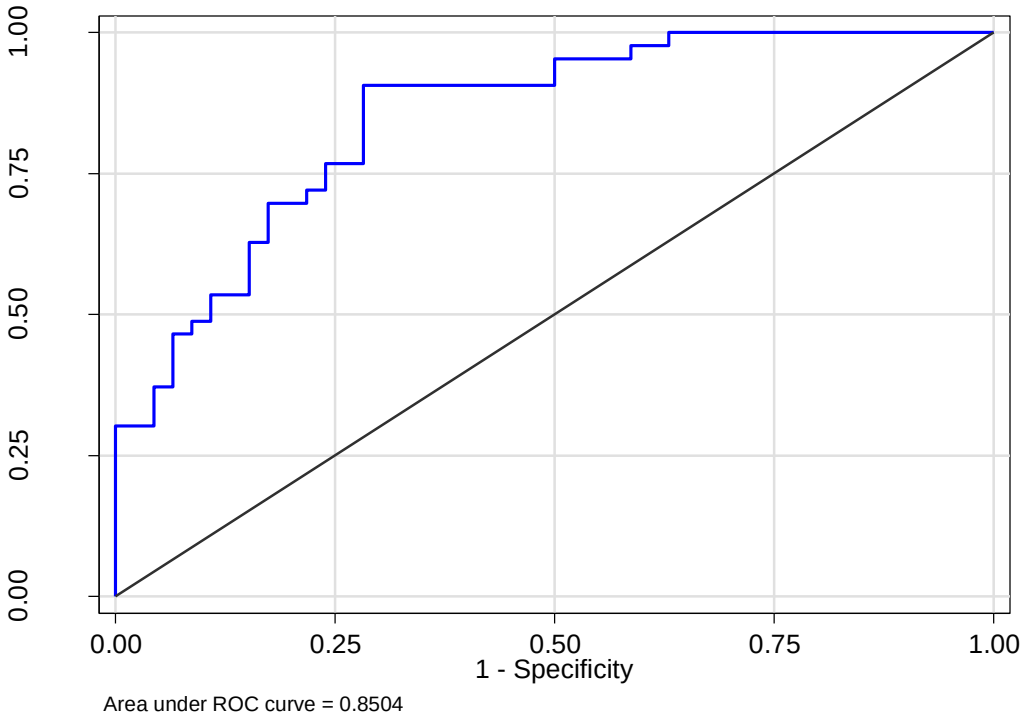
Logistic model for yd

Classified	True D	True ~D	Total
+	31	11	42
-	12	35	47
Total	43	46	89

Classified + if predicted Pr(D) >= .5
True D defined as yd != 0

Sensitivity	Pr(+| D)	72.09%
Specificity	Pr(-|~D)	76.09%
Positive predictive value	Pr(D| +)	73.81%
Negative predictive value	Pr(~D| -)	74.47%
False + rate for true ~D	Pr(+|~D)	23.91%
False - rate for true D	Pr(-| D)	27.91%
False + rate for classified +	Pr(~D| +)	26.19%
False - rate for classified -	Pr(D| -)	25.53%
Correctly classified		74.16%

Figure 32
AUC and ROC curves
(Diversified set of variables model)



Question 20

Now, you got a job for providing credit to private firms. What weight would you give to the score of this firm for giving or not credit to this firm with respect to its business and financial plans and financial analysis?

Lending money is a profitable yet risky business. To keep it as much profitable as possible, it is important for the banker to assess the risks and allocate optimally its funding towards the most optimal firms: those that have the highest yields and that do not default. Finding the solution to the tradeoff between the potential profits and the firms' forecasted likelihood to default is the main task of credit providers. The banker has however some help coming from credit rating agencies (Moody's, Standard & Poor's and Fitch Ratings being the largest ones) which publish the creditworthiness of firms, ranging from AAA (prime grade) to D (default). He also gets help coming from the academic literature, as for example Altman's Z-score (which could have prevented investors losses in the Enron fiasco). Those tools provide some useful information based on the financial analysis of the firm. A financially healthy firm is awarded high ratings and is thus less likely to default in the following period, giving a positive signal to the banker to grant the credit loan to the healthy firm. Conversely, a financially unhealthy firm is scolded with low ratings, refraining the banker to lend the banks' money to this firm.

However, if lending money was as easy as just optimizing a threshold, many of us would be bankers! In practice, many factors must be taken into account: the firm's financial data, its business plan, the macroeconomic and financial conditions, public policies, etc. The banker also needs the analytical and econometric tools to build and improve the performance metrics of his models and deal with his bank's business model (aggressive or conservative lending) as well as his own risk aversion. Not all those factors can be favorable at the same time and some decisions need to be taken in the light of the available data, the forecasted default probability and the banker's understanding of the situation. If we take the example of a highly indebted firm asking for a credit but whose business plan seems credible enough for the banker to risk lending the firm the desired loan, the banker would poorly weight the credit rating of the firm. On the contrary, a financially healthy firm led by an incompetent board of directors may have easy access to credit due to its low likelihood of defaulting, but may find itself in a financial distress situation in a couple of periods of time.

The credit score of a firm is one of the many informational tools that helps the banker to assess -at a glance– the financial situation at time t of a given firm, yet other factors need to be processed to take the final lending decision. Ultimately, after reviewing the output of his models and his own knowledge, the banker will take the investment decision. Nevertheless, remember that bankers are humans, and humans make mistakes!

Question 21
Our own original estimations.

This question will be split in two sections and four sub-sections: we first explore the underlying trends in our dataset by conducting (i) an analysis on unsupervised learning through a principal component analysis (PCA), and then have a look at supervised learning by (ii) working on tree models, (iii) running a random forest and (iv) developing a Bayesian network.

I. Unsupervised learning: Principal Component Analysis (PCA)

Before running additional classification models and testing their performance, it is useful to firstly evaluate whether there exists any underlying trends in our data, which would enable us to refine our subsequent analysis. Therefore, we start this section of the analysis by running a Principal Component Analysis (PCA) on our entire dataset.

A PCA is indeed not a classification algorithm. It is rather an unsupervised learning tool that aims at identifying and visualizing patterns – such as clusters – in a dataset, and is particularly applicable for high-dimensional data (i.e. a dataset where the number of variables – or 'features' – "p" is greater than the number of observations "N". It does so by reducing dimensionality, namely by finding the smallest possible number of components that could capture the most variability possible in our data. This is the reason why we estimate it on the whole dataset. Although our dataset is not highly dimensional (15 variables for 180 observations), a PCA could still be very informative, given that we have a sufficiently big number of variables among which patterns can be identified.

The question we therefore ask is the following: regardless of the "financial distress variable" yd, can the observations be separated into distinct groups? Are there underlying patterns in our data that could underpin default trends? We therefore aim at building a model that will reduce our 14 continuous variables into as few principal components as possible – ideally 2, as we would like to find that our data could be clustered into a defaulting and non-defaulting group, without our interference – that would capture the most variability in our data. We therefore hope that our model will enable to identify clusters of observations in our data.

As we can only use continuous variables, we run our PCA model excluding the "default" binary variable. Our continuous variables are centered and scaled (mean = 0; units are in standard deviation change), so that the unit of measurement would not interfere with our results. We run the following line of code in R, to build our PCA for our entire dataset:

R-Code 1

```
PCA_T <- prcomp(Defaut2000[,-1], scale. = TRUE)
```

Our original (scaled) variables have now combined to form 14 linear combinations (or 14 principal components), in the following manner:

$$PC_n = a_{1,1}X_1 + a_{1,2}X_2 + \cdots + a_{1,14}X_{14} \quad (13)$$

where each principal component PC_n is explained by the following linear combination of our variables $X_1, \dots, X_{14}$

The principal components are numbered and ordered according to their ability to explain variability in the data (PC_1 explains the most variability in the data, followed by PC_2, and so on). We obtain the following results:

R-Code 2

Principal Component Analysis result – Standard deviations

```
> PCA_T
Standard deviations (1, .., p=14):
 [1] 2.0192781 1.7801001 1.1065323 1.0651735 1.0093764 0.9666117 0.9278657
 [8] 0.8141100 0.6783886 0.4804603 0.3558596 0.2572719 0.1450143 0.1138215
```

This first section related to the standard deviation of our data against one single PC, in other words of the variability of our data across that principal component. This section gives us little information apart from the fact that at first sight, the optimal number of principal components may be difficult to find, as we cannot see that the 2 or 3 first components capture a very significant portion of variability, while the rest capture nearly none of it. Instead, we see that many more components appear to capture significant amounts of standard deviation.

R-Code 3

Principal Component Analysis result – Rotation

```
Rotation (n x k) = (14 x 14):
               PC1          PC2         PC3         PC4          PC5
tdta     0.39605939  0.127432896 -0.05733711 -0.08127132  0.283236352
reta    -0.30979586 -0.222496706  0.01364912 -0.01298408 -0.268070716
opita   -0.29106442 -0.362512782 -0.29546764  0.08272757  0.052393896
ebita   -0.29868589 -0.332310342 -0.33605478  0.03846456  0.095515849
lsls     0.12623341 -0.455935053  0.21070247 -0.28372649  0.167645608
lta      0.10641636 -0.446334496  0.22650949 -0.33832936  0.178967735
gempl   -0.16590104 -0.091061190 -0.42345122 -0.24240436  0.211087805
invsls  -0.01164642  0.252928469 -0.38796024 -0.54747935  0.026566809
nwcta   -0.39140981  0.199589168  0.04206477 -0.16111534 -0.008835941
cacl    -0.41120239  0.130302662  0.26262201 -0.10098926 -0.045067499
qacl    -0.38825060  0.094143953  0.36104410  0.07178448  0.057837186
fata     0.08113511  0.005551888  0.10884348 -0.54572515 -0.636163413
ltdta   -0.13755304  0.076945710  0.36704467 -0.27156333  0.512964408
mveltd   0.14064958 -0.376065178  0.15200247  0.13626979 -0.239494970
```

```
               PC6          PC7         PC8         PC9          PC10
tdta    -0.26420505 -0.25171454  0.17583579 -0.10310458  0.090836780
reta     0.48839116  0.31051601 -0.11217647  0.17380130 -0.012115240
opita   -0.08805561 -0.13101051  0.33214395 -0.18428391 -0.062624423
ebita   -0.06059077 -0.13375437  0.38878850 -0.12892257 -0.084145380
lsls     0.09945792 -0.24727256 -0.09098538  0.19595927  0.160000039
lta      0.09490857 -0.22814710 -0.18704928  0.08192285 -0.172464694
gempl   -0.51653600  0.37307034 -0.42018977  0.30313802  0.074041926
invsls   0.34209278 -0.09913169 -0.22756800 -0.49044565 -0.181666638
nwcta    0.03308444 -0.29804054  0.04414203  0.09066523  0.738265266
cacl    -0.17968619 -0.22358277 -0.20610219 -0.20130827 -0.002127474
qacl    -0.29346396 -0.16336606 -0.14590097 -0.06412133 -0.471479940
fata    -0.31917684  0.08943028  0.38681169  0.10798995 -0.085610715
ltdta    0.06647727  0.56304068  0.37569210 -0.18351765  0.065243337
mveltd  -0.22740365  0.22900603 -0.25870473 -0.65790780  0.328407107
```

```
              PC11          PC12          PC13          PC14
tdta   -0.729339182  0.134575507 -0.040780192  0.045919154
reta   -0.621689873  0.124016259 -0.049072503  0.011625640
opita   0.051910260  0.018463698 -0.501270085  0.509459592
ebita  -0.059800054 -0.057194029  0.491999958 -0.481134446
lsls    0.043616322 -0.230828860 -0.467574758 -0.453941144
lta     0.077222934  0.217177599  0.462387777  0.438434463
gempl  -0.038384043  0.005382638 -0.009413389  0.006299734
invsls  0.055950242  0.077466571 -0.113604642 -0.103537456
nwcta   0.077605450  0.350870297  0.061426839  0.033525760
cacl   -0.229614377 -0.687950872  0.116583037  0.166909466
qacl   -0.044964317  0.488463123 -0.196705004 -0.248465274
fata    0.006270254  0.018299120 -0.012479206 -0.001923601
ltdta   0.042360933 -0.027472368 -0.010751836  0.019463096
mveltd -0.014996198  0.142643059  0.028998585 -0.081052837
```

For this second section, we see that within each PC, we have a value associated with each variable. These values are eigenvectors, namely the values by which our dataset variables are multiplied by to calculate the Principal Components score for any observation.

For instance, the vector for PC_1 would be:

$$(0.396 \times tdta) + (-0.309 \times reta) + (-0.291 \times opita) + \cdots + (0.140 \times mveltd) \quad (14)$$

and so on for the remaining PC_n. This is how the linear combination that forms principal components is derived. Calling the $Summary()$ function gives us more information about our principal components.

R-Code 4

```
summary(PCA_T)
```

As shown in the below table, we have, for each principal component:

- Standard deviation: the same result as above, in the first section of the PCA results.
- Proportion of variance: the proportion of all the variability in the original data explained by each single principal component. For instance, 34% of variance in the data is explained by PC1.
- Cumulative proportion: taking PC1 and PC2 together, we can explain 57% of the variability in the data.

R-Code 5

Principal Component Analysis – Summary

```
> summary(PCA_T)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.0193 1.7801 1.10653 1.06517 1.00938 0.96661 0.9279
Proportion of Variance 0.2913 0.2263 0.08746 0.08104 0.07277 0.06674 0.0615
Cumulative Proportion  0.2913 0.5176 0.60505 0.68609 0.75886 0.82560 0.8871
                           PC8     PC9     PC10    PC11    PC12   PC13    PC14
Standard deviation     0.81411 0.67839 0.48046 0.35586 0.25727 0.1450 0.11382
Proportion of Variance 0.04734 0.03287 0.01649 0.00905 0.00473 0.0015 0.00093
Cumulative Proportion  0.93444 0.96731 0.98380 0.99284 0.99757 0.9991 1.00000
```

The “Cumulative Proportion” line of this table suggests that it is unlikely that we can explain most of the variability in our data using a small number of component (such as 2 or 3). For instance, we need to use at least 9 components to explain 95% of the variability in our data. This a first hint that we are unlikely to find any significant pattern in our data.

We plot a line chart of this cumulative variance in order to identify whether we can find a number of components where (1) most of the variability in our data is explained and (2) adding an additional PC does not explain significantly more variance.

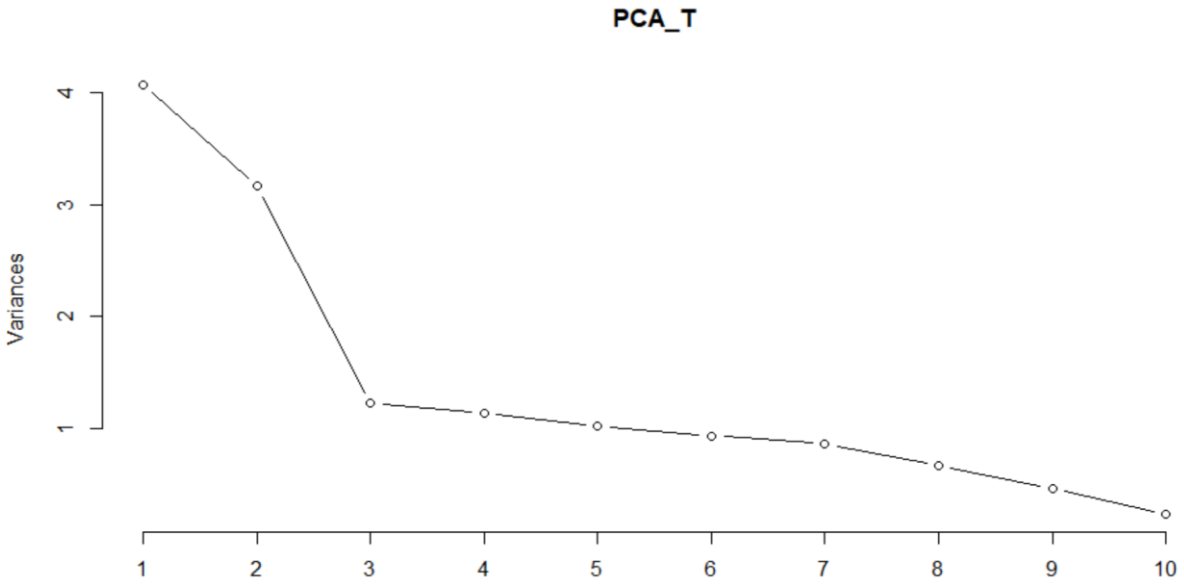
R-Code 6

```
plot(PCA_T, type="l")
```

Although we notice a sharp drop in explained cumulative variability after the 2nd PC, those two explain only nearly 60% of the total variance. We would therefore need many more components to account for the remaining 40%. However, for the sake of the intended dimensionality reduction, we will stick with those two components, and further inquire how they help us identifying patterns in our data.

Figure 33

Explained cumulative variance



Now that we have chosen two Principal Components, we have a first look at the correlation between each variable in our dataset and the estimate two Principal Components.

R-Code 7

Correlations between variables andthe Principal Components.

```
> cor(Defaut2000[,-1], Total_DefautData[,16:17])
                PC1          PC2
tdta     0.79975404  0.226843318
reta    -0.62556400 -0.396066419
opita   -0.58774001 -0.645309056
ebita   -0.60312988 -0.591545688
lsls     0.25490035 -0.811610054
lta      0.21488423 -0.794520102
gempl   -0.33500033 -0.162098037
invsls  -0.02351736  0.450238004
nwcta   -0.79036525  0.355288706
cacl    -0.83033198  0.231951789
qacl    -0.78398592  0.167585664
fata     0.16383435  0.009882917
ltdta   -0.27775785  0.136971070
mveltd   0.28401062 -0.669433679
```

We obtain a visual representation of the information provided in the R-Code 7 by running the following line of code.

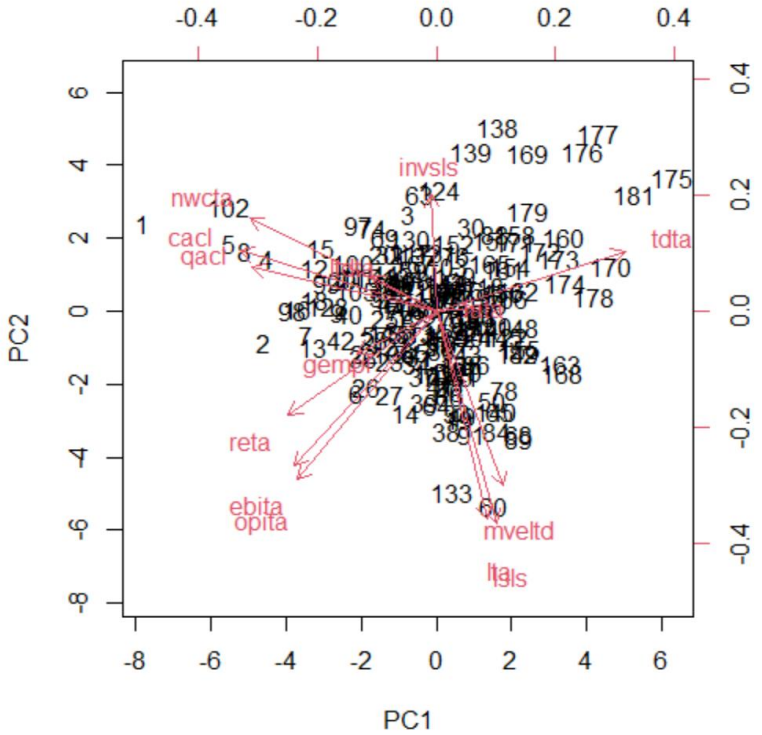
R-Code 8

```
biplot(PCA_T, scale=0)
```

We to obtain a plot that helps us visualize the relationship between the different variables and the principal components. Although the results displayed in the chart below make sense economically, it is still difficult to identify patterns in our data.

Figure 34

Observations against PC_1 and PC_2



Each number represents one observation from our data frame. Each observation is plotted with its values for principal component 1 and principal component 2. The red arrows represent the eigenvectors for each variable in our data frame.

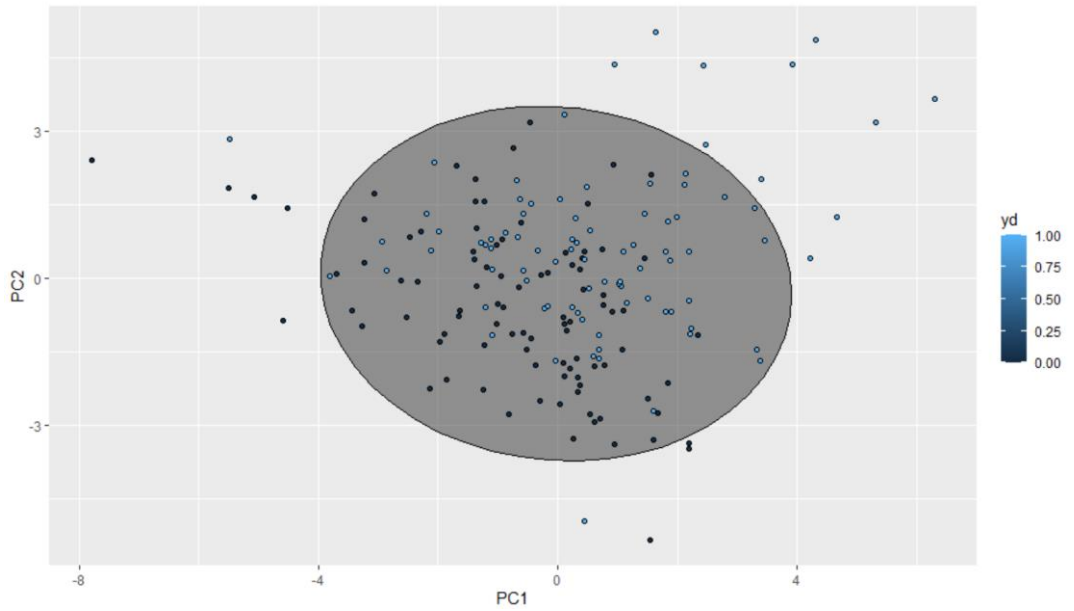
As an observation has a higher value for PC_1, it has a lower $reta$, $ebita$, $opita$, $nwcta$, $cacl$, $ltdta$, $qacl$ and a high $mveltd$, lta, $lsls$, $fata$, $tdta$ and no real difference for $invsls$.

As an observation has a higher value for PC2, it has a lower $reta$, $ebita$, $opita$, $mveltd$, lta, $lsls$ and $gempl$ and a high $invsls$, $tdta$, $nwcta$, $cacl$, $ltdta$, $qacl$ and no real difference for $fata$.

We can indeed see that variables which are highly correlation ($ebita$ and $opita$, or $cacl$ and $qacl$ for instance) have the same correlation direction with the two principal components. However, the second result that immediately stands out is that observations are much centered. This suggests that those two Principal Components fail to cluster our data. This intuition is confirmed by the below plot, where we color our observations according to whether they are financially distressed or not.

Figure 35

Observations against "yd" group membership



The above chart also maps each observation according to the two principal components, on top of which we have added a 95% confidence ellipse. The chart shows that the PCA – i.e. condensing our 14 variables into 2 different linear combinations – did not manage to identify a pattern in our data that enables us to separate our observations according to their financial distress status.

Conclusion of our Principal Component Analysis:

The PCA did not manage to find significant underlying patterns in our data. We therefore fail to find relationships between our variables that would enables us to separate our observations according to their financial distress status.

We therefore complement this PCA with a cluster analysis of our dataset, to confirm the difficulty to cluster our dataset between defaulting and non-defaulting firms.

Cluster analysis aims at finding patterns in data in the form of "clusters" – namely sets of observations that are more similar to each other than they are to observations belonging to other clusters. For instance, in our firm data, without any prior information on the structure of our data, we may perform cluster analysis to determine if there are groups of individuals which are more similar to each other, which may signify the presence of underlying patterns.

We will use a K-means algorithm, as it is the simplest of all tools for performing cluster analysis. It is the most used clustering method for splitting a dataset into a set of K groups. We specify it a value of 2, as we know we have 2 categories ($y = 0$ and $y = 1$, namely financially distressed and non-distressed firms). We run the following line of R code:

R-Code 9

```
FitK <- kmeans(Defaut2000_Scaled, 2)
```

R-Code 10

K-means clustering

```
> FitK
K-means clustering with 2 clusters of sizes 71, 110

Cluster means:
        tdta       reta      opita      ebita       lsls        lta       gempl
1 -0.6856119  0.5004991  0.3790061  0.3974390 -0.4963284 -0.4744245  0.2440820
2  0.4425313 -0.3230494 -0.2446312 -0.2565288  0.3203574  0.3062194 -0.1575438
       invsls      nwcta       cacl       qacl       fata      ltdta     mveltd
1  0.08837556  0.8257027  0.7961676  0.7104068 -0.1700731  0.2713601 -0.4855583
2 -0.05704241 -0.5329535 -0.5138900 -0.4585353  0.1097744 -0.1751506  0.3134058

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 2 1 1 1 1 2 1 1 1 1 1 2 2 1 2 2 1 1 2 2 2
 [40] 1 2 1 1 1 2 2 2 1 2 2 2 1 2 1 2 1 1 2 1 2 2 2 1 2 2 2 2 2 1 1 2 1 2 1 2 2 2 2
 [79] 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 2 1 1 1 2 1 1 2 1 2 2 1 1 1 1
[118] 2 2 2 2 2 1 1 1 2 2 1 2 1 1 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[157] 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1]  826.3945 1259.0493
 (between_SS / total_SS =  17.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

The algorithm returns us 2 clusters of sizes of 71 and 110 observations. The important point on the above table is that this method for assigning cluster membership accounts for little of the variability in our data: the ratio of "Between Sum of Squares" / "Total Sum of Squares" is of 17.2% only.

The "Between Sum of Squares" is the sum of squares between clusters (i.e. how well the clusters spatially distanced from each other). We divide it by the Total Sum of Squares, namely the total variability in our data. We find that, with a ratio of 17.2%, the model has very low accuracy if we rely on our financial distress dummy variable.

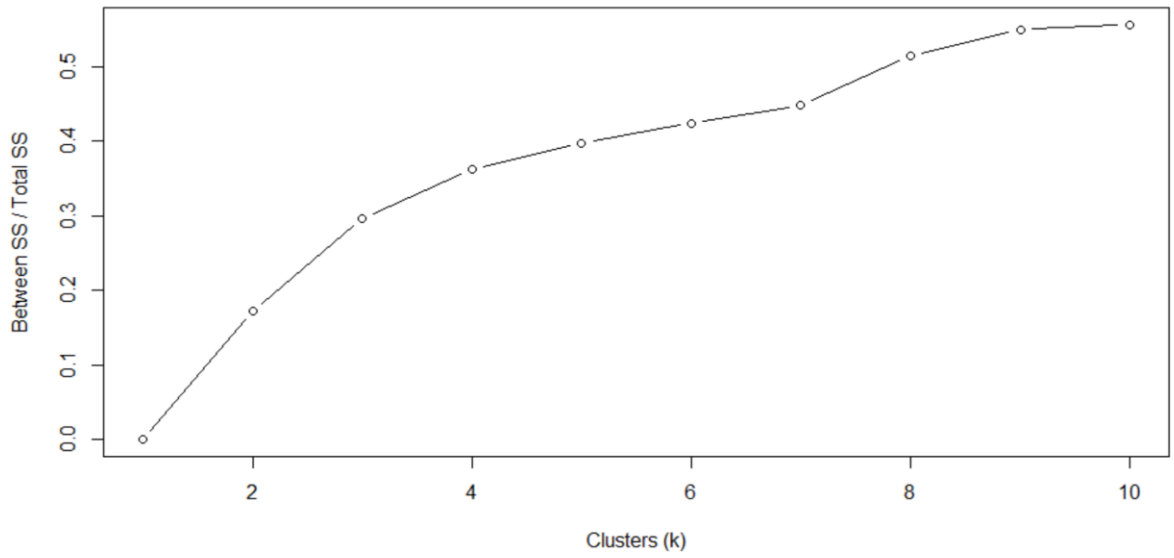
The low performance of the previous model, which relied on $K = 2$, pushes us to ask the question of the appropriate number of clusters. We therefore run the following line of code, in order to plot the number of K against the "Between SS / Total SS" ratio, to identify which number of clusters K is the most appropriate for our model.

R-Code 11

```
plot(1:10, betweenss_totss, type = "b",
     ylab = "Between SS / Total SS", xlab = "Clusters (k)")
```

Figure 36

Optimal number of K clusters



What is immediately visible from this chart is that:

1- Even choosing a high number of clusters (10) does not allow us to explain significant variability in our data (we hardly go above 50% of explained variability).
2- Moreover, we see no "shoulder effect" in our curve: in other words, we do not see a sharp increase in explained variability which would allowing us to identify the right number K immediately.

Therefore, this result confirms that no significant pattern emerges from our data. We can visually depict the low performance of this model based on the below graph, with a number of clusters set to $K = 2$.

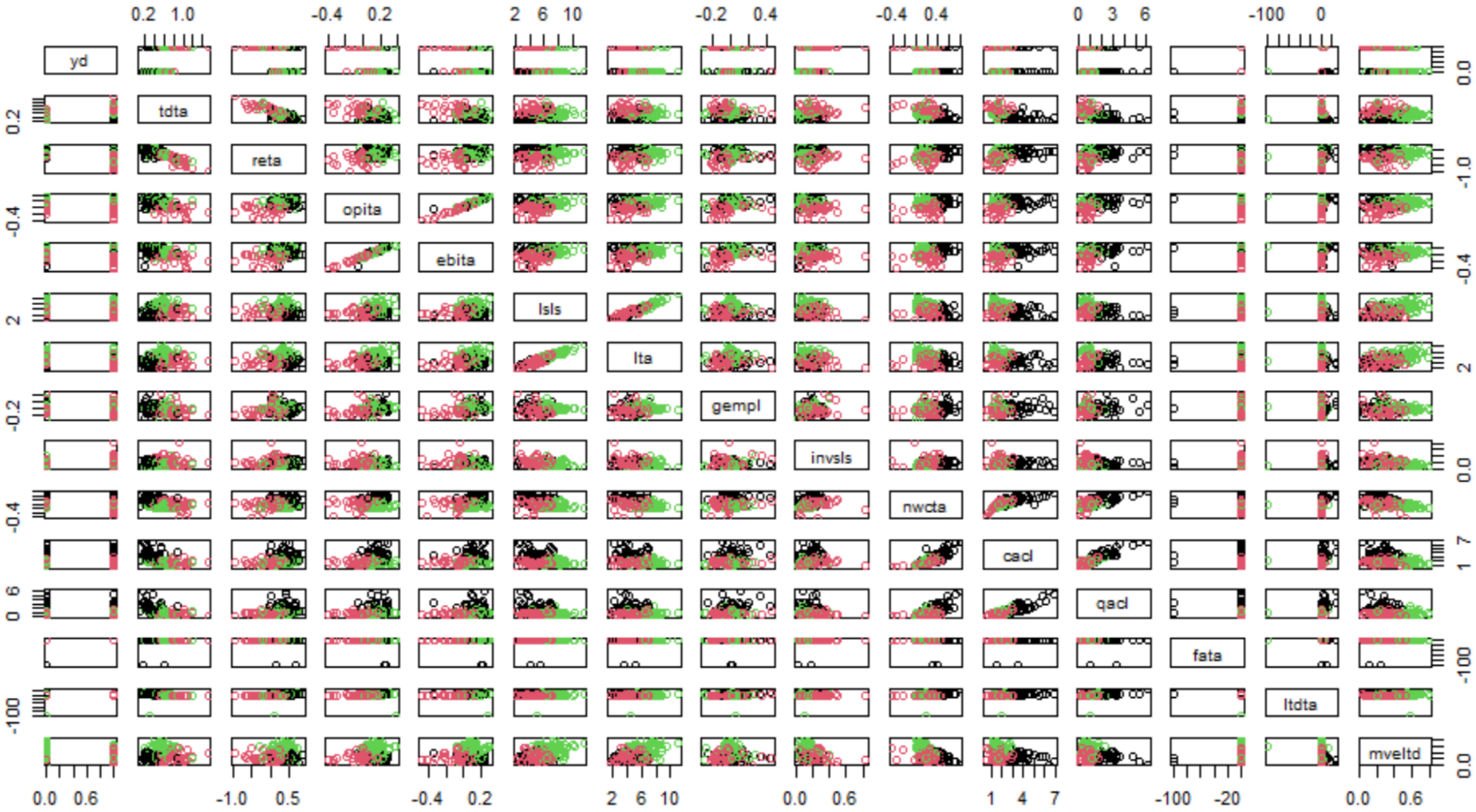
In particular, the charts for yd depict the failure of our model to cluster our data between $yd = 0$ and $yd = 1$. As we can see in the first line and in the first column of charts, there are observations of the "red cluster" mixed up with the "green" one, while we ideally would have seen only green for $yd = 0$, and only red for $yd = 1$.

Conclusion of our K-means clustering analysis:

The model fails to cluster our data – and more generally to identify underlying patterns – regardless of the number of clusters that we allow the model to have. This could be attributed to the relatively small size of our dataset. Using a larger dataset could have yielded more significant results.

Figure 37

Optimal number of K clusters



II. Supervised learning: Tree Model, Random Forest and Bayesian network

As we failed to identify patterns underlying our data based on those models, we examine this dataset through a different angle, and put again the financial distress variable at the center of our analysis. We therefore turn to supervised learning algorithms, as we now have one response variable. We will apply three classification techniques: (i) a tree model, (ii) a random forest model and (iii) a Bayesian network.

First, to capture the potential improvements obtained by our models, we laid down a simple logit model that will be used as a benchmark for the rest of this development. To build this model, we decided to use all available explanatory variables of our data set, i.e., *tdta, reta, opita, ebita, lsls, lta, gempl, invsls, nwcta, cacl, qacl, fata, ltdta* and *mveltd*.

Table 34
Simple logit model

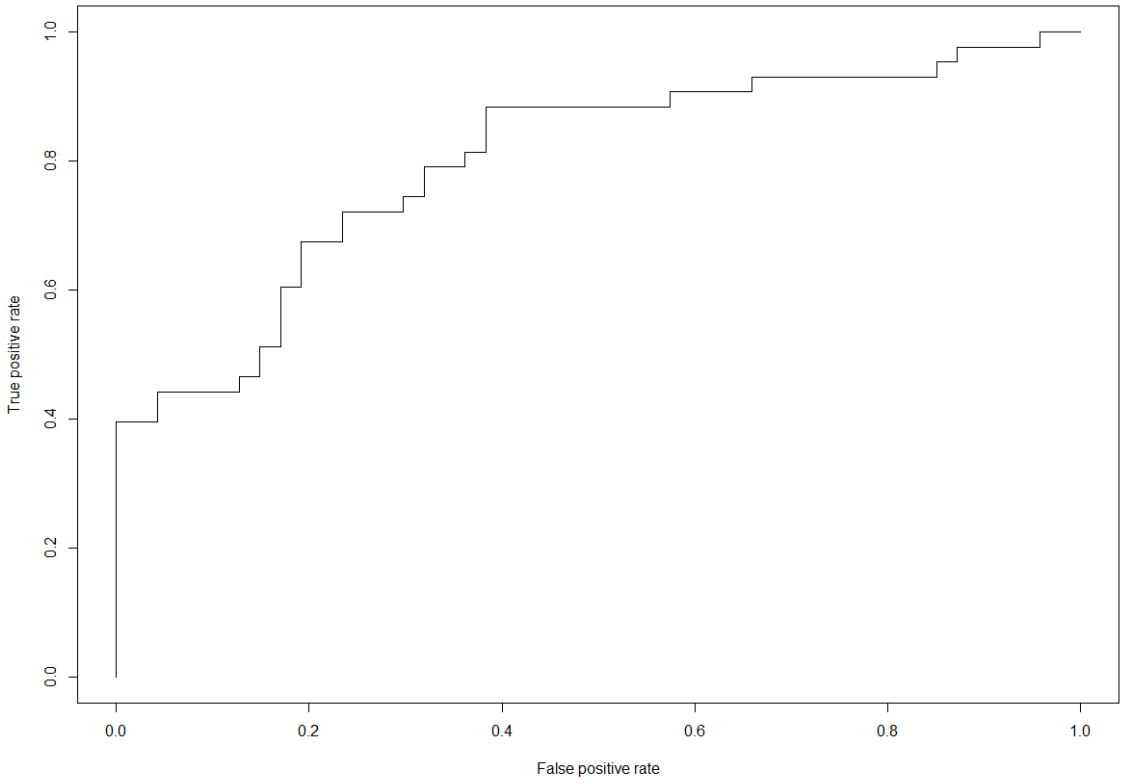
	Model 1		
(Intercept)	0.70098		
	(4.36934)		
tdta	-1.53231	*invsls*	17.23997
	(5.02324)		(9.61283)
reta	-4.10528	*nwcta*	-6.66078
	(2.65803)		(6.47751)
opita	-21.74981	*cacl*	-1.54211
	(11.83998)		(1.18473)
ebita	18.52437	*qacl*	3.46873
	(10.54735)		(1.79820)
lsls	1.62383	*fata*	1.18893
	(1.34503)		(3.79034)
lta	-1.58976	*ltdta*	-0.55965
	(1.34998)		(0.33548)
gempl	-8.19632 *	*mveltd*	0.76295
	(3.48838)		(3.35064)
N	91	BIC	145.34872
AIC	107.68583	Pseudo R2	0.54875

*Note: *** p < 0.001; ** p < 0.01; * p < 0.05.*

Figure 38

ROC curve

Simple logit model



Our model is statistically significant (($X^2(14) = 48.192$, $p = 0.000$) with an associated Pseudo R^2 (Cragg-Uhler) of 0.524. However, it seems that our results obtained with R are slightly different from the ones obtained through STATA in Question 16. Here again, most of the individual β estimates are not significant except for *gempl*. Although we find a slight difference with the previous estimation, the value of the area under the ROC curve (Figure 39) is 0.80, which is in line with we had previously found and confirms the strong predictive potential of the variable *gempl*.

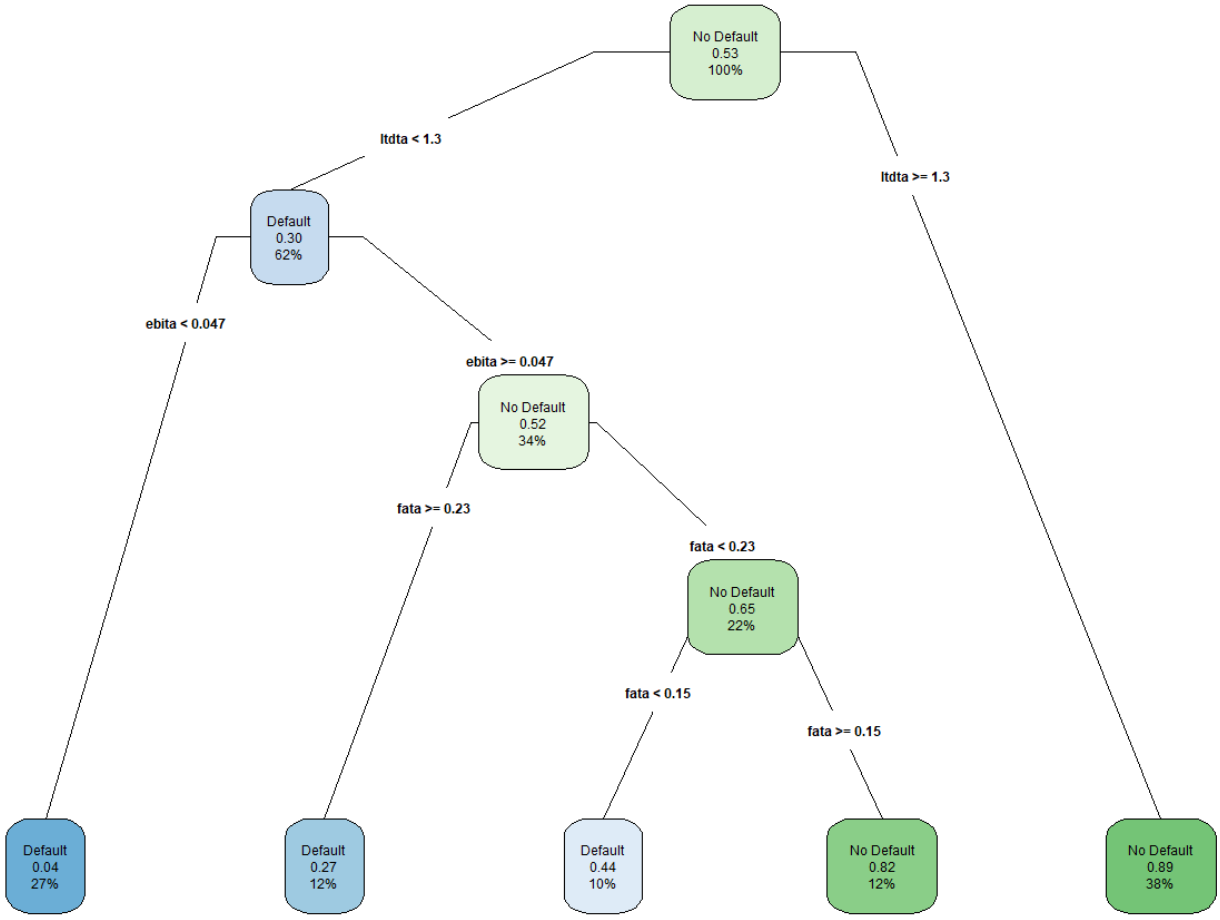
i. Tree models

As our journey through the woods begins now, we decided to start our walk by implementing a basic recursive partitioning tree.

A recursive partitioning process generates a decision tree that aims to classify members of a given population by splitting it into sub-populations based on several *criteria*. Therefore, by creating a tree model, we can predict the value of our outcome variable based on our explanatory variables. This process is entitled "recursive" since each sub-population could be split up an indefinite number of times until the process terminates after a peculiar stopping *criterion* is reached. Thanks to the recursive partition we can create a set of rules based on individuals' characteristics that allows us (i) to split our sample into sub-categories and (ii) to compute the probability of occurrence of an outcome given the level of X parameters at each nod and for each leaf.

Before we start chopping some wood, we need to identify two categories of trees: classification and regression trees. For the first category, the predicted outcome is a discrete variable whereas for the second category, the predicted outcome is a real number. We will build both models in our analysis, starting with a classification tree.

Figure 39
Recursive Partitioning Tree
Classification tree



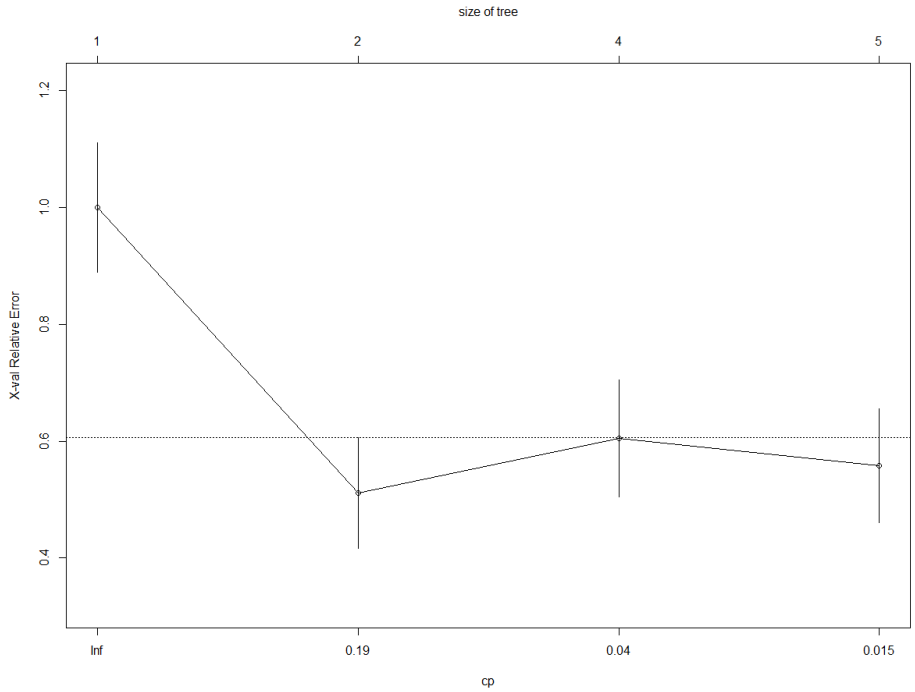
Thanks to the *rpart.plot* function we know that each node shows the predicted class (Default, No Default), the predicted probability of not defaulting and the percentage of observations in the node.

As shown on the first node, it seems that the *criterion* that helps dividing our sample the most is the long-term debt over total asset ratio. Indeed, after only one split, the rightmost node contains a sub-population of almost 40% with an associated probability of not defaulting of 0.89. Therefore, we could assume that this subset is reasonably homogenous. When talking about homogeneity, we imply that the sub-population, with respect to our outcome variable, is composed of mostly non-defaulting firms as proven by its associated probability.

The other two financial ratios used to split our sample are *ebita* and *fata*. Thanks to the rules generated by our model, we now know that companies with a *ltdta* < 1.3 and an *ebita* < 0.047 represents 27% of our sample and that they have a default probability of almost 1 (0.96 to be more precise). Surprisingly enough, with the *fata* ratio, we can create two other rules that split our sample into three distinct groups. For instance, if *fata* is included between]0.23;0.15] then we can create a homogenous group of mostly non-defaulting firms (0.82). However, if *fata* < 0.15, we end up creating a more heterogeneous subset of firms with respect to the default outcome (associated probability of default of 0.44).

Figure 40

Graphical representation of the matrix of information on optimal pruning given complexity parameters



R-Code 12

Matrix of information on optimal pruning given complexity parameters

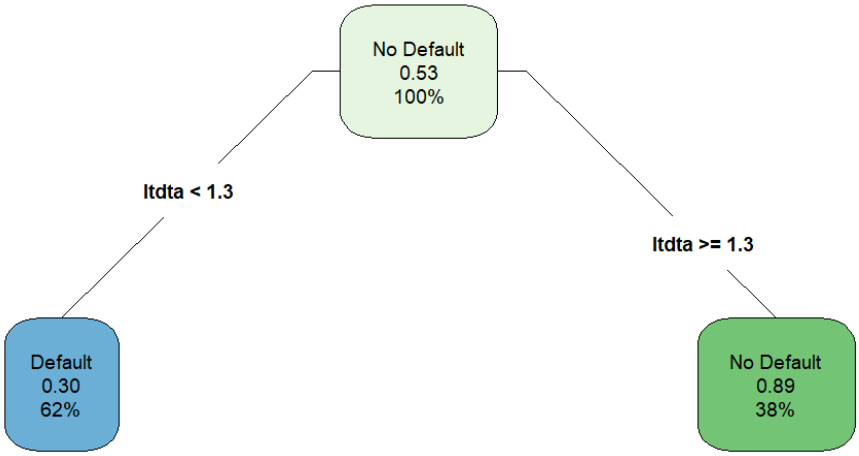
```
Classification tree:
rpart(formula = yd ~ ., data = Estim)
Variables actually used in tree construction:
[1] ebita fata  ltdta
Root node error: 43/91 = 0.47253
n= 91
       CP  nsplit   rel error xerror   xstd
1 0.511628      0   1.00000 1.00000 0.110756
2 0.069767      1   0.48837 0.51163 0.094983
3 0.023256      3   0.34884 0.60465 0.100220
4 0.010000      4   0.32558 0.55814 0.097758
```

One of the main criticisms made to the recursive partitioning tree is its tendency to overfit the data. To avoid such an issue, we must validate our model by using the complexity parameter and the cross-validated error.

On the Pruning plot, the complexity parameter values (CP) are plotted against the cross-validated error. The dashed line represents the highest cross-validated error minus the minimum cross-validated error, plus the standard deviation of the error at that tree. In order to avoid the aforementioned problem, we select the complexity parameter that minimize the cross-validated error to prune our tree. Either from the list of CP values or from the Pruning plot, we can state that the parameter having the least cross-validated error is $CP = 0.069$. Thus, we can now build our pruned tree.

Figure 41

Pruned tree

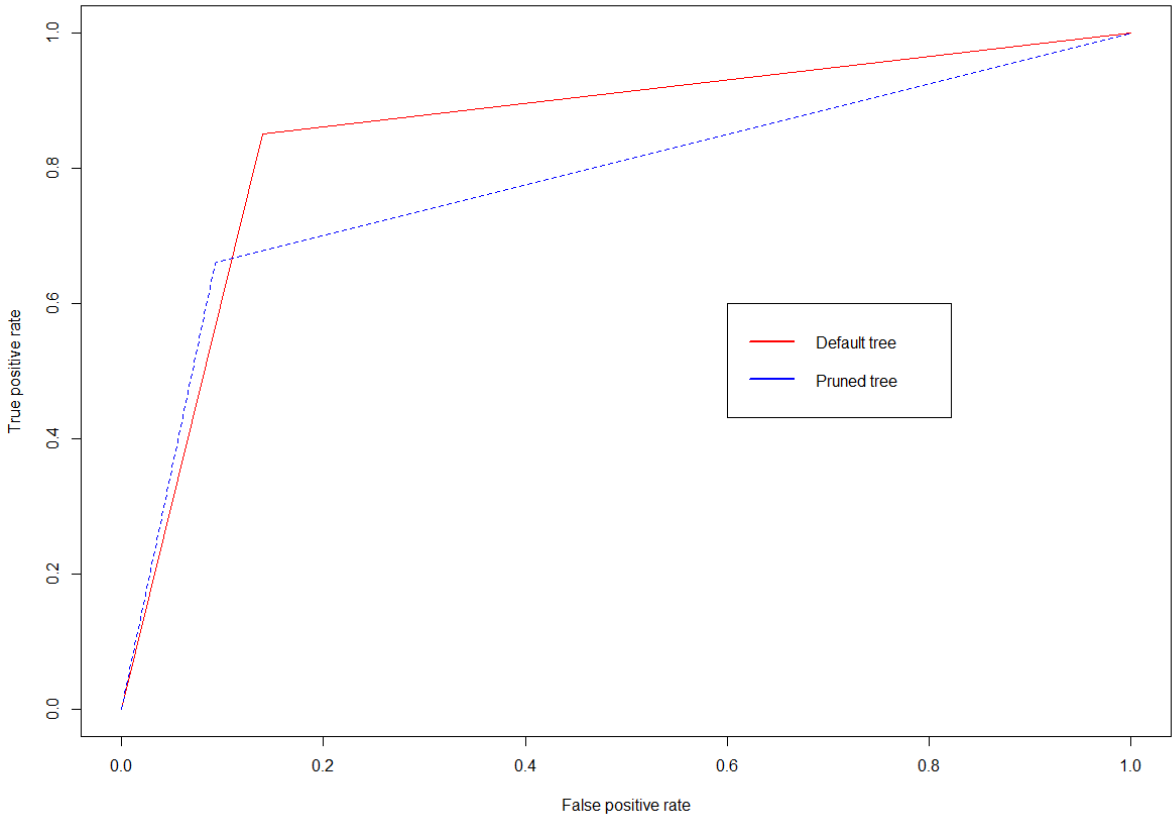


As one can notice, the number of leaves has been drastically reduced. However, our model still creates two coherent subsets of population thanks to the *ltdta* parameter. Indeed, we obtain a first group that included almost 40% of the sample with an associated probability of non-defaulting of 0.89, the remaining 62% of the sample have them an associated probability of defaulting of 0.7. By pruning our tree, we might prevent the overfitting of our data, and we can now move to evaluating the fit thanks to ROC curves.

From the graph hereunder, we conclude that the default tree ROC curve stochastically dominates the pruned tree ROC curve (0.855>0.783). Therefore, we can reject the hypothesis of the overfitting of our data by our default tree model since it predicts in a more accurate manner the results of our testing sample than our pruned model.

Figure 42

ROC curves: Default vs pruned tree



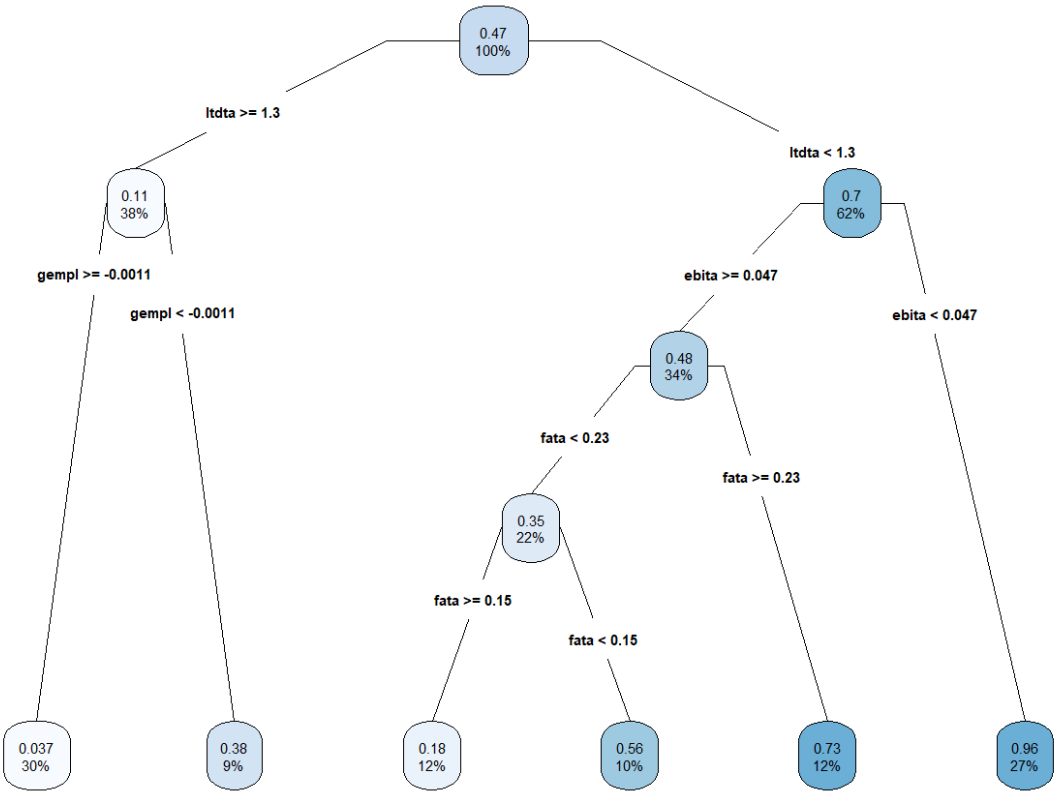
As our little walk through the woods continues, we encounter a new species of tree: a regression tree.

The first thing we can notice is that it has (i) more leaves and (ii) it is the opposite symmetry of our classification tree. The new variables used, *gempl*, allows us to create a new rule that generate a large (30% of the sample) homogenous group of non-defaulting firms (associated probability of default: 0.087) and a small (9% of the sample) more heterogenous sub-sample of firms. One may notice that by adding a new variable we create more homogenous sub-categories on the extremity of the tree, but we end up increasing the heterogeneity of the middle leaves and also decreasing their size.

Figure 43

Recursive partitioning tree

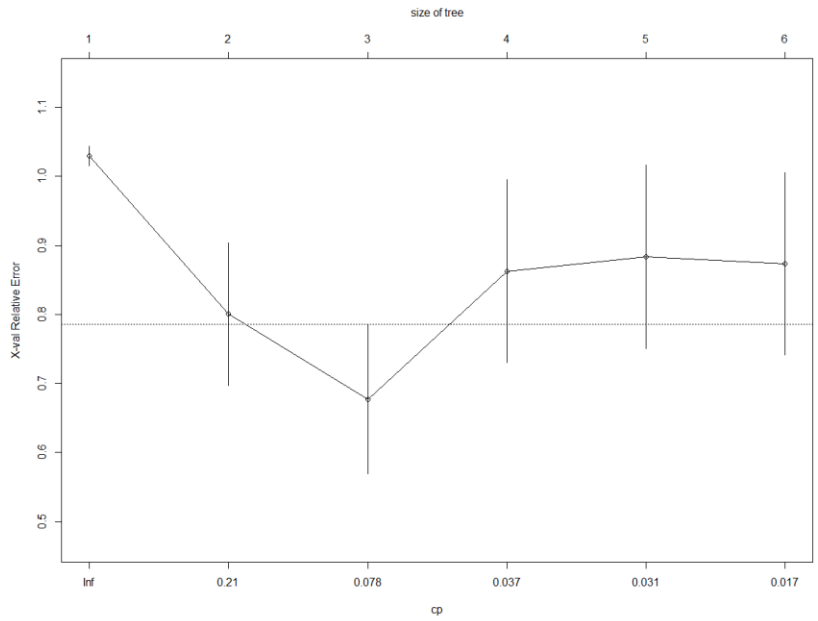
Regression tree



Once again, we need to check for any overfitting matters by studying the Matrix of information on optimal pruning given complexity parameters.

Figure 44

Graphical representation of the matrix of information on optimal pruning given complexity parameters



R-Code 13

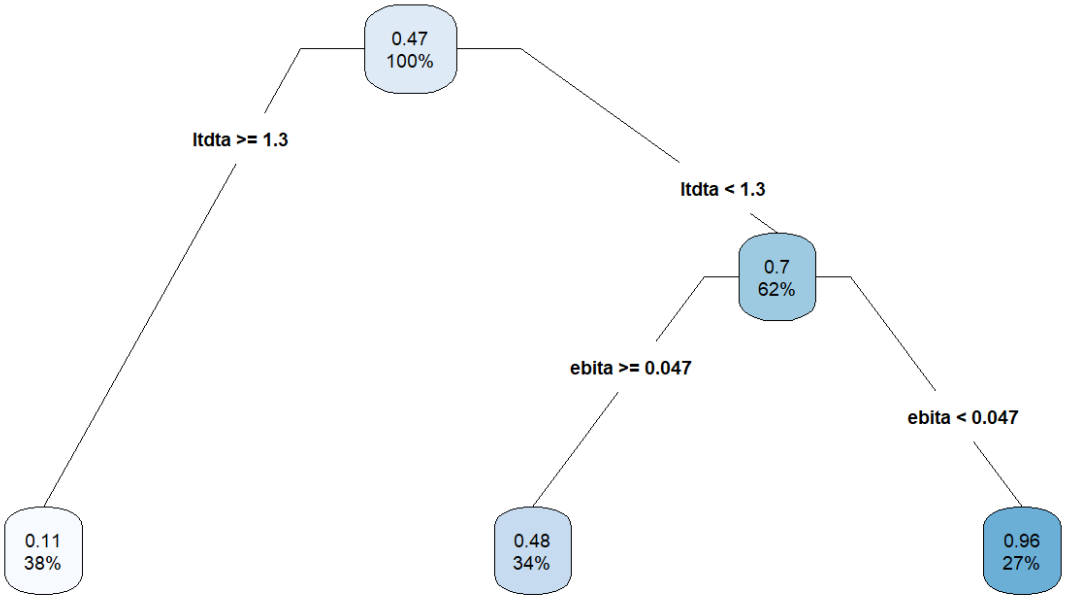
Matrix of information on optimal pruning given complexity parameters

```
                    Regression tree:
rpart(formula = yd ~ ., data = Estim, method = "anova")
     Variables actually used in tree construction:
          [1] ebita fata  gempl ltdta

      Root node error: 22.681/91 = 0.24925

                       n= 91

          CP nsplit  rel error xerror    xstd
1 0.321814      0   1.00000 1.02909 0.014174
2 0.138323      1   0.67819 0.80053 0.103132
3 0.044535      2   0.53986 0.67744 0.108644
4 0.031078      3   0.49533 0.86273 0.132193
5 0.030484      4   0.46425 0.88341 0.132479
6 0.010000      5   0.43377 0.87305 0.131966
```

In order to avoid the aforementioned problem, we select the complexity parameter that minimize the cross-validated error to prune our tree. Either from the list of CP values or from the Pruning plot, we can state that the parameter having the least cross-validated error is $CP = 0.044$. Therefore, we can now build our pruned tree.

Figure 45

Pruned tree

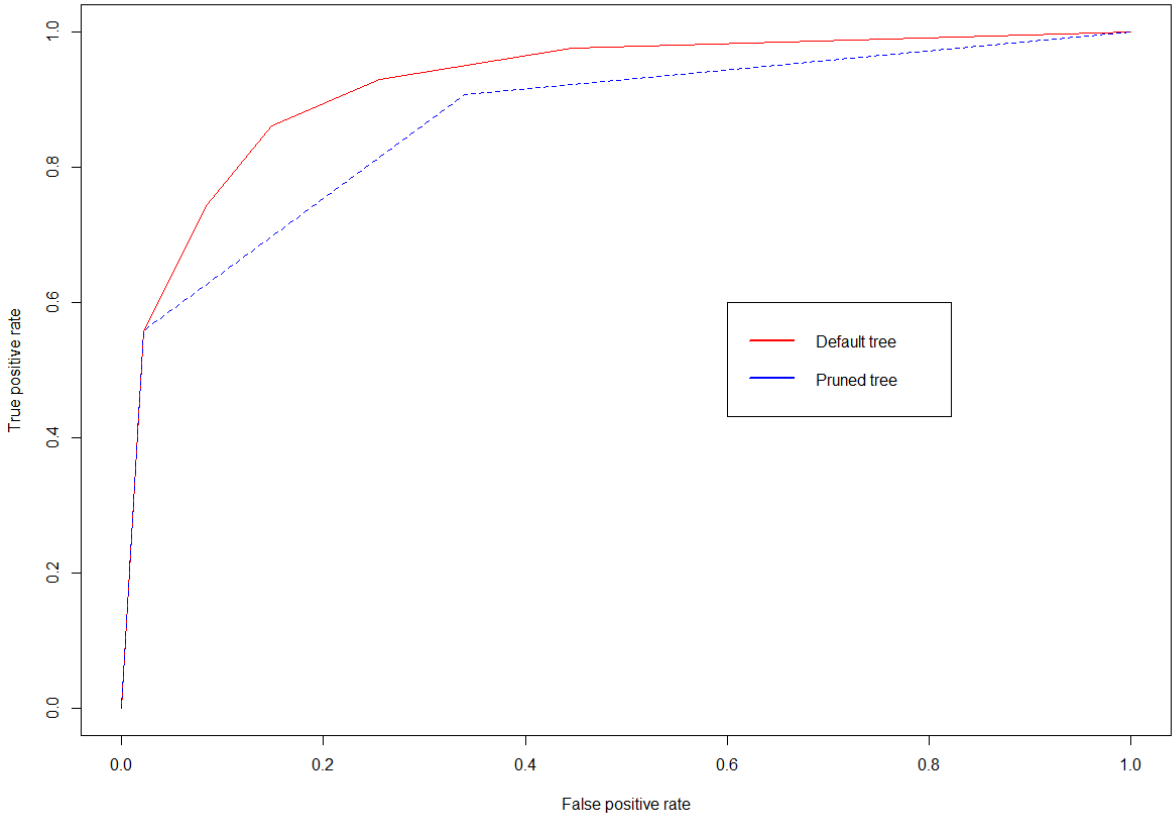


The number of splits associated with our complexity parameter is higher than our classification model, therefore, we allow for more leaves and nodes to be created. One may notice that we mostly obtain the same sub-populations as the ones from our classification tree meaning that, in our case, classification and regression trees do not really differ from each other.

Let us now move on to evaluating the fit of our regression tree model.

Figure 46

ROC curves : default versus pruned tree



Here again, the default tree ROC curve stochastically dominates the pruned tree ROC curve (0.923>0.868). Therefore, we can also reject the hypothesis of the overfitting of our data by our default tree model since it predicts in a more accurate manner the results of our testing sample than our pruned model. Moreover, we can also state from all our previous results that, for our study, the regression tree is the best option to choose when it comes to implementing a recursive partitioning model.

Just before exiting the underwood, we encounter a last variety of tree that triggered our curiosity: the conditional inference trees.

Conditional inference tree also known as unbiased recursive partitioning, is a non-parametric class of decision trees used for continuous and multivariate response variables in a conditional inference framework. Conditional inference trees use significance test to select input variables whereas conditional trees use the Gini coefficient for selecting the variable that maximizes the information measure.

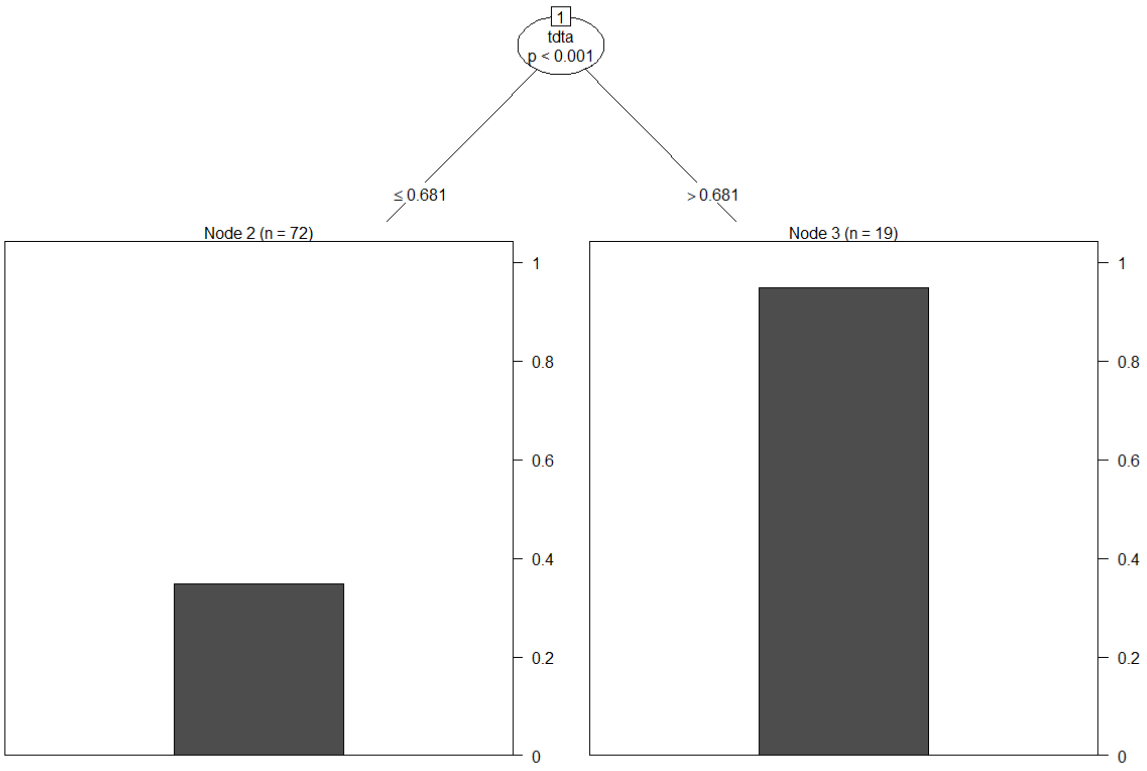
R-Code 14

Conditional inference model

```
        Conditional inference tree with 2 terminal nodes


Response:  yd

Inputs:  tdta, reta, opita, ebita, lsls, lta, gempl, invsls, nwcta, cacl, q
acl, fata, ltdta, mveltd

Number of observations:  91


1) tdta <= 0.6812963; criterion = 1, statistic = 24.223

  2)*  weights = 72

1) tdta > 0.6812963

  3)*  weights = 19
```

Figure 47

Conditional inference tree



The code from the package "*party*" allows us to produce a graph of a conditional inference tree that displays the value of *yd* in a form of a bar plot in each node. The only *criterion* retained by the algorithm to build our model is the *total debt over total asset* ratio with a critical value of 0.681. On the one hand, the first node contains 72 firms with a

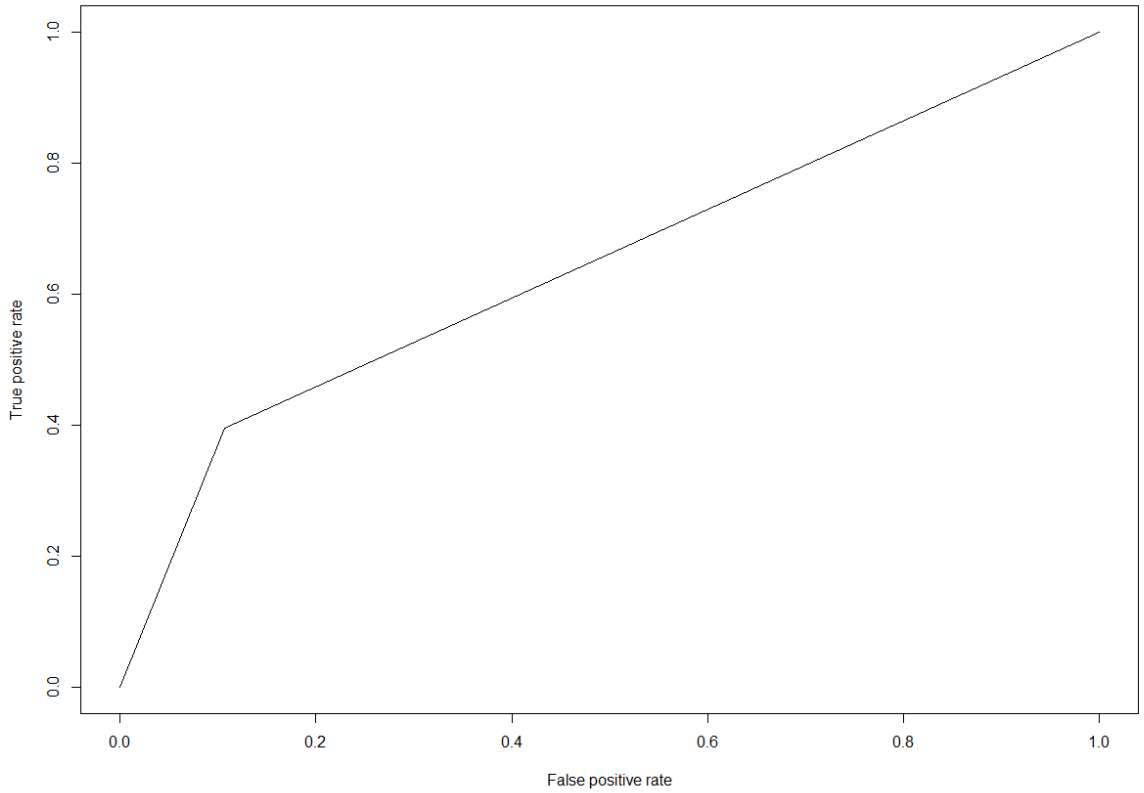
$tdta \leq 0.681$, graphically we can state that this node includes *mostly* firms in good financial health. Whereas on the other hand, the second node contains practically only firms in bad financial health.

However, we can question the validation of this model as it obviously lacks explanatory variables to operate the splits in our sample. Moreover, the average value of *yd* for the first node's firms remains too high but before jumping to conclusion, let's first study the ROC curve of our model.

Figure 48

ROC Curve

Conditional inference model



With an associated area of "only" 0.644, our conditional inference tree is our less explanatory model. This could be explained by the lack of specification of this model: by using only one *criterion* it fails to generate coherent sub-populations and therefore fails to predict the value of *yd* in our validation sample.

ii. Random Forest

As our curiosity grew unsatisfied, we continued our journey through the woods until we ended up in the Random Forest region.

The first Random Forest algorithm was developed in the 1990s. Its main purpose is to classify a sample population given its intrinsic characteristics or features by using a multitude of individual decision trees that operate as an ensemble. By constructing a large number of uncorrelated trees, Random Forest algorithms produce more accurate predictions than individual trees.

A low correlation between each individual tree is one of the key components to ensure a better predictability. Thanks to what is called bagging or bootstrap aggregation (assigning a random training set of the same size for each tree that includes the original training set and some replacement), trees will remain well rooted and will not co-move in the same direction, therefore protecting each other from their individual errors.

Another main aspect of the Random Forest algorithm is that it uses "Feature Randomness". In a regular decision tree model, for each node, the algorithm selects the best possible feature that produces the most separations between the observations. In the case of the Random Forest algorithm, each tree can only pick a random subset of features. Therefore, Feature Randomness generates more variations across the trees and eventually reduces the correlation between trees.

Now that we have finished the topography of our Random Forest, we can apply it to our sample by using the marvelous *randomForest* and *randomForestExplained* packages. Recall that, thanks to bagging, we can use our whole sample to generate our model and that each model is by essence unique and therefore non-reproducible.

R-Code 15

Random Forest short summary

```
Call:
 randomForest(formula = yd ~ ., data = Defaut2000_cleaned, importance = TRU
E,      proximity = TRUE, ntree = 500, keep.forest = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 0.1869226
                    % Var explained: 25.05
```

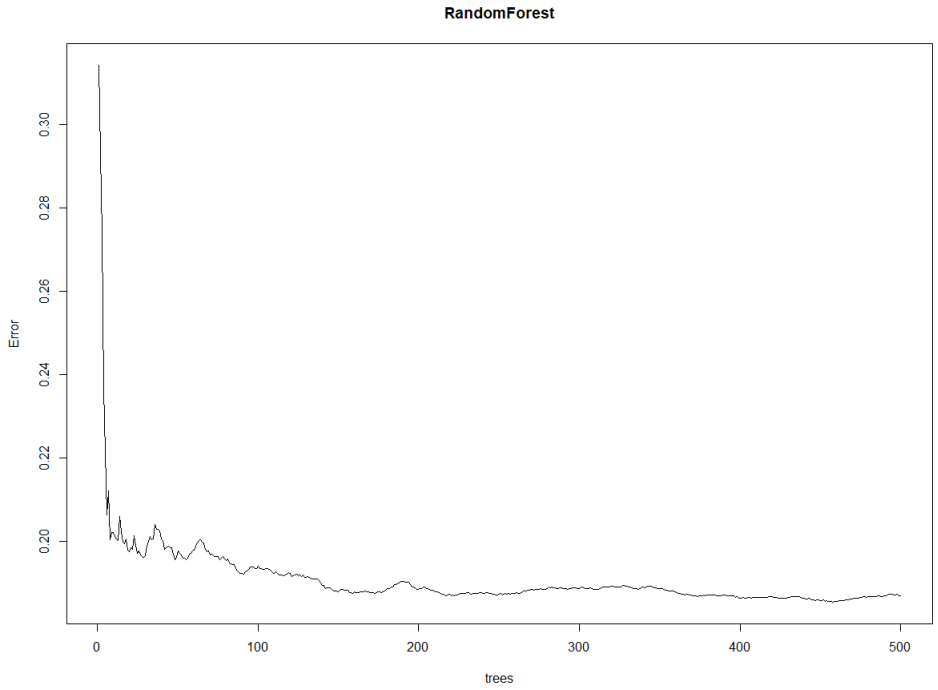
Remember that, since we use a dummy variable to express financial distress, the algorithm will automatically consider using a regression model instead of a classification model. Therefore, we constructed our Random Forest with the help of 500 decision trees, which is the default parameter. We tried to generate models with less trees, but we obtained results that are more accurate by increasing the number of decision trees. Regarding the number of variables tried at each split, we also kept the default parameter which is automatically computed

by the algorithm: "For regression models, it is the number of predictor variables divided by 3 (rounded down)."

The %Var explained is a measure of how well out-of-bag predictions explain the variance of the training set. We find a %Var of 25.05% which is quite low. Unexplained variance could be due to true random behaviour or lack of fit. We will conduct some robustness checks taking into account unexplained variance in the following sections. In particular by finding the optimal number of trees for our model as too many trees might overfit our model.

Figure 49

Number of trees versus Mean of squared residuals



Graphically we can see that the MSR keeps decreasing overall and reach its low point around 450. With the formula *which.min(model$mse)*, we established that the optimal number of trees for our model was 458. Now that we found the best spot of the Random Forest to settle our camp, we can continue our little stroll.

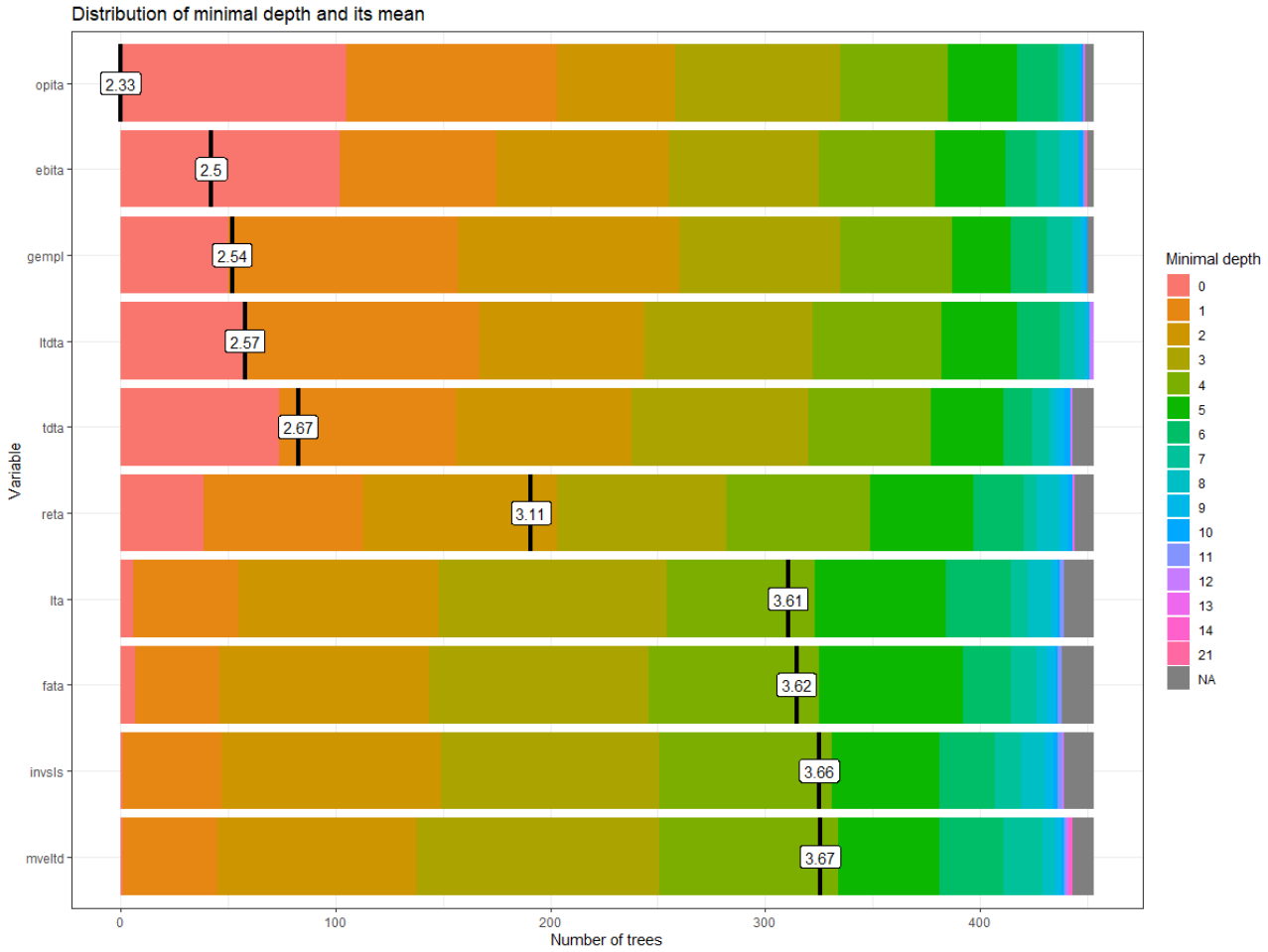
R-Code 16

Random forest short summary

```
Call:
 randomForest(formula = yd ~ ., data = Defaut2000_cleaned, importance = TRU
E,      proximity = TRUE, ntree = 458, keep.forest = TRUE)
               Type of random forest: regression
                     Number of trees: 458
No. of variables tried at each split: 4
          Mean of squared residuals: 0.1859937
                    % Var explained: 25.42
```

We reduced the number of trees by 42, yet the results have not significantly changed. However, the mean of squared residuals has decreased and the %Var explained has mechanically increased. Despite the %Var explained remaining quite low, we move on to our interpretation which is based on the article "*Understanding random forests with randomForestExplainer*" written by Aleksandra Paluszyńska**.**

Figure 50
Distribution of minimal depth and its mean



The plot above shows the distribution of minimal depth among the trees of our forest. Note that:

- the mean of the distribution is marked by a vertical bar with a value label on it (the scale for it is different than for the rest of the plot),
- the scale of the X axis goes from zero to the maximum number of trees in which any variable was used for splitting.
- we used the default parameter of the function which consider only the top ten variables according to their mean minimal depth.

By definition, "minimal depth for a variable in a tree equals to the depth of the node which splits on that variable and is the closest to the root of the tree. If it is low, then a lot of observations are divided into groups on the basis of this variable". In simpler words, the distribution of the mean minimal depth allows us to appreciate the variable's role in the random forest's structure and prediction.

Interpreting results from a Random Forest model can be quite tedious. However, if we focus on "variable importance", i.e., whether that variable was selected to split on during the tree

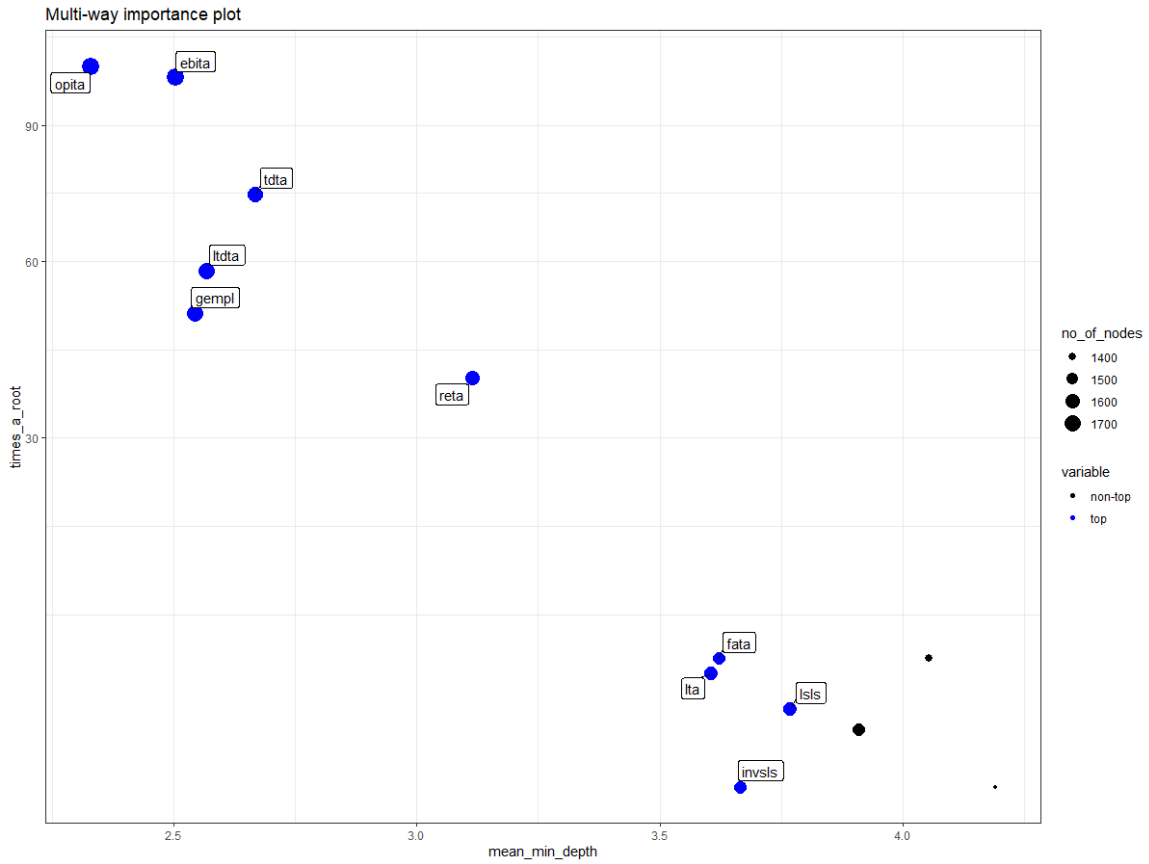
building process, and how much the squared error (over all trees) improved (decreased) as a result, we should get out of the woods safely.

As one may notice, when it comes to implementing our Random Forest, the algorithm selects the five variables that appear to be the most important: *opita*, *ebita*, *gempl*, *ltdta* and *tdta*. To further explore variables importance measures, we produced two more graphs which you will find below.

Figure 51

Multi-way importance plot

No_of_nodes versus Times_a_root versus Mean_min_depth



First, we must define the measurement tools used to produce this graph. Let X_j, an explanatory variable of our model:

- *no_of_nodes*: total number of nodes that use X_j for splitting (it is usually equal to *no_of_trees* if trees are shallow).
- *times_a_root*: total number of trees in which X_j is used for splitting the root node (i.e., the whole sample is divided into two based on the value of X_j).
- *mean_min_depth*: mean minimal depth.

Recall that at each node, a subset of the full set of explanatory variables is evaluated for their strength of association with the dependent variable. The most strongly associated explanatory variable is then used to split the data. Both *times_a_root* and *mean_min_depth* are straightforward ways of measuring variable importance: a predictor that is closer to the root, or one that on average occurs closer to the root, is one that is strongly associated with the

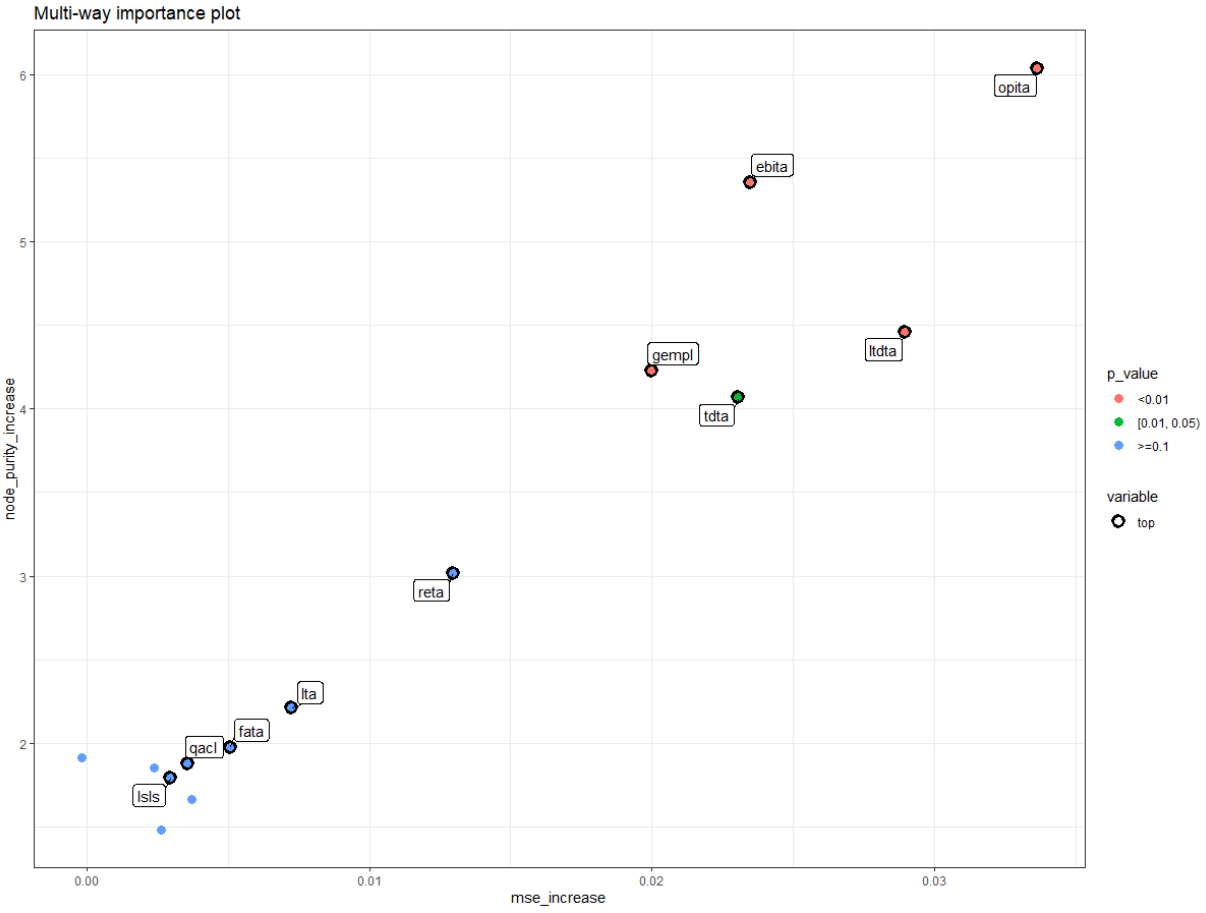
dependent variable. On our graph, we can observe a strong negative relationship between both those metrics, meaning that if a variable is often used for splitting the root node, the average depth of the node which splits on that variable should be smaller.

When considering our three metrics, we can conclude that our five important variables cited earlier clearly dominate the other variables.

Figure 52

Multi-way importance plot

Node purity increase versus MSE increase versus P-value



Let us define our three metrics used to produce this graph:

- *node_purity_increase*: measure of importance based on loss function
- *mse_increase*: measure of importance based on the decrease in predictive accuracy post variable perturbation.
- *p-value*: number of nodes in which predictor X_j was used for splitting exceeds the theoretical number of successes if they were random, following the binomial distribution given (i.e., if a variable is significant, it means that the variable is used for splitting more often than would be the case if the selection was random).

As in the previous plot, the two measures used as coordinates seem correlated. However, in this case instead of having two metrics related to the structure of the Random Forest, we have one related to its structure and the other one to its prediction.

In the previous plot, it seemed that *opita* and *ebita* were rather similar regarding their mean minimal depth and number of root splits, but Figure 52 suggest that *opita* dominates the other variables. The variable *ltdta* could be a candidate for the second place. However, its p-value is higher, and its node purity increase is smaller compared to *ebita*'s.

Before going to further on variable importance and interactions, lets plot the different relationships between our measures of importance to ensure ourselves we correctly selected our top variables.

Figure 53 suggests that all our importance measures are highly correlated with each on another. Note that the "smallest" correlation being *corr(mse_increase,no_of_nodes)*= 0.851 still remains quite high. Thus, we are quite limited when it comes to building a coherent "importance graph" since that, according to A. Paluszyńska, we should "select three [importance measures] that least agree with each other and use them in the multi-way importance plot to select top variables". When it comes to defining the rank of our variables, Figure 54 indicates that the correlation between importance measures is even stronger.

Having highly correlated importance measures will not prevent us to move on our development. The two previous plots are useful when selecting a set of important variables reveals itself tedious. Indeed, by studying interactions between importance measures we can build more efficient multi-way plots to identify our most important variables.

Figure 53

Correlation matrix: Measures of importance

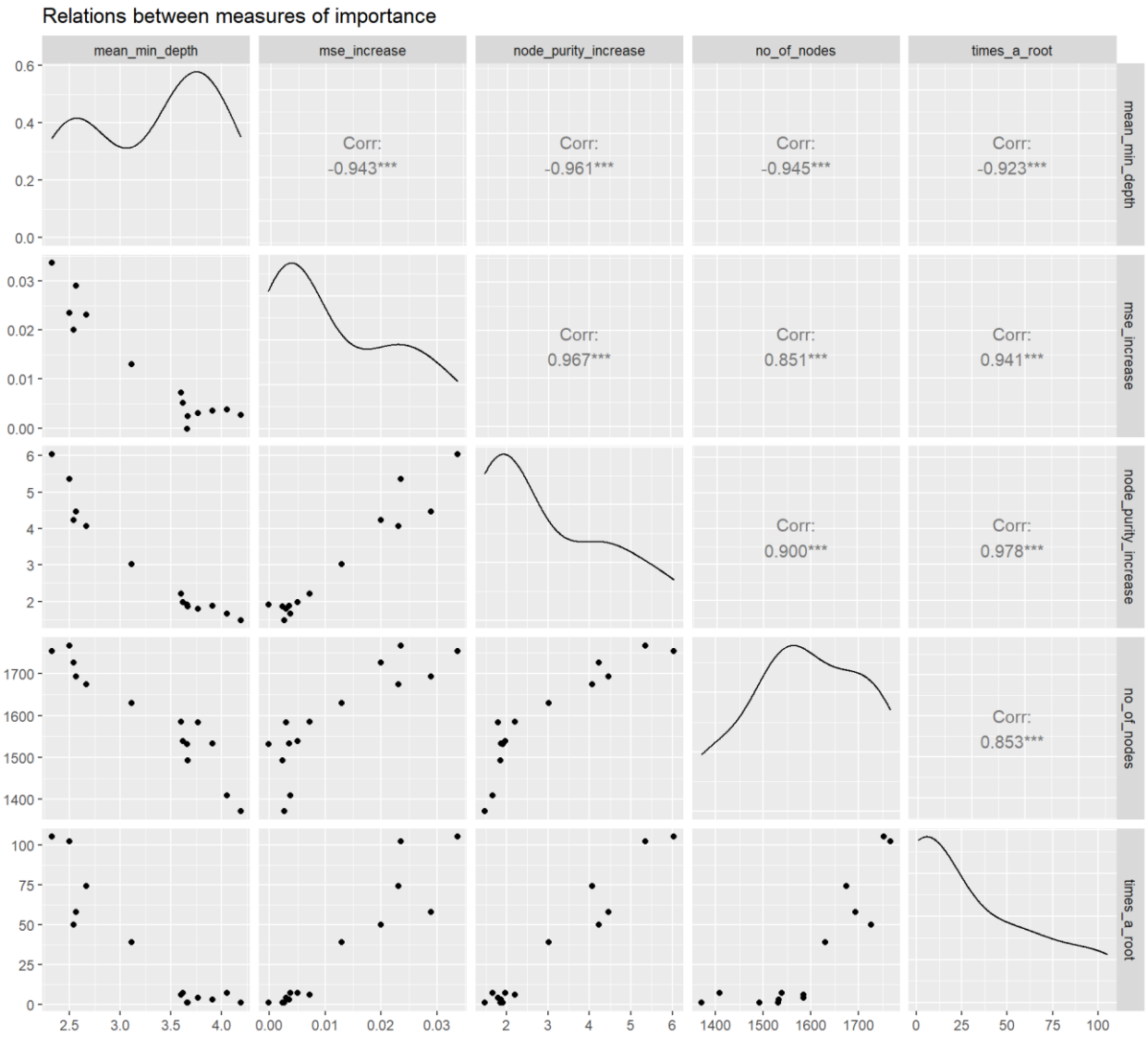
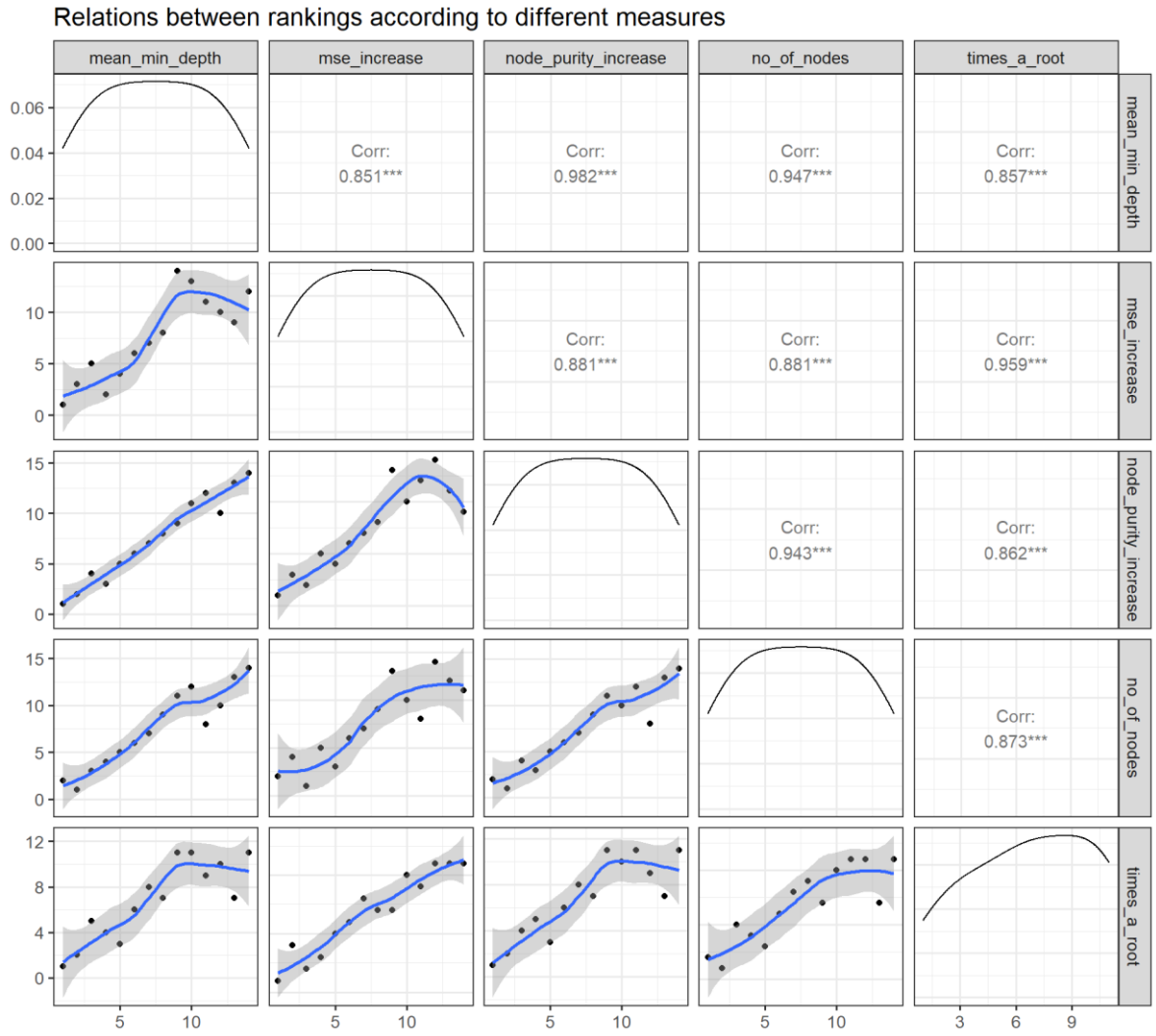


Figure 54

Correlation matrix and fitted LOESS curve: Ranking according to different measures



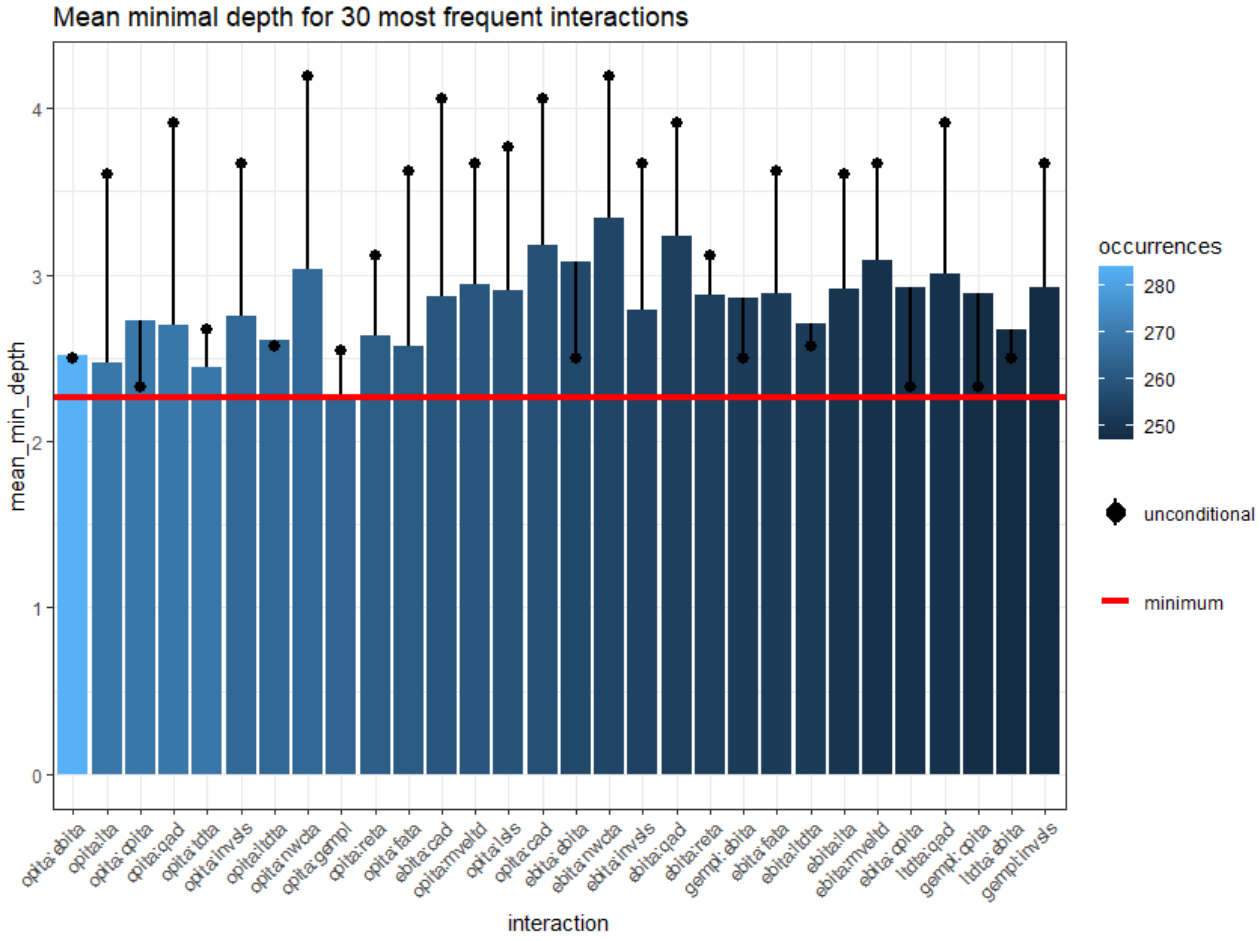
Since indentifying our set of most important variables was rather simple as their repeatedly dinstinguish themselves against the other variables, we can move on to the next step: variables interactions.

The plot below reports 30 top interactions according to mean of conditional minimal depth – a generalization of minimal depth that measures the depth of the second variable in a tree of which the first variable is a root (a subtree of a tree from the forest). In order to be comparable to normal minimal depth 1 is subtracted so that 0 is the minimum.

For example value of 0 for interaction x: y in a tree means that if we take the highest subtree with the root splitting on x then y is used for splitting immediately after x (minimal depth of x in this subtree is 1). The values presented are means over all trees in the forest. Note that:

- the plot shows only 30 interactions that appeared most frequently,
- the horizontal line shows the minimal value of the depicted statistic among interactions for which it was calculated,
- the interactions considered are ones with the following variables as first (root variables): *opita*, *ebita*, *ltdta*, *gempl*, *tdta*, *reta*, *lta*, *fata*, *lsls*, *qacl*, *invsls*, *mveltd*, *cacl*, *nwcta* and all possible values of the second variable.

Figure 55
Variables interactions



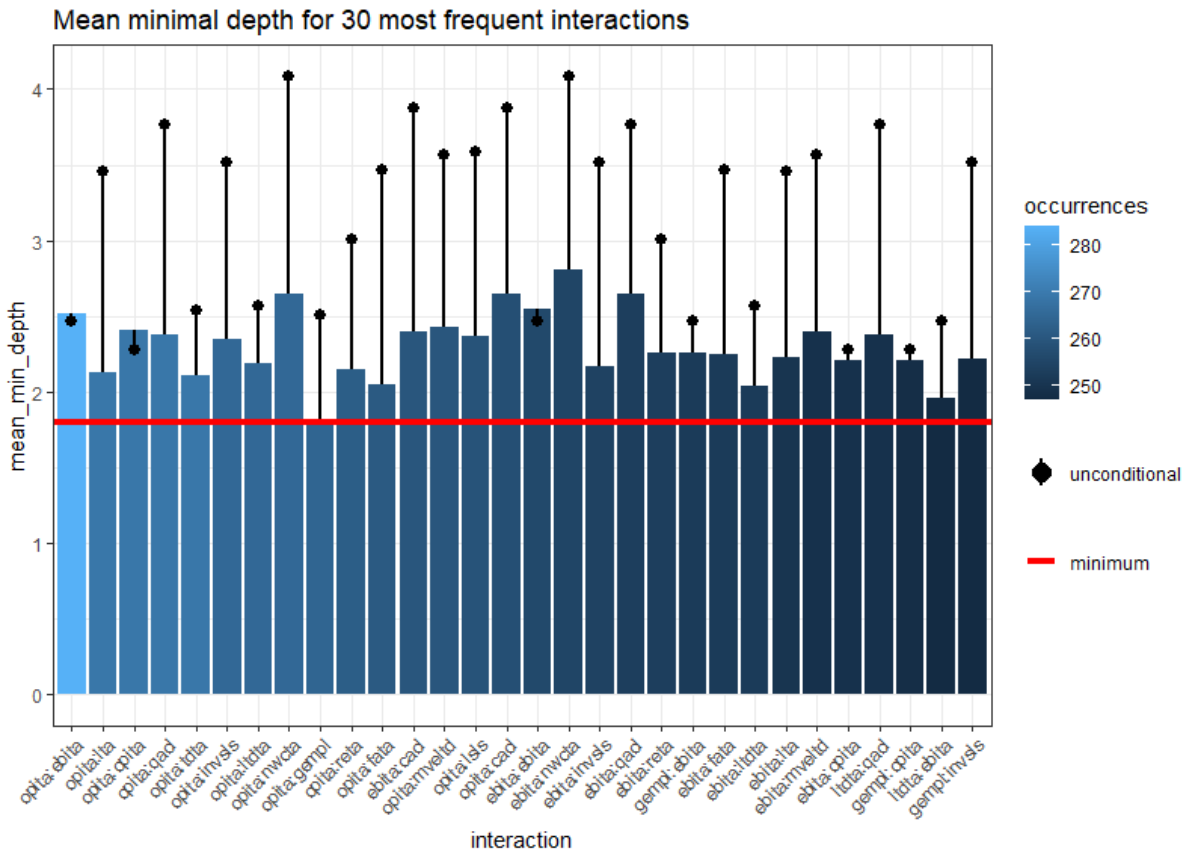
Note that the interactions are ordered by decreasing number of occurences and that regarding mean minimal depth we have a rather homogenous distribution across variables. The most frequent interactions depend on four of the five variables we retained earlier : *opita, gempl, ebita, ltdta*.

To complete our study of variables interactions, we repeat the computation of means using only "*relevant_trees*", a parameter that allow the algorithm to ignore missing values.

Figure 56

Variables intereactions

Without NAs



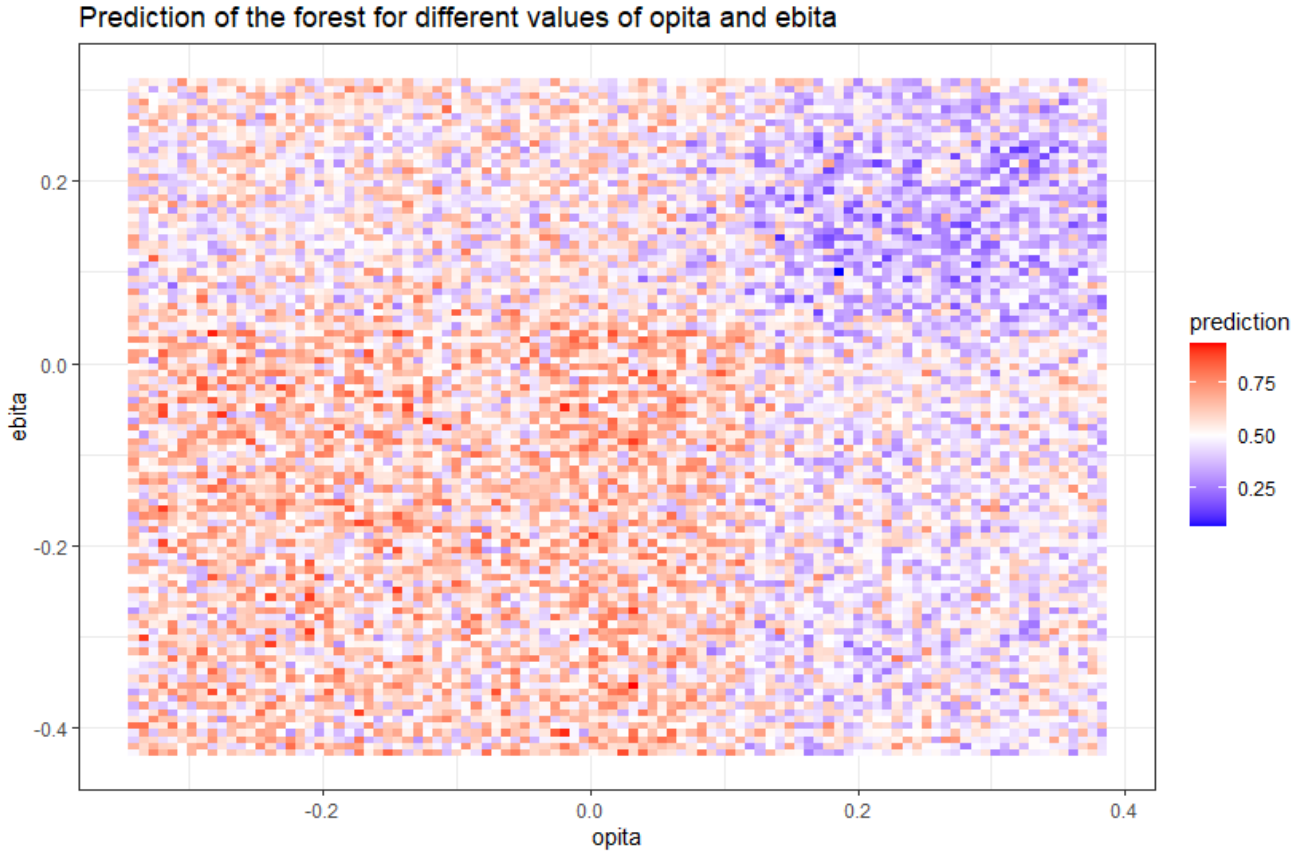
Comparing this plot with the previous one, we see that removing penalization of missing values lowers the mean conditional minimal depth of all interactions except the most frequent one. Now regarding the classification of interactions, it seemed that it is not affected by this change.

To further investigate the most frequent interaction *opita:ebita*, we plot the prediction of our forest on a grid of values for the components of each interaction. The function requires the forest, training data, variable to use on x and y-axis, respectively.

Remember that *opita* is the ratio of income over debt. Therefore, having most of variable's interactions depending on *opita* makes also sense from a financial point of view.

Figure 57

Prediction of the forest for different values of opita and ebita



In the above plot we can clearly see the effect of interaction: the predicted financial distress is highest when both *ebita* and *opita* are negative or low.

Thanks to the Random Forest and the now clearly identified variables interactions, we can improve our previous logit model by taking into account those interactions.

Our model is statistically significant ($(X^2(17) = 54.25,\ p = 0.000)$ with an associated Pseudo R^2 (Cragg-Uhler) of 0.6 (which is > 0.524, the previous R^2 we obtained from our simple logit model). Here again, most of the individual β estimates are not significant. However, compared to the simple logit model, we obtain four new significant estimates : *ebita*, *qacl*, *cacl*, *opita:qacl*. Moreover, the AUROC curve (Figure 58) is 0.895, which is again superior to the value we obtained from our first logit model. Thus, we can reasonably affirm that using results obtained from a Random Forest model can improve the fitness and predictability of a logit model.

Figure 35

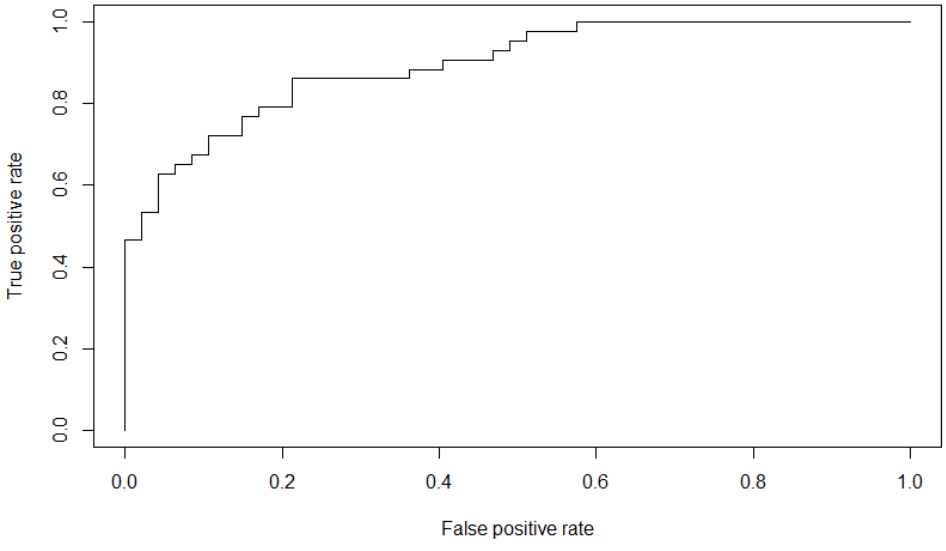
Improved logit model with interactions

	Model 2		
(Intercept)	-5.32963	*nwcta*	-6.80908
	(5.88147)		(6.26797)
tdta	-0.87959	*cacl*	-2.63844 *
	(6.54580)		(1.34542)
reta	-3.01224	*qacl*	7.79014 *
	(2.86051)		(3.05765)
opita	-24.81128	*fata*	4.25518
	(14.90326)		(5.18680)
ebita	50.69643 **	*ltdta*	-0.45141
	(19.22081)		(0.37509)
lsls	2.29400	*mveltd*	3.64494
	(1.61782)		(3.84305)
lta	-1.82191	*opita:ebita*	44.39090
	(1.63877)		(29.66772)
gempl	-7.47507 *	*opita:lta*	-3.03092
	(3.55199)		(2.19898)
invsls	20.39010	*opita:qacl*	-21.47882 *
	(10.94496)		(10.63895)
N	91	BIC	152.82199
AIC	107.62652	Pseudo R2	0.59938

**** p < 0.001; ** p < 0.01; * p < 0.05.*

Figure 58

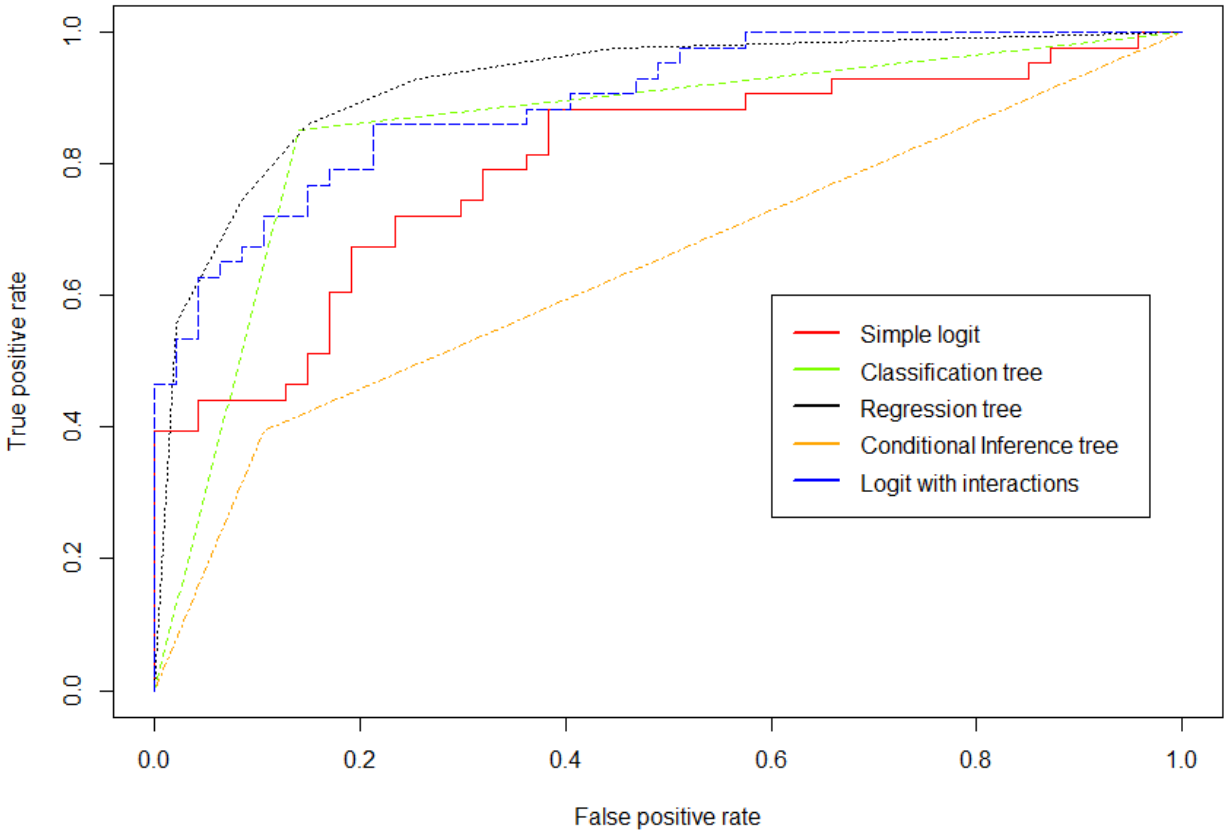
ROC curve : Improved logit model with interactions



As our journey through the woods ends here, we wanted to sum up our results one last time. The following figure plots all the tested models so far for a better visual comparison.

Figure 59

ROC Curves: Simple logit, Classification tree, Regression tree, Conditional inference tree, Logit with interactions



In conclusion to this second sub-section on decision trees and Random Forests, it appears that the regression tree model seems to better suit our data. The second best model is the augmented logit with interactions and the least performing model is the conditional inference tree. All models have, by construction, their strengths and weaknesses and having the opportunity to work with so many models gave us some interesting insights on their underlying methodology as well as how to implement them in an empirical framework using R.

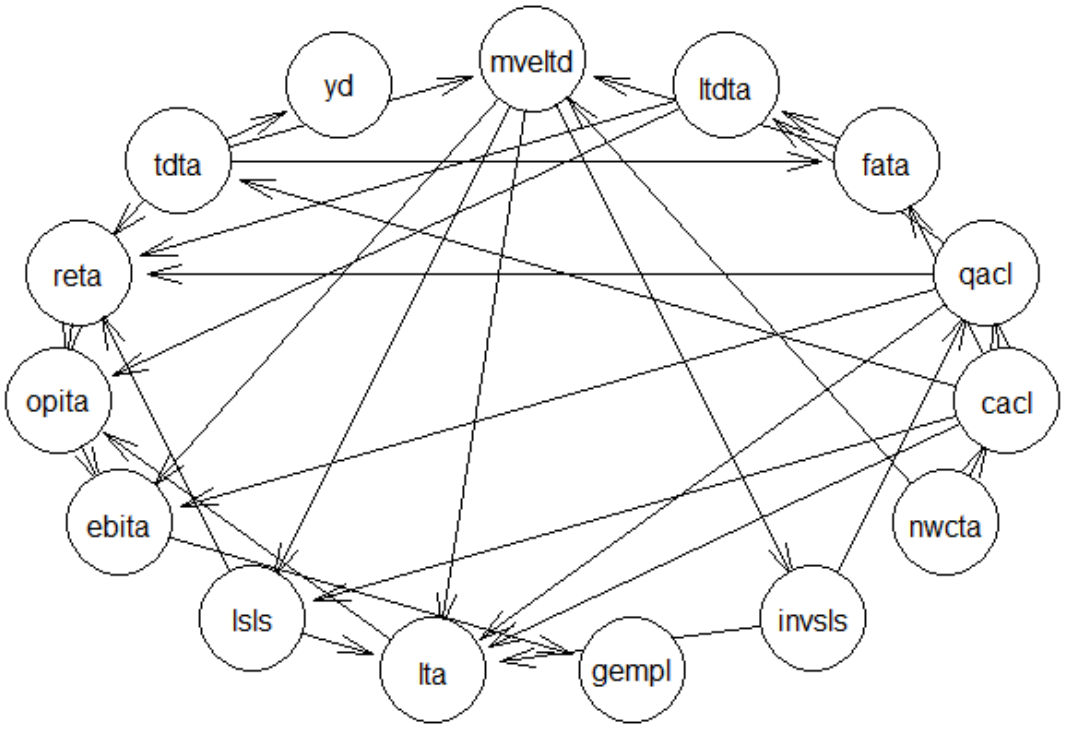
iii. Bayesian Network

The Bayesian Network gave us more of a challenge than the two previous sub-sections. Indeed, many questions popped in our heads due to the complicated interactions between the variables in the sample, especially regarding the understanding of the statistical output and its economic interpretation. Yet, we still wanted to present our results.

First, a short definition: a Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies. In our case, we could use this model to represent the probabilistic relationships between financial distress and financial indicators. Given a certain level of an indicator, we could potentially compute the occurrence's probability of financial distress.

Figure 60

Bayesian Network



The primary aspect that stands out of our Bayesian network is that our variable of interest *yd* depends only on the variable *tdta* meaning that the algorithm considers that *yd* is conditionally independent from all the other variables included in our set. The maximum likelihood estimates of the regression coefficients in the local distribution of an outcome variables against all the explanatory variable is identical to those produced by the command *lm()* when the option *na.action = na.omit* is selected. Therefore, having *yd* depending on *tdta* is quite an expected result since we never obtained a significant estimate for *tdta* in our previous model. However, we failed to reproduce this result when including all explanatory variables into a linear probability model. Yet, we succeeded when only including the *tdta* variable (R-Code 19) and again we obtained a surprising result as both models share a similar adjusted R^2 of around 0.26. This finding gives us an insight on our previous models: we might have considered too many variables to construct them and therefore decreased their predictability power.

R-Code 17

Conditional dependencies

```
       Bayesian network parameters
  Parameters of node yd (Gaussian distribution)

Conditional density: yd | tdta
Coefficients:
(Intercept)         tdta
  -0.187342     1.204772
Standard deviation of the residuals: 0.4315745

  Parameters of node tdta (Gaussian distribution)

Conditional density: tdta | cacl
Coefficients:
(Intercept)         cacl
  0.8078751   -0.1097202
Standard deviation of the residuals: 0.1734254

  Parameters of node reta (Gaussian distribution)

Conditional density: reta | tdta + lsls + qacl + ltdta
Coefficients:
(Intercept)         tdta         lsls         qacl        ltdta
 0.86522169  -1.26612232   0.03093307  -0.05940230  -0.01185555
Standard deviation of the residuals: 0.1403845

 Parameters of node opita (Gaussian distribution)

Conditional density: opita | reta + lta + ltdta
Coefficients:
 (Intercept)          reta           lta         ltdta
-0.025613832   0.214868333   0.013581418   0.006991201
Standard deviation of the residuals: 0.09087961
```

```
  Parameters of node ebita (Gaussian distribution)

Conditional density: ebita | opita + qacl + mveltd
Coefficients:
 (Intercept)         opita          qacl        mveltd
-0.012068767   1.086626027  -0.008652885  -0.092954681
Standard deviation of the residuals: 0.03142905

  Parameters of node lsls (Gaussian distribution)

Conditional density: lsls | cacl + mveltd
Coefficients:
(Intercept)         cacl       mveltd
  4.7545171   -0.3102058    4.8714353
Standard deviation of the residuals: 1.582516

  Parameters of node lta (Gaussian distribution)

Conditional density: lta | lsls + invsls + cacl + qacl + mveltd
Coefficients:
(Intercept)         lsls       invsls         cacl         qacl       mvelt
d
 -1.6684705    1.0040414    4.1293111   -0.3760829    0.7144308    1.334351
3
Standard deviation of the residuals: 0.2574704

  Parameters of node gempl (Gaussian distribution)

Conditional density: gempl | ebita
Coefficients:
 (Intercept)         ebita
-0.008655849   0.287962942
Standard deviation of the residuals: 0.1107535
```

```
  Parameters of node invsls (Gaussian distribution)

Conditional density: invsls | mveltd
Coefficients:
(Intercept)       mveltd
  0.2502352   -0.2144210
Standard deviation of the residuals: 0.08445869

  Parameters of node nwcta (Gaussian distribution)

Conditional density: nwcta
Coefficients:
(Intercept)
  0.2863189
Standard deviation of the residuals: 0.1906568

  Parameters of node cacl (Gaussian distribution)

Conditional density: cacl | nwcta
Coefficients:
(Intercept)        nwcta
  0.9842563    4.8438346
Standard deviation of the residuals: 0.7513547

  Parameters of node qacl (Gaussian distribution)

Conditional density: qacl | invsls + cacl
Coefficients:
(Intercept)       invsls         cacl
 0.05506245  -2.66578185   0.74020417
Standard deviation of the residuals: 0.3905957
```

```
  Parameters of node fata (Gaussian distribution)

Conditional density: fata | tdta + cacl
Coefficients:
(Intercept)         tdta         cacl
-0.21723667   0.60054716   0.03424007
Standard deviation of the residuals: 0.1175938

  Parameters of node ltdta (Gaussian distribution)

Conditional density: ltdta | qacl + fata
Coefficients:
(Intercept)         qacl         fata
 -0.6641083    2.8138954   -4.7047801
Standard deviation of the residuals: 3.105172

  Parameters of node mveltd (Gaussian distribution)

Conditional density: mveltd | tdta + nwcta + fata
Coefficients:
(Intercept)         tdta        nwcta         fata
  0.7551563   -0.5470945   -0.6806156    0.4059630
Standard deviation of the residuals: 0.111639
```

R-Code 18

Linear probability model: All explanatory variables

```
Call:
lm(formula = yd ~ ., data = Estim, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75897 -0.32733 -0.06365  0.33048  0.87872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09450    0.60498   0.156   0.8763
tdta         0.33118    0.68706   0.482   0.6312
reta        -0.36668    0.35714  -1.027   0.3078
opita       -3.25134    1.81055  -1.796   0.0765 .
ebita        2.81479    1.62608   1.731   0.0875 .
lsls         0.10422    0.20120   0.518   0.6060
lta         -0.09382    0.19617  -0.478   0.6338
gempl       -1.10096    0.47070  -2.339   0.0220 *
invsls       1.74857    1.18929   1.470   0.1456
nwcta       -0.15898    0.56759  -0.280   0.7802
cacl        -0.22059    0.14015  -1.574   0.1197
qacl         0.39068    0.22011   1.775   0.0799 .
fata         0.15045    0.48925   0.308   0.7593
ltdta       -0.02456    0.01714  -1.433   0.1560
mveltd       0.40937    0.50095   0.817   0.4164
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4287 on 76 degrees of freedom
Multiple R-squared:  0.3841,   Adjusted R-squared:  0.2706
F-statistic: 3.385 on 14 and 76 DF,  p-value: 0.0002974
```

R-Code 19

Linear probability model: yd on tdta

```
Call:
lm(formula = yd ~ tdta, data = Estim, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7761 -0.3926 -0.0786  0.4293  0.9304

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1873     0.1238  -1.513    0.134
tdta          1.2048     0.2104   5.725 1.38e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4316 on 89 degrees of freedom
Multiple R-squared:  0.2691,   Adjusted R-squared:  0.2609
F-statistic: 32.77 on 1 and 89 DF,  p-value: 1.379e-07
```

Annexes – Question 21

R-Code 21.1

Classification tree: Rstudio output

```
Call:
rpart(formula = yd ~ ., data = Estim, method = "class")
  n= 91


          CP nsplit rel error    xerror      xstd
1 0.51162791      0 1.0000000 1.1162791 0.1107556
2 0.06976744      1 0.4883721 0.7441860 0.1059283
3 0.02325581      3 0.3488372 0.6511628 0.1023906
4 0.01000000      4 0.3255814 0.6279070 0.1013404

Variable importance
 ltdta   tdta   reta  ebita   fata  gempl  opita   qacl   cacl invsls  nwct
a   lsls    lta
    20     19     14     14     11      8      8      2      1      1
1      1      1

Node number 1: 91 observations,    complexity param=0.5116279
  predicted class=No Default  expected loss=0.4725275  P(node) =1
    class counts:    43    48
   probabilities: 0.473 0.527
  left son=2 (56 obs) right son=3 (35 obs)
  Primary splits:
      ltdta < 1.3214     to the left,  improve=14.59835, (0 missing)
      opita < 0.1232996  to the left,  improve=12.23594, (0 missing)
      tdta  < 0.6894356  to the right, improve=10.82901, (0 missing)
      ebita < 0.04608445 to the left,  improve=10.48522, (0 missing)
      reta  < 0.09310425 to the left,  improve= 7.39294, (0 missing)
  Surrogate splits:
      tdta  < 0.4618699  to the right, agree=0.857, adj=0.629, (0 split)
      reta  < 0.4291527  to the left,  agree=0.747, adj=0.343, (0 split)
      fata  < 0.1298109  to the right, agree=0.747, adj=0.343, (0 split)
      ebita < 0.0878356  to the left,  agree=0.714, adj=0.257, (0 split)
      gempl < 0.04146995 to the left,  agree=0.714, adj=0.257, (0 split)
```

```
Node number 2: 56 observations,    complexity param=0.06976744
  predicted class=Default     expected loss=0.3035714  P(node) =0.6153846
    class counts:    39    17
   probabilities: 0.696 0.304
  left son=4 (25 obs) right son=5 (31 obs)
  Primary splits:
      ebita < 0.04718255 to the left,  improve=6.274700, (0 missing)
      opita < 0.1406256  to the left,  improve=5.760019, (0 missing)
      reta  < 0.0879118  to the left,  improve=4.000794, (0 missing)
      tdta  < 0.6894356  to the right, improve=3.621672, (0 missing)
      lta   < 7.8047     to the left,  improve=2.698980, (0 missing)
  Surrogate splits:
      opita < 0.0870639  to the left,  agree=0.929, adj=0.84, (0 split)
      reta  < 0.0879118  to the left,  agree=0.875, adj=0.72, (0 split)
      tdta  < 0.7009355  to the right, agree=0.804, adj=0.56, (0 split)
      gempl < -0.0755221 to the left,  agree=0.732, adj=0.40, (0 split)
      qacl  < 0.9219013  to the left,  agree=0.679, adj=0.28, (0 split)

Node number 3: 35 observations
  predicted class=No Default  expected loss=0.1142857  P(node) =0.3846154
    class counts:     4    31
   probabilities: 0.114 0.886

Node number 4: 25 observations
  predicted class=Default     expected loss=0.04  P(node) =0.2747253
    class counts:    24     1
   probabilities: 0.960 0.040

Node number 5: 31 observations,    complexity param=0.06976744
  predicted class=No Default  expected loss=0.483871  P(node) =0.3406593
    class counts:    15    16
   probabilities: 0.484 0.516
  left son=10 (11 obs) right son=11 (20 obs)
```

```
  Primary splits:
      fata < 0.2332357  to the right, improve=2.020235, (0 missing)
      reta < 0.3774628  to the left,  improve=1.736396, (0 missing)
      lsls < 4.856544   to the left,  improve=1.527349, (0 missing)
      lta  < 4.318782   to the left,  improve=1.527349, (0 missing)
      qacl < 1.084636   to the right, improve=1.308432, (0 missing)
  Surrogate splits:
      reta   < 0.2156669  to the left,  agree=0.839, adj=0.545, (0 split)
      tdta   < 0.6114259  to the right, agree=0.806, adj=0.455, (0 split)
      invsls < 0.2188934  to the right, agree=0.742, adj=0.273, (0 split)
      nwcta  < 0.5000626  to the right, agree=0.742, adj=0.273, (0 split)
      cacl   < 2.437241   to the right, agree=0.742, adj=0.273, (0 split)

Node number 10: 11 observations
  predicted class=Default     expected loss=0.2727273  P(node) =0.1208791
    class counts:     8     3
   probabilities: 0.727 0.273

Node number 11: 20 observations,    complexity param=0.02325581
  predicted class=No Default  expected loss=0.35  P(node) =0.2197802
    class counts:     7    13
   probabilities: 0.350 0.650
  left son=22 (9 obs) right son=23 (11 obs)
  Primary splits:
      fata   < 0.1509046  to the left,  improve=1.382828, (0 missing)
      qacl   < 1.036181   to the right, improve=1.350000, (0 missing)
      tdta   < 0.5148624  to the left,  improve=1.056044, (0 missing)
      invsls < 0.1136279  to the left,  improve=1.056044, (0 missing)
      nwcta  < 0.338116   to the right, improve=1.056044, (0 missing)
  Surrogate splits:
      opita < 0.2175087  to the right, agree=0.75, adj=0.444, (0 split)
      tdta  < 0.473561   to the left,  agree=0.70, adj=0.333, (0 split)
      ebita < 0.1346948  to the right, agree=0.70, adj=0.333, (0 split)
      lsls  < 5.148528   to the left,  agree=0.70, adj=0.333, (0 split)
      lta   < 4.521986   to the left,  agree=0.70, adj=0.333, (0 split)
```

```
Node number 22: 9 observations
  predicted class=Default     expected loss=0.4444444  P(node) =0.0989011
    class counts:     5     4
   probabilities: 0.556 0.444

Node number 23: 11 observations
  predicted class=No Default  expected loss=0.1818182  P(node) =0.1208791
    class counts:     2     9
   probabilities: 0.182 0.818
```

R-Code 21.2

Regression tree: Rstudio output

```
Call:
rpart(formula = yd ~ ., data = Estim, method = "anova")
  n= 91

          CP nsplit rel error    xerror       xstd
1 0.32181444      0 1.0000000 1.0290859 0.01417371
2 0.13832310      1 0.6781856 0.8005321 0.10313226
3 0.04453521      2 0.5398625 0.6774385 0.10864360
4 0.03107818      3 0.4953273 0.8627342 0.13219303
5 0.03048386      4 0.4642491 0.8834112 0.13247855
6 0.01000000      5 0.4337652 0.8730456 0.13196567

Variable importance
 ltdta   tdta  ebita   reta   fata  gempl  opita   qacl   cacl invsls  nwct
a   lsls    lta
    19     19     14     14     11     10      8      2      1      1
1      1      1

Node number 1: 91 observations,    complexity param=0.3218144
  mean=0.4725275, MSE=0.2492453
  left son=2 (35 obs) right son=3 (56 obs)
```

```
  Primary splits:
      ltdta < 1.3214     to the right, improve=0.3218144, (0 missing)
      opita < 0.1232996  to the right, improve=0.2697362, (0 missing)
      tdta  < 0.6894356  to the left,  improve=0.2387209, (0 missing)
      ebita < 0.04608445 to the right, improve=0.2311422, (0 missing)
      reta  < 0.09310425 to the right, improve=0.1629742, (0 missing)
  Surrogate splits:
      tdta  < 0.4618699  to the left,  agree=0.857, adj=0.629, (0 split)
      reta  < 0.4291527  to the right, agree=0.747, adj=0.343, (0 split)
      fata  < 0.1298109  to the left,  agree=0.747, adj=0.343, (0 split)
      ebita < 0.0878356  to the right, agree=0.714, adj=0.257, (0 split)
      gempl < 0.04146995 to the right, agree=0.714, adj=0.257, (0 split)

Node number 2: 35 observations,    complexity param=0.03107818
  mean=0.1142857, MSE=0.1012245
  left son=4 (27 obs) right son=5 (8 obs)
  Primary splits:
      gempl  < -0.0010579 to the right, improve=0.19896210, (0 missing)
      cacl   < 2.025956   to the right, improve=0.19896210, (0 missing)
      mveltd < 0.2226897  to the right, improve=0.19896210, (0 missing)
      fata   < 0.04843845 to the right, improve=0.19354840, (0 missing)
      opita  < 0.1433486  to the right, improve=0.07920708, (0 missing)
  Surrogate splits:
      tdta  < 0.2170261  to the right, agree=0.8, adj=0.125, (0 split)
      opita < 0.00529425 to the right, agree=0.8, adj=0.125, (0 split)
      ebita < -0.0388773 to the right, agree=0.8, adj=0.125, (0 split)
      ltdta < 1.37937    to the right, agree=0.8, adj=0.125, (0 split)

Node number 3: 56 observations,    complexity param=0.1383231
  mean=0.6964286, MSE=0.2114158
  left son=6 (31 obs) right son=7 (25 obs)
  Primary splits:
      ebita < 0.04718255 to the right, improve=0.2649949, (0 missing)
      opita < 0.1406256  to the right, improve=0.2432587, (0 missing)
      reta  < 0.0879118  to the right, improve=0.1689626, (0 missing)
```

```
      tdta  < 0.6894356  to the left,  improve=0.1529515, (0 missing)
      lta   < 7.8047     to the right, improve=0.1139841, (0 missing)
  Surrogate splits:
      opita < 0.0870639  to the right, agree=0.929, adj=0.84, (0 split)
      reta  < 0.0879118  to the right, agree=0.875, adj=0.72, (0 split)
      tdta  < 0.7009355  to the left,  agree=0.804, adj=0.56, (0 split)
      gempl < -0.0755221 to the right, agree=0.732, adj=0.40, (0 split)
      qacl  < 0.9219013  to the right, agree=0.679, adj=0.28, (0 split)

Node number 4: 27 observations
  mean=0.03703704, MSE=0.03566529

Node number 5: 8 observations
  mean=0.375, MSE=0.234375

Node number 6: 31 observations,    complexity param=0.04453521
  mean=0.483871, MSE=0.2497399
  left son=12 (20 obs) right son=13 (11 obs)
  Primary splits:
      fata < 0.2332357  to the left,  improve=0.13047350, (0 missing)
      reta < 0.3774628  to the right, improve=0.11214230, (0 missing)
      lsls < 4.856544   to the right, improve=0.09864130, (0 missing)
      lta  < 4.318782   to the right, improve=0.09864130, (0 missing)
      qacl < 1.084636   to the left,  improve=0.08450292, (0 missing)
  Surrogate splits:
      reta   < 0.2156669  to the right, agree=0.839, adj=0.545, (0 split)
      tdta   < 0.6114259  to the left,  agree=0.806, adj=0.455, (0 split)
      invsls < 0.2188934  to the left,  agree=0.742, adj=0.273, (0 split)
      nwcta  < 0.5000626  to the left,  agree=0.742, adj=0.273, (0 split)
      cacl   < 2.437241   to the left,  agree=0.742, adj=0.273, (0 split)
```

```
Node number 7: 25 observations
  mean=0.96, MSE=0.0384

Node number 12: 20 observations,    complexity param=0.03048386
  mean=0.35, MSE=0.2275
  left son=24 (11 obs) right son=25 (9 obs)
  Primary splits:
      fata   < 0.1509046  to the right, improve=0.1519592, (0 missing)
      qacl   < 1.036181   to the left,  improve=0.1483516, (0 missing)
      tdta   < 0.5148624  to the right, improve=0.1160488, (0 missing)
      invsls < 0.1136279  to the right, improve=0.1160488, (0 missing)
      nwcta  < 0.338116   to the left,  improve=0.1160488, (0 missing)
  Surrogate splits:
      opita < 0.2175087  to the left,  agree=0.75, adj=0.444, (0 split)
      tdta  < 0.473561   to the right, agree=0.70, adj=0.333, (0 split)
      ebita < 0.1346948  to the left,  agree=0.70, adj=0.333, (0 split)
      lsls  < 5.148528   to the right, agree=0.70, adj=0.333, (0 split)
      lta   < 4.521986   to the right, agree=0.70, adj=0.333, (0 split)

Node number 13: 11 observations
  mean=0.7272727, MSE=0.1983471

Node number 24: 11 observations
  mean=0.1818182, MSE=0.1487603

Node number 25: 9 observations
  mean=0.5555556, MSE=0.2469136
```

Question 22
Practical changes to the STATA code

Working with the default dataset and the provided do-file gave us many ideas on how to improve the STATA code. Our rework on the code includes adding, modifying, deleting, rewriting, simplifying, installing new packages and replacing some parts of the code. As listing those modifications would be too long (and certainly boring to read), we attached to this application the reworked STATA code as well as the R code used in Question 21. Authors are available at any time to comment and share their experience on how the chunks of code work.

Thank you.

Roland BOUILLOT
Khalil JANBEK
Mehdi LOUAFI