VRIJE UNIVERSITEIT AMSTERDAM

EMPIRICAL FINANCE

Case 1

Authors: **Group 56**

Soheel Gholamy: 2668182

Mark de Kwaasteniet: 2649271

S.gholamy@student.vu.nl

M.G.de.Kwaasteniet@student.vu.nl

Supervisor:

Anne Opschoor

Amsterdam, September 20, 2021

1 Part 1

1.1 Question 1a

The data set contains 406 companies, without any missing values for the default probability variable (DEF). DEF is a variable that can only take on the values 0 (not defaulting) and 1 (defaulting). The maximum and minimum of the DEF variable is 0 and 1, which is conform with the description of the variable. The Skewness and Kurtosis test indicates that the probability that DEF is normally distributed is lower than 1%. Furthermore, the average value of DEF is 0.108, which indicates that approximately 10% (44 companies) of the data set defaulted.

Table 1

Summary Statistics

Variable Names	Mean	Std. Deviation	Min	Max	P-Value
DEF	.108	.311	0.000	1	0
sic2	42.911	18.676	10.000	99	0
mv	.878	1.36	-3.780	4.16	.13
ebitta	.077	.086	-0.330	.4	0
wkta	.18	.205	-0.510	.73	0
reta	.134	.332	-1.043	1.56	0

This table contains the summary statistics of the variables in the data set of case 1 - group 56. The mean, standard deviation, minimum value and maximum values are represented in column 2,3,4 and 5 respectively (N = 406). P-value denotes the probability that the Chi-Squared distributed variables is greater than the test statistic. The p-value is calculated with the Jarque Bera test, which jointly tests the skewness and kurtosis. It represents the probability of accepting the null-hypothesis of the test, which is a joint hypothesis of the skewness being equal to zero and kurtosis being equal to three.

1.2 Question 1b

Model 2:

$$DEF_i = 0.208 - 0.432EBITTA_i - 0.059MV_i + 0.014WKTA_i - 0.131RETA_i \quad (1)$$

Table 2

Linear Probability Model: Ratios affecting the default probability

	Model 1	Model 2
ebitta	-0.615***	-0.432**
	(0.174)	(0.187)
mv	-0.0664***	-0.0587***
	(0.0110)	(0.0116)
wkta		0.0143
		(0.0714)
reta		-0.131***
		(0.0497)
Constant	0.214***	0.208***
	(0.0201)	(0.0234)
R^2	0.142	0.156
$R^2 - Adjusted$	0.137	0.148
F-statistic	33.22	18.56

This table presents the regression results of model 1 and 2. Model 1 is the regression of the default probability on the EBITTA (Earnings Before Interest and Taxes / Total Assets) and the MV (Market to Book Value). Model 2 is an extension of model 1, including the factors WKTA (Working Capital / Total Assets) and RETA (Retained Earnings / Total Assets). The standard errors are reported in the parentheses. Statistical significance at the 1%, 5% and 10% levels are denoted by ***, ** and *, respectively. (N=406)

1.3 Question 1c

If the EBITTA ratio (Earnings before interest and taxes divided by Total Assets) of a company increases by 1 amount, the probability of this company defaulting will decrease by 0.432, corrected for the other factors.

1.4 Question 1d

$$\textbf{BIC: } ln\left(N^{-1}\sum_{i=1}^{N}\hat{\epsilon}_i^2\right) + \left(\frac{K}{N}\right)ln(N) \tag{2}$$

$$\textbf{AIC: } ln\left(N^{-1}\sum_{i=1}^{N}\hat{\epsilon}_i^2\right) + 2\left(\frac{K}{N}\right) \tag{3}$$

Table 6

Information Criteria

	Model 1	Model 2
AIC	-2.475	-2.482
BIC	-2.445	-2.433

This table contains the AIC and BIC (information criteria) calculated values for model 1 and 2.

Since the BIC penalizes additional parameters stronger compared to the AIC, we will be using the BIC as our model selection value, selecting the most parsimonious model.

In order to obtain the best fitted model, the BIC value should be the lowest of the two model, indicating a better fitting model. In our case this is model 1, concluding that model 1 has a better fit compared to the over fitted model 2.

1.5 Question 1e

Interpretation of the dummy variable for code 13 SIC (Oil Gas Extraction Industry):

If a company is operating in the Oil Gas Extraction industry the probability of defaulting is approximately 0.494 less compared to a company that is operating in the Metal, Mining industry (SIC 10), corrected for the other factors.

Hypotheses

- H_0: $\delta_1 = \delta_1 = ... = \delta_{49} = 0$
- H_A: At least one of the δ_i is not equal to zero.

The appropriate test in this case is the f-test in equation 4. Using this test we are able to test whether or not the coefficient of the all company dummy variables are equal to zero. We will be using a significance level of 5% to determine whether we should reject the null-hypothesis or not. Equation 5 and 6 show that the calculated f_{value} is larger than the f_{crit}, meaning that at least one of the dummy coefficients in model 3 is significantly different from zero. Therefore, we can reject the null-hypothesis and accept the alternative hypothesis, indicating that the probability of default varies across industries based on the SIC codes.

$$F - Statistic = \frac{RSS_0 - RSS_1}{RSS_1}\frac{N - k_1}{k_1 - k_0} \sim F(k_1 - k_0, N - k_1) \tag{4}$$

where the RSS_0, RSS_1, k_0 and k_1 denote the Residual Sum of Squares and the number of parameters in model 2 and 3, respectively.

$$F = \frac{33.10 - 27.20}{27.20}\frac{406 - 54}{54 - 5} = 1.56 \sim F_{5\%(49,352)} \tag{5}$$

$$F_{crit,5\%} = 1.39 \tag{6}$$

Model 3:

$$DEF_i = 0.825 - 0.353EBITTA_i - 0.087MV_i + 0.107WKTA_i - 0.055RETA_i + \sum_{j=2}^{50} \delta_j sic2_{i,j} \tag{7}$$

1.6 Question 1f

Hypothesis Set 1

- H_0: $\beta_5 = 0$
- H_A: $\beta_5 \neq 0$

Hypothesis Set 2

- H_0: $\beta_6 = 0$
- H_A: $\beta_6 \neq 0$

In order to test the proposed hypotheses, we use the student t-test in equation 8 on the coefficients of both interaction variables. The results are shown in equation 9 and 10. Both β_5 and β_6 are not statistically significant, as the T-statistics are lower than the T-crit. Using a significance level of 5%, we are not able to reject the null-hypotheses in both hypothesis set 1 and 2. This means that the effect of reta and wkta, separately, is not significantly different for industries with a sic2 code lower than 40, corrected for the other factors.

$$T - Statistic = \frac{\hat{\beta}_i - \beta_i}{S.E.} \sim T(0, 1, N - k) \tag{8}$$

$$T_{\beta_5} = \frac{-0.1811}{0.1189} = -1.523 \tag{9}$$

$$T_{\beta_6} = \frac{0.0263}{0.0936} = 0.28 \tag{10}$$

$$T_{crit,5\%} = 1.96 \tag{11}$$

$\boldsymbol{\beta_4}$ **:** For a company in an industry sic2 code of higher than 40, if the RETA ratio increases by 1 amount, the probability of defaulting will decrease by 0.145, corrected for the other factors.

$\boldsymbol{\beta_6}$ **:** For a company in an industry sic2 code of lower than 40, if the RETA ratio increases by 1 amount, the probability of defaulting will increase by 0.0263 on top of the effect of a company in an industry sic2 code of higher than 40, corrected for the other factors.

Model 4:

$$\begin{aligned} DEF_i = 0.207 - 0.397 EBITTA_i - 0.061 MV_i + 0.140 WKTA_i \\ -0.145 RETA_i - 0.181 WKTA_i lowSic_i + 0.0263 RETA_i lowSic_i \end{aligned} \tag{12}$$

Table 3

Linear Probability Model: Ratios affecting the default probability (Including SIC interaction dummies)

	Model 1	Model 2	Model 4
ebitta	-0.615***	-0.432**	-0.397**
	(0.174)	(0.187)	(0.190)
mv	-0.0664***	-0.0587***	-0.0613***
	(0.0110)	(0.0116)	(0.0117)
wkta		0.0143	0.140
		(0.0714)	(0.109)
reta		-0.131***	-0.145*
		(0.0497)	(0.0841)
c.wkta#c.lowSic			-0.181
			(0.119)
c.reta#c.lowSic			0.0263
			(0.0936)
Constant	0.214***	0.208***	0.207***
	(0.0201)	(0.0234)	(0.0235)
R-squared	0.142	0.156	0.161
F-stat	33.22	18.56	12.78
R-adjusted	0.137	0.148	0.149

This table presents the regression results of model 1, 2 and 4. This table is an extension of Table 2, with model 4 as addition to the table. Model 4 is the regression of the default probability on the EBITTA (Earnings Before Interest and Taxes / Total Assets), the MV (Market to Book Value), WKTA (Working Capital / Total Assets), RETA (Retained Earnings / Total Assets), the cross product of WKTA and LowSic dummy and the cross product of RETA and the LowSic dummy, where the lowSic dummy equals to 1 for companies in industries with a Sic2 code of lower than 40 and 0 for companies in other industries. The standard errors are reported in the parentheses. Statistical significance at the 1%, 5% and 10% levels are denoted by ***, ** and *, respectively. (N=406)

2 Part 2

2.1 Question 2a

Model 1 and Model 2 are linear probability models, meaning that the estimated relationships are linear. However, the true relationship between (a) continuous variable(s) and a binary dependent variable is nonlinear. Therefore, the drawbacks of estimating a linear relationship is that the marginal effect remains constant over all the values of X when the marginal effect should increase/decrease with respect to X. A second drawback is that it could lead to a probability of greater than 1 or less than 0, because the linear relationship does not have an upper or lower bound, when the true probability cannot be larger than 1 or smaller than 0.

2.2 Question 2b

Table 4

Logistic Regression: Ratios affecting the default probability

	(1)	(2)
VARIABLES	Model 5	Model 6
ebitta	-7.832***	-5.992***
	(2.088)	(2.201)
mv	-0.807***	-0.701***
	(0.144)	(0.148)
wkta_s		-0.00170
		(0.190)
reta		-1.499**
		(0.655)
Constant	-1.448***	-1.527***
	(0.193)	(0.203)
Pseudo R-squared	0.221	0.240

This table presents the logistic regression results of model 5 and 6. Model 5 is the regression of the default probability on the EBITTA (Earnings Before Interest and Taxes / Total Assets) and the MV (Market to Book Value). Model 6 is an extension of model 5, including the standardized WKTA ratio (Working Capital / Total Assets) and the RETA (Retained Earnings / Total Assets). The standard errors are reported in the parentheses. Statistical significance at the 1%, 5% and 10% levels are denoted by ***, ** and *, respectively. (N=406)

Model 6:

$$Pr[DEF_i = 1] = F(Z_i) = \frac{\exp^{Z_i}}{1 + \exp^{Z_i}} \tag{13}$$

where DEF_i denotes the probability of default of company i and Z_i is equal to:

$$Z_i = -1.527 - 5.992EBITTA_i - 0.701MV_i - 0.002WKTA - S_i - 1.499RETA_i \tag{14}$$

2.3 Question 2c

Table 5

Margin Effects: Ratios affecting the default probability

	Model 5 Marginal Effects overall average	Model 5 Marginal Effects at means	Model 6 Marginal Effects overall average	Model 6 Marginal Effects at means
ebitta	-0.605	-0.437	-0.450	-0.322
mv	-0.0624	-0.0451	-0.0526	-0.0377
wkta_s			-0.000127	-9.12e-05
reta			-0.113	-0.0806

This table presents the marginal effects of the logistic regression results of model 5 and 6. Model 5 is the regression of the default probability on the EBITTA (Earnings Before Interest and Taxes / Total Assets) and the MV (Market to Book Value). Model 6 is an extension of model 5, including the standardized WKTA ratio (Working Capital / Total Assets) and the RETA (Retained Earnings / Total Assets). The standard errors are reported in the parentheses. Statistical significance at the 1%, 5% and 10% levels are denoted by ***, ** and *, respectively. (N=406)

Method I: If the Working capital / Total assets ratio increases by 1 standard deviation, then on average the probability of the company defaulting will decrease by 0.01 percentage points.
Method II: For a company with an average Working capital / Total assets ratio, if the Working capital / Total assets ratio increases by 1 standard deviation, then the probability of the company defaulting will decrease by 0.009 percentage points.

The main difference between the two methods is that method 1 estimates all the marginal effects and then calculates the average marginal effect, where as method 2 only takes the average value of X and calculates the marginal effect corresponding with that value. Therefore, this results in values that are not identical.
The main difference between model 5 and 6 is that we see that the coefficients of Model 6 are smaller than the coefficients of Model 5, meaning that the average slope or the slope at the average X values is less negative. This is probably due to the fact that the WKTA-S and the RETA are omitted in Model 5, therefore their effect is incorporated into the coefficients of EBITTA and MV.

2.4 Question 2d

Model 5: We see that the AUROC of the in-sample set(0.8313) differs a little from the AUROC of the out-sample set(0.7681), therefore we conclude that the model is quite robust.
Model 6: We see that the AUROC of the in-sample set(0.8517) differs more from the AUROC of the out-sample set(0.7711), compared to Model 5. Therefore we conclude that Model 5 is more robust than Model 6.

However, when we look at the difference between the AUROC of Model 5 and 6, we see that both the in- and out-sample AUROC is higher in Model 6, indicating that Model 6 is the better model in explaining the Default probability, because it has a higher probability of predicting the right values.
In this case we would prefer Model 6 over Model 5. Even though, Model 6 is less robust, it is more capable in predicting the right values.

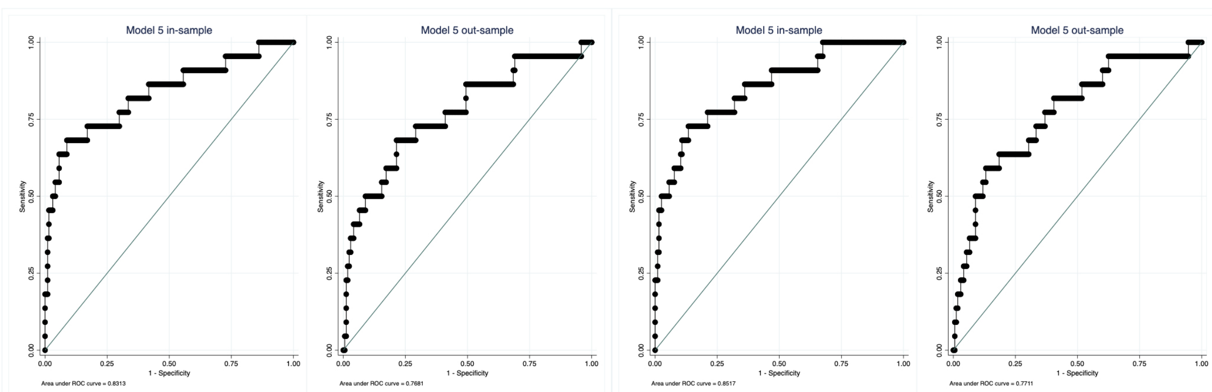

Figure 1: ROC Graphs of Model 5 and Model 6

2.5 Question 2e

According to our analysis the Earnings Before Interest and Taxes, Market to Book Value and Retained Earnings ratios are important factors that could explain probability of defaults. When looking at the logistic regression models, which are superior to the linear probability models, concluded in question 2a, the factors were significant in relation to the probability of default (significance level of 5%). This suggests that an increase in these factors could lead to a decrease in the probability of default. Lastly, the ROC curve of our models shows a value higher than 0.75 for the AUROC which suggests that the logit models are good at predicting true defaults.

```
//Prepare Workspace
clear all
ssc install outreg2
ssc install asdoc

//Change directory
cd "/Users/markdekwaasteniet/Documents/Master Finance/Empirical Finance/Case 1"

//Data structurizing
import excel "/Users/markdekwaasteniet/Documents/Master Finance/Empirical
Finance/Case 1/data_case_I_group_56.xls", sheet("Data") firstrow


// Exercise 1a

//Test the normality of all variables except Date
sktest DEF sic2 mv ebitta wkta reta

mat sktest = r(Utest)

//Subtract the last column of the matrix including the p-values
matrix Pvalues = sktest[.,4...]

//Summarize all the variables except Date
//Sut the variables into a table and convert to matrix
tabstat DEF sic2 mv ebitta wkta reta, statistic(mean sd min max) columns(statistics) save
mat statistics = r(StatTotal)'

//Merge the two matrices together
mat table1 = statistics,Pvalues\Pvalues,statistics
mat table1 = table1[1..6,1...]

//Output the matrix
asdoc wmat, mat(table1) title(Summary Statistics Table 1) dec(3) replace
//Rename Myfile.doc name to Table 1.doc

//Count Missing values in DEF.
egen number_missing = total(missing(DEF))
summarize number_missin
```

```
// Exercise 1b

// Regress the two models. Using outreg2 to create output.
regres DEF ebitta mv
outreg2 using "Table 2.tex", replace word addstat(F-stat, e(F), R-adjusted, e(r2_a))
ctitle(Model 1) title(Ratios affecting the default probability)

regres DEF ebitta mv wkta reta
outreg2 using "Table 2.tex", append word addstat(F-stat, e(F), R-adjusted, e(r2_a))
ctitle(Model 2)




// Exercise 1e

// regress model 2, incorporating dummy variables for SIC code with i.sic2.
regres DEF ebitta mv wkta reta i.sic2





// Exercise 1f

// Generate dummy variable. Siclower is 1 if the SIC code is below 40
gen lowSic = 0
replace lowSic = 1 if sic2 < 40


// Regress with the new dummy variable including interaction variables.
regress DEF ebitta mv wkta reta c.wkta#c.lowSic c.reta#c.lowSic
test c.wkta#c.lowSic c.reta#c.lowSic
outreg2 using "Table 2.tex", append word addstat(F-stat, e(F), R-adjusted, e(r2_a))
ctitle(Model 4) title(Ratios affecting the default probability (Including SIC dummy))
```

```
// Exercise 2b

//Create logit model.
logit DEF ebitta mv
outreg2 using "Table 4.tex", replace addstat(Pseudo R-squared, `e(r2_p)') ctitle(Model 5)

//Standardize wkta
egen wkta_s = std(wkta)
sum wkta_s

//Logit model including standardized wkta and normal reta
logit DEF ebitta mv wkta_s reta
outreg2 using "Table 4.tex", append addstat(Pseudo R-squared, `e(r2_p)') ctitle(Model 6)




// Exercise 2c

// estimate the marginal effects.
logit DEF ebitta mv //model 5
margins, dydx(*) post //average effect of all the x's observations
outreg2 using "Table 5.tex", word replace noaster sideway noparen stats(coef) ctitle(Model
5 - Marginal Effects overall average)
logit DEF ebitta mv //model 5
margins, dydx(*) atmeans post //the effect at the average x observation.
outreg2 using "Table 5.tex", word append noaster sideway noparen stats(coef) ctitle(Model
5 - Marginal Effects atmeans)

logit DEF ebitta mv wkta_s reta  //model 6
margins, dydx(*) post //average effect of all the x's observations
outreg2 using "Table 5.tex", word append noaster sideway noparen stats(coef) ctitle(Model
6 - Marginal Effects overall average)
logit DEF ebitta mv wkta_s reta  //model 6
margins, dydx(*) atmeans post //the effect at the average x observation.
outreg2 using "Table 5.tex", word append noaster sideway noparen stats(coef) ctitle(Model
6 - Marginal Effects atmeans)
```

```
//Exercise 2d

// Making the graphs using lroc and saving them in the CD.
logit DEF ebitta mv if lowSic == 1
lroc if lowSic == 1, title("Model 5 in-sample")
graph save "Graph_model5_insample.gph", replace
lroc if lowSic == 0, title("Model 5 out-sample")
graph save "Graph_model5_outsample.gph", replace
graph combine Graph_model5_insample.gph Graph_model5_outsample.gph
graph save "Graph_model5.gph", replace

logit DEF ebitta mv wkta_s reta if lowSic == 1
lroc if lowSic == 1, title("Model 5 in-sample")
graph save "Graph_model6_insample.gph", replace
lroc if lowSic == 0, title("Model 5 out-sample")
graph save "Graph_model6_outsample.gph", replace
graph combine Graph_model6_insample.gph Graph_model6_outsample.gph
graph save "Graph_model6.gph", replace

graph combine Graph_model5.gph Graph_model6.gph
```