Croatian Operational Research Review
CRORR **8**(2017), 631–652

Logistic regression modelling: procedures and pitfalls in developing and interpreting prediction models

Nataša Šarlija[1], **Ana Bilandžić**[1] **and Marina Stanić**[1,†]

[1]*Faculty of Economics in Osijek, J.J. Strossmayer University of Osijek, Trg Ljudevita Gaja 7, 31000 Osijek, Croatia*
E-mail: ⟨ {*natasa, anag, marina*}*@efos.hr* ⟩

Abstract. This study sheds light on the most common issues related to applying logistic regression in prediction models for company growth. The purpose of the paper is 1) to provide a detailed demonstration of the steps in developing a growth prediction model based on logistic regression analysis, 2) to discuss common pitfalls and methodological errors in developing a model, and 3) to provide solutions and possible ways of overcoming these issues. Special attention is devoted to the question of satisfying logistic regression assumptions, selecting and defining dependent and independent variables, using classification tables and ROC curves, for reporting model strength, interpreting odds ratios as effect measures and evaluating performance of the prediction model.
Development of a logistic regression model in this paper focuses on a prediction model of company growth. The analysis is based on predominantly financial data from a sample of 1471 small and medium-sized Croatian companies active between 2009 and 2014. The financial data is presented in the form of financial ratios divided into nine main groups depicting following areas of business: liquidity, leverage, activity, profitability, research and development, investing and export. The growth prediction model indicates aspects of a business critical for achieving high growth. In that respect, the contribution of this paper is twofold. First, methodological, in terms of pointing out pitfalls and potential solutions in logistic regression modelling, and secondly, theoretical, in terms of identifying factors responsible for high growth of small and medium-sized companies.

Keywords: logistic regression, prediction, firm growth, financial ratios

Received: June 06, 2017; accepted: December 21, 2017; available online: December 30, 2017

DOI: 10.17535/crorr.2017.0041

1. Introduction

Small and medium sized companies with high growth have been recognized as important drivers of employment (Henrekson and Johansson, 2009), as well as drivers of economic and structural change (Hölzl, 2009). High-growth companies

† Corresponding author

are those with annualized growth (in sales, employees or assets) greater than 20% a year over a three-year period (OECD, 2010).

The topic of predicting company growth has received considerable attention from both researchers in economics and entrepreneurship (Davidsson et al, 2010; Delmar, 2006). Aside from the academic world, policy makers and financial institutions are also interested in company growth as it results in the emergence and growth of new value-adding and job-creating businesses (Davidsson and Wiklund, 2000). Finally, growth-oriented managers and entrepreneurs can use these models to evaluate and modify their business activities and strategies.

Company growth is complex and can be observed from multiple perspectives: at the entrepreneur/manager level, company level and environmental level. Theoretically, a plethora of factors are potentially predictors of company growth. On the one hand, these numerous valuable predictors leave space for researchers to test theoretical hypotheses and, without too much difficulty, to develop a prediction model with good predictive power. On the other hand, the complexity of company growth calls for systematic research strongly founded in theory. Otherwise, comparability and interpretability of the model is significantly reduced. Inconsistencies in growth prediction methodologies have led to mixed results, with a lack of understanding of specific methodological differences that potentially hinder theory development (Davidsson and Wiklund, 2000; Weinzimmer, 1998).

The general aim of this study is conduct an empirical investigation of the most common issues concerning prediction modelling of company growth. This paper presents a detailed demonstration of the steps necessary for developing and testing a high-growth prediction model. Furthermore, special attention is directed to common pitfalls and methodological errors in developing the model and suggestions on how to overcome these issues are given. Our research hypothesis is that if company high-growth modelling is done with good theoretical knowledge of growth and statistical methodology with taking care of multicolinearity, overfitting and underfitting on a large data set where error conditions are independent then high-growth model has good performance quality.

The structure of the paper proceeds as follows. The following section provides an overview of previous research on growth prediction. Section 3 is a theoretical and empirical explanation of the logistic regression and focuses on common pitfalls and mistakes, particularly in regard to defining dependent variables, using logistic regression in predicting, the underlying assumptions of logistic regression, variable inclusion, as well as selection and multicollinearity. This section also includes an interpretation of the results of growth prediction model for companies in Croatia, as well as validation of the model. The last section provides a discussion, conclusions and implications for further research.

2. Previous research on growth prediction

Most studies in the field of company growth are oriented toward making theoretical progress. In that context, factors influencing growth potential are usually identified at three levels: the entrepreneur, the company and the environment. With respect to entrepreneur's characteristics, previous studies have singled out a willingness to become involved in situations with uncertain outcomes, mid-management experience, education and the entrepreneur's aspiration to grow to be relevant growth factors (Cassia et al., 2009; Kolvereid and Bullvag, 1996). From a company perspective, a positive relationship exists between the growth potential of a company and R&D investments, innovation capacity and productivity. Additionally, strategical orientation, financial structure, age and size of a company are significant factors in the potential for growth (Mateev and Aatanasov, 2010; Barringer et al., 2005; Freel and Robson, 2004). Finally, the macroeconomic environment and its stakeholders play an important role in facilitating or obstructing the growth of SMEs.

Studies that focus on the methodological aspect of assessing a company's growth are primarily concerned with defining and measuring dependent variables (Shepherd et al., 2009; Janssen, 2009; Weinzimmer, 1998). Conceptualization of the growth variable may be the most frequently discussed topic among scholars. Researchers tend to omit the theoretical justification for selecting a particular method for measuring growth, and interchangeably use different dependent variables (Janssen, 2009). This approach may hinder the process of theory development and lead to inconsistencies in findings and implications for both scholars and policy makers. From a research standpoint, operationalization of measuring growth can vary based on 1) its conceptualized (growth measured as an increase in revenue, assets, employment, capital, added value or market share), 2) the way it is computed, and 3) its complexity (one dimension or composite index). The most frequently used measures of growth relate to increase in sales (revenues), employment and assets. Each measure has its strengths and downsides. Sales seems to be a weak measure in the very early stage of a venture development when assets and employment may very well grow before the actual company starts generating revenue from selling products or services. On the other hand, sales have a high generality as sales increases usually precede an increase in assets and employment, and sales driven by increased demand for a company's products or services reflects the company's level of efficiency and effectiveness (Davidsson et al., 2006). Changes in the number of employees seems to be the best fit for a dependent variable when analyzing company growth from an economics perspective, and the study aims to provide recommendations for policy makers. Employment is not a good measure of company growth when focusing on small and micro businesses given that when such businesses hire an additional single employee it can represent a high percentage increase in employment.

Finally, growth measures related to changes in assets are contentious when studying service industries. Moreover, multicollinearity is a special challenge when modelling asset growth based on financial data. Furthermore, previous studies warn that the practice of using solely absolute measures for growth, such as $(t_f - t_0)$ and $(t_f - t_0)/n$ (where t_f *and* t_0, represent company size during the initial and final period of observation, and n represents the number of periods of observation) can be misleading (Delmar, 2006). These give a distorted picture of real growth as it benefits large companies as opposed to the relative measures, such as $(t_f - t_0)/t_0$, that favor more small companies when calculating growth rates (Weinzimmer et al., 1998).

Regarding the methodology used for modelling company growth, the most frequently used methods are discriminant analysis, logit and probit regressions. Delmar et al. (2003) used correlations and regression analysis to model company growth. Geroski (2005) used static and dynamic optimizing models of company output choice, production functions for modelling corporate learning, models for R&D competition and diversification, and examined their influence on corporate growth rates. Moreno and Casillas (2007) used discriminant analysis to find discriminating variables between high growth and non-high-growth companies. Almus (2002) used probit regression in analyzing factors that influence the probability that a company will achieve fast growth. Probit regression was also used by Arrighetti and Lasagni (2013) in investigating determinants of high growth and calculating the probability of achieving high growth. Sampagnaro and Lavadera (2013) used quantile regression and TOBIT analysis to distinguish between high growth and non-high-growth companies. Mateev and Anastasov (2010) used panel regression analysis in their research in determinants of fast growing SMEs. Heimonen (2012) used logistic regression in his study to identify factors that discriminate between growing innovative SMEs and their non-innovative counterparts.

3. Logistic regression for growth prediction

The method to predict growth depends on the data available and characteristics of the variables used in the modelling process. The main motivation for using logistic regression for growth prediction is the fact that it predicts the probability of a company achieving high growth. It gives insight into variables that are important in predicting growth. Positive and negative values of the estimated coefficients from logistic regression reveal whether a particular variable increases or decreases the probability of high growth. Moreover, interpretation of the odds ratio provides additional information on the degree of impact by the variable.

In general, regression for r independent variables $x_1, x_2, \ldots x_r$ is used to obtain $r \in \mathbb{N}$ coefficients. In logistic regression a dependent variable is binary. In modelling growth prediction, 1 denotes a high-growth company and 0 otherwise. Logistic re-

gression is used to predicts the probability of a company achieving high growth. Since the dependent variable is binary variable, the relationship between the dependent and independent variables is non-linear. The logistic function, which describes this relationship, is of the form:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_r x_r}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_r x_r}} \tag{1}$$

where p is the probability that the dependent variable is equivalent to 1, meaning high growth. Regression coefficients β_i, $i = 1,2,\ldots.,r$, are unknown parameters that need to be estimated. The usual approach to estimating them is logistic transformation:

$$logit(y) = \ln\frac{p}{1-p} = \ln e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_r x_r} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r = g(x) \tag{2}$$

where $\frac{p}{1-p}$ is called odds and its logarithm is called log odds. This transformation provides a linear function, $g(x)$. It too needs to be estimated. Since the assumptions of linear regression such as normality and homoscedasticity are not met, the least square estimation should be avoided. By denoting y_i to be a realization of the dependent variable, and $x_i' = (1, x_{i,1}, \ldots, x_{i,r})$ as observed corresponding to r explanatory variables, where $i = 1, \ldots n$ and n being the sample size, where $p_i = \frac{e^{g(x_i)}}{1+e^{g(x_i)}}$, the entire sample likelihood function conditional on $\boldsymbol{x_i}$ is (Jobson,2012):

$$L(\beta|\boldsymbol{y}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \tag{3}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_r)$. The logarithm is used to obtain a more manageable form:

$$\ln L(\beta|\boldsymbol{y}) = \ln\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \sum_{i=1}^{n}\ln p_i^{y_i}(1-p_i)^{1-y_i} = \sum_{i=1}^{n}[y_i \ln p_i + (1-y_i)\ln(1-p_i)] \tag{4}$$

To estimate β, (4) is maximized through partial differentiation. The solution is obtained using iterative processes (Czepiel, 2002).

Growth modelling has its own specificity, no matter which method is used for developing the model. First, there are lots of variables that can be used as predictors of high growth. This is due to the fact that growth is a complex

phenomenon. There are a multitude of theoretical frameworks focused on explaining growth stages, determinants and future prospects (Davidsson, Wiklund, 2013). It is widely accepted that the growth of a company is most likely nonlinear, temporal and related to the variety of factors that reflect the individual, company, organizational, industry or environmental determinants relating to growth. Second, predictors are often mutually related which means there is a high correlation between them. This is especially true for financial variables. Growth prediction models may also include a range of financial indicators such as lag determinants of a company's performance extracted from the national register of financial statements. Third, the relationship between particular predictors and measure of growth is inconsistent, i.e., the same predictor relating positively to one growth measure may have a negative impact on a different measure of growth. Some studies have confirmed that growth prediction models based on different measures of growth have different set of predictors (Šarlija et al, 2016; Weinzimmer, 1998). Fourth, it can happen that the sample size in relation to the number of candidate variables is not large enough. An example of such a variable is an industry sector which can be shredded. Fifth, the presence of outliers in data. This is quite often the case when financial variables are included in a data set. Sixth, the presence of missing data.

This paper will focus on the more important and common pitfalls and mistakes that can appear when applying logistic regression to model growth prediction. These are: (i) assumptions of logistic regression; (ii) multicollinearity and variable selection; (iii) definition of dependent variable; (iv) interpretation of the results; and (v) use of the receiver operating characteristic curves, so called ROC curve and confusion matrix.

3.1. Assumptions of logistic regression

Logistic regression has much more leeway for method assumptions compared to linear regression. It can handle all sorts of relationships because it applies a non-linear log transformation to the predicted odds ratio. However, there are some assumptions concerning logistic regression that are particularly important in modelling growth (Garson, 2014):

(I) Avoid over fitting and under fitting: Developing a meaningful logistic regression model for growth prediction requires good theoretical knowledge of growth and statistical methodologies. They both go hand in hand.

(II) The error conditions must be independent: Logistic regression requires that each observation be independent. In modelling growth this means that each company should be independently selected in the sample.

(III) The model should have no multicollinearity, meaning that the independent variables should be independent from each other. This is quite a challenge when modelling growth. It will be described in more detail further on.

(IV) Logistic regression requires a large sample size. In modelling growth, there are cases when it is necessary to predict growth for certain data groups, for example, an industry group. For some industry groups, the sample size can be very small and in such cases researchers must be cautious. If the sample size is not large enough, prediction will be biased and it becomes impossible to test the model with standard accuracy measures such as ROC, KS (Kolmogorov–Smirnov statistic) statistics, or t the confusion matrix. Besides an adequate sample size, an adequate count in each cell based on factors should also be satisfied. The presence of small or empty cells may cause the logistic model to be unstable, thus reporting implausibly large b coefficients and odds ratios for independent variables.

3.2. Multicollinearity and selection of variables

Multicollinearity exists when there are correlations between independent variables in the regression model. In growth modelling, researchers always have to handle the multicollinearity problem, especially if interested more in interpretability rather than predictability. Datasets for growth modelling are usually composed of many variables arranged into three main categories that influence growth potential: entrepreneur, company and strategy (Storey, 1994). Variables between groups are usually less correlated than variables in each of the groups. This is particularly noticeable between financial variables. Some financial ratios among predictors have identical numerators, some identical denominators, whereas for others ratios, numerators in ratio are identical to denominators in another ratio. It is obvious that multicollinearity is present in such datasets. When present, it may divert and limit research results and conclusions. Multicollinearity may seriously distort the interpretation of a model. Some of the pitfalls that may appear are (Kutner et al, 2004):
(I) The estimated regression coefficient of one variable depends on other predictors included in the model. Removing or adding just one variable that correlates with other variables may change the significance as well as the sign of regression coefficients.
(II) The precision of the estimated regression coefficients decreases as more predictors are added to the model.
(III) The marginal contribution of one predictor variable in reducing the error sum of squares depends on which other predictors are already in the model.
(IV) Hypothesis tests for $\beta_k = 0$ may yield different conclusions, depending on which predictors are in the model (Mason et al., 1991). Regression coefficients biased by collinearity might lead to variables that demonstrate no significant relationship with the outcome when considered in isolation. However, they do become highly significant in conjunction with collinear variables, yielding an elevated risk of false-positive results (Type I error). Alternatively, multiple

regression coefficients might show no statistical significance due to incorrectly estimated wide confidence intervals, yielding an elevated risk of false-negative results (Type II error) (Tu, 2005).

Multicollinearity is directly related to variable inclusion and selection, which might be the most important issue when modelling growth. There are lots of variables relating to an individual, company, organizational, industry and environmental determinants of growth. A major problem when building a logistic model for growth prediction is which variable to select and include. There are several modelling strategies that can be used. First, collect as many variables as possible and then insert all of them into the modelling procedure to find something that has significance. This is not a good approach for growth modelling as there can be many variables, and in such cases either fraudulent result may appear or results that cannot be interpreted and not related to growth theory. Second, start from growth theory and combine it with different selection procedures available in statistical software. For example, this approach begins by selecting all the important variables available in the dataset, and then drops them out one by one, preferably the less significant ones. This can be done manually by the researcher or using computer-assisted selection procedures. It starts with putting all variables into the model, and leaving out the one with the highest p-value. This step is repeated until the desired number of variables remains in the model (Bursac at al., 2008). There is another procedure similar to the previous, but it starts with choosing one variable with the lowest p-value and adding it the model. Variables are added one by one, each with the lowest p-value, until the desired number of variables is reached (Bursac at al., 2008).

When there is a limited sample size in relation to the number of candidate variables, a pre-selection should be performed. One way to do this is by developing models with just one explanatory variable at a time, and afterwards include in the multivariate model all variables that exhibit a relaxed p-value (for instance, $p \leq 0.25$). This relaxed p-value criterion allows reducing the initial number of variables in the model, thus reducing the risk of missing important variables (Sperandei, 2014).

In our research, the independent variables consisted of financial ratios and control variables as well as variables related to R&D, investment and export which is especially important for high-growth companies. To get the financial variables, in cooperation with the Financial Agency (FINA) in Croatia, we collected financial statements (balance sheets and income statements) of all micro, small and medium companies in Croatia over the period 2009-2014. In all, 36 variables were placed into 9 groups: (i) R&D; (ii) investment; (iii) liquidity; (iv) export; (v) productivity; (vi) capital structure; (vii) profitability; (viii) turnover ratios, (ix) control variables. According to Croatian regulations, a micro company has less than 10 employees, with annual sales and total annual balance sheet not exceeding 2 million EUR. A small company is defined as a legal entity which employs from

11 to 50 employees annually, with total annual income and total annual balance sheet not exceeding 10 million EUR. Finally, the category of medium-sized enterprises is made up of legal entities which employ from 51 to 250 persons annually, with annual sales not exceeding 50 million EUR and total annual balance sheet not exceeding 43 million EUR.

Data set for our research consisted of 1471 small and medium-sized Croatian companies active between 2009 and 2014. The sample size used for development and validation purposes is shown in Table 1. As it can be seen, in both subsets the ratio of high growth vs non-high growth companies is 50:50. Total number of companies that were active was 56596 and the total number of companies with the high growth in sales was 800. The same number of non high-growth companies are randomly selected. The numbers in the subsamples were reduced to 745 and 726 after controlling for outliers. Sampling was done with R Revolution, using its function sample(). The ratio of dividing the total sample into development and validation subsets is 80:20.

	Development subset of the sample			Validation subset of the sample		
Sample	High growth	Non-high growth	Total in the development sample	High growth	Non-high growth	Total in the validation sample
SMEs	616	632	1248	113	110	223

Table 1: *Sample sizes for development and validation*

Descriptive statistics for the developing dataset for high-growth and non-high-growth companies can be found in the Table 2 and Table 3.

Variable code	Description of variable	High-growth	Non-high growth
		Mean (st.dev.)	Mean (st.dev.)
Research and development			
NematImovina**	intangible assets/total assets	0.0079 (0.0281)	0.0054 (0.0199)
ExpdnTA	R&D/total assets	0.00002 (0.0004)	0.00001 (0.0003)
CPLTA	concessions, patents, licenses, trademarks, software /total assets	0.0046 (0.0259)	0.0053 (0.0255)

GWTA	goodwill/ total assets	0.0005 (0.0156)	0.00006 (0.0015)
Investments			
InvLATA	investment in long-term assets/ total assets	0.0135 (0.0355)	0.0142 (0.0343)
InvLATE	investment in long-term assets / total expenditures	0.0147 (0.040)	0.0154 (0.0405)
Liquidity ratios			
Kimovkobv***	current assets/current liabilities	1.4292 (1.6159)	1.824 (1.7830)
Lubrl***	(current assets-inventory)/current liabilities	1.2081 (1.6324)	1.4998 (1.7321)
Lkiui	current assets/total assets	0.6376 (0.3210)	0.644 (0.3194)
crenl	cash/current liabilities	0.3038 (0.5718)	0.3441 (0.5777)
Export			
prihosal	domestic sales/ total sales	0.9217 (0.2377)	0.9448 (0.1980)
prihlsal	export/ total sales	0.0144 (0.0620)	0.0109 (0.0477)
implata	import/ total assets	0.0055 (0.0241)	0.0051 (0.02083)
implate	import/ total expenditures	0.0053 (0.0239)	0.0064 (0.0249)
Productivity			
Pprihzapos***	sales/number of employees	305453.66 (257511.05)	420584.51 (358858.12)
Turnover ratios			
Aukupni	total revenue/total assets	1.4259 (1.5781)	1.5918 (1.7126)
Adug**	total revenue /fixed assets	5.0838 (5.9609)	5.8549 (5.3089)
Akrat**	total revenue /current assets	2.6471 (3.7385)	3.1312 (3.0614)
Asalta	sales/total assets	1.3754 (1.5497)	1.5210 (1.9597)
Asalwc	sales/net working capital	0.6561 (7.4561)	1.2349 (7.3279)

Anap1*	365/receivables turnover	92.96 (88.70)	83.46 (107.70)
Akrdob	365/payables turnover	96.06 (89.50)	88.46 (100.87)
Capital structure			
Zkz***	total debt/total assets	1.3792 (3.0533)	0.86328 (1.7419)
Zdk**	total debt/equity	2.1929 (8.7848)	1.2257 (8.4439)
Blta***	bank loan/total assets	0.0570 (0.1257)	0.0818 (0.1510)
Lclnw*	current liabilities/equity	1.6047 (5.8024)	1.0739 (5.5984)
Prearnta	retained earnings/total assets	-0.3622 (1.4885)	-0.0431 (0.8453)
zlongdca	long-term liabilities/short-term assets	0.7236 (2.3307)	0.5429 (1.7401)
Profitability ratios			
Rosd	net income/sales	0.1031 (0.4028)	0.0687 (0.1375)
pnmdg	(net income or loss/ total revenue)*100 (%)	2.4592 (12.7822)	3.2453 (12.4058)
pnroadg	(net income or loss/ total assets) *100	1.5879 (14.734)	3.2516 (13.0139)
pnroed	net income/equity (%)	6.7937 (12.6929)	7.6172 (11.7660)

* statistically significant at 10%
** statistically significant at 5%
*** statistically significant at 1%

Table 2: *Descriptive statistics and t-test for differences in means for independent variables*

As it can be noticed from Table 2, high-growth companies compared to non high-growth companies have higher mean values of intangible assets and leverage while lower value is present in current and quick ratios, turnover of fixed and short term assets as well as in ratio of bank loans to total assets and productivity.

Variable code	Description of variable		High-growth	Non-high growth
			Mean (st.dev.)	Mean (st.dev.)
Control variables				
Age***	>= 7 years old		47.61%	52.39%
	<7 years old		59.75%	40.25%
Size***	Micro		52.38%	47.62%
	Small		35.26%	64.74%
	Medium		10%	90%
High tech	Low tech industry		50.11%	49.89%
	High tech industry		50.93%	49.07%
IND*	Industry sector	Agriculture	47.62%	52.38%
		Manufacturing	49.50%	50.50%
		Construction	52.20%	47.80%
		Trade	46.47%	53.53%
		Transportation and storage	52.70%	47.30%
		Accommodation and food service	55.71%	44.29%
		Information and communication	66.67%	33.33%
		Financial activities	49.09%	50.91%
		Professional and scientific services	46.38%	53.62%
		Social, education and other services	53.57%	46.43%

* statistically significant at 10%
** statistically significant at 5%
*** statistically significant at 1%

Table 3: *Percentages of high-growth and non high-growth companies and chi-square test of independence for control variables*

It is shown in Table 3 that high-growth companies are younger and smaller. The highest percentage of high-growth companies is from ICT sector.

3.3. Definition of the dependent variable

Growth can be measured in various ways, depending on the particular business's focus: revenue generation, assets or physical output expansion, employment boost and market share increase. Sales is taken to be the best measure of growth according to the most researchers (Davidsson and Wiklund, 2000). Except during the

very early start-up phase of venture development, when is possible for assets and employment to grow before any sales occur (Delmar et al. 2003), sales volume is the most common performance indicator used by entrepreneurs and business owners (Barkham et al., 2002). Growth can be measured as an absolute and relative measure. In both cases, growth measure is sensitive to a company's initial size. There is a positive association between a company's initial size and absolute growth, and negative association between a company's initial size and relative growth rate (Weinzimmer et al., 1998). To overcome this issue, the recommenddation is to use initial size as a control variable.

The main advantage of using logistic regression for growth modelling is its ability to predict the likelihood of a company becoming achieving high growth. Linear regression can also be used for growth prediction, but it does not give a probability but prediction of the dependent variable in the same measuring unit as the variable. In linear regression models, a set of input variables is used to predict a continuous response variable. In the logistic regression, the dependent variable is an indicator variable, whereas it is a continuous variable in linear regression. When modelling growth with linear regression, the dependent variable could, for example, be a growth percentage from one year to next. In that case, the growth percentage of a company is based on a set of predictors. In contrast to that, when modelling growth using logistic regression, the dependent variable is binomial, indicating whether a company is high growth or non-high growth. In that case, the probability that a company will achieve high growth, as defined above, is obtained.

Using logistic regression for growth prediction requires careful definition of the dependent variable. This is usually done such that the continuous variable, for example sales or assets, is converted to a dummy variable, as was done in our research. Since logistic regression assumes that P(Y=1) is the probability of the event occurring, the dependent variable should be coded accordingly. That is, for a binary regression, the factor level 1 of the dependent variable should represent the desired outcome, in our case high growth. When defining growth as an indicator variable, a better distinction between high-growth and non-high-growth companies is achieved, and consequently better identifying the variables that influence high growth. On the other hand, since linear regression has a dependent continuous variable, linear modelling will probably provide a more precise model.

Another important issue in defining variables is the time frame that is covered. Using logistic regression for predicting growth leads to a time lag in the dataset. Based on the data set from year t, the prediction is done for $t+1$ or $t+2$, and depends on the time lag.

In our research, the dependent variable covers the period from 2011 to 2014 and the independent variables comprise the financial ratios from 2011. The dependent variable is binomial, indicating a company is experiencing high growth, if it has

annualized sales growth greater than 20% a year over a three-year period from 2011 to 2014. Otherwise, a company is defined as non-high-growth. (OECD, 2010).

3.4. Interpretation of the results

The Logistic regression model for high growth prediction developed in our research is presented in Table 3. The model was developed keeping in mind four important issues:
(I) Variable inclusion and selection. First, bivariate analysis is performed, then several different combinations of variables as well as selection procedures are tested keeping in mind growth theory and previous research results.
(II) Assumptions of logistic regression.
(III) Multicollinearity issue.
(IV) Interpretability of the models. The model is in line with the theory and as such is easier to interpret. In analyzing the companies' growth, both issues are relevant - interpretation and prediction. Whether it involves a researcher or an entrepreneur, both want to know which factors are relevant for a company's growth rather than just predicting which company will grow and which one will not.

Variable code	Variable	Regression coefficient	Lower CL Upper CL 95% - 95%	p-value
Research and development				
NematImovina	intangible assets/total assets	5.887	0.495 11.278	0.0324
Export				
prihosal	domestic sales/ total sales	-0.672	-1.267 -0.076	0.0269
Productivity				
Pprihzapos	sales/number of employees	-0.0001	-0.0001 0.0000	<0.0001
Capital structure				
Zdk	total debt/equity	0.022	0.007 0.036	0.0023
Blta	bank loan/total assets	-0.936	-1.825 -0.047	0.0391

Prearnta	retained earnings/total assets	-0.229	-0.357 -0.101	0.0004
Profitability ratios				
Rosd	net income/sales	0.493	-0.045 1.032	0.0727
Control variables				
Age	>= 7 years old	-0.554	-0.879 -0.229	0.0008
	<7 years old	0.000		
Size	Micro	1.832	0.329 3.334	0.0168
	Small	1.247	-0.284 2.779	0.1104
	Medium	0.000		
High-tech	High tech industry	0.224	-0.020 0.468	0.0721
	Low tech industry	0.000		
				AIC=1612.41; R^2=0.139

Table 3: *The logistic regression model for high growth prediction*

One of the issues with logistic regression is interpreting it. Unlike linear regression, it is not obvious what impact independent variables have on the dependent variable. There is more than one approach for this. One way is to observe signs and absolute values of the regression coefficients. If the coefficient sign is positive for a specific independent variable, then an increase in the variable value will result in a higher probability for the dependent variable to have a positive outcome. In our model, what is noticeable is that the probability of achieving high growth increases with a decrease in bank loans over total assets, retained earnings over total assets and revenue per employee and with an increase in profit margin, share of intangible assets in total assets and total debt over total assets. Also, a higher probability for growth is attributed to companies that are export-oriented, younger, smaller and adopt high technology. In regards to the absolute value, it is used to get a sense of which variables have a bigger impact. This is done with care, because the range of values for the variables can vary greatly. Another way is to use equation (1) for computing probabilities by changing only one variable by 1, but the problem here is that this equation is non-linear, it has an S-shaped (or conversely S-shaped) graph, so the difference in probability on the lower and upper end of the range of values is quite small compared to the remainder. A third way is using log odds ratios for interpreting. The log odds ratio is a natural logarithm of odds ratio. By increasing the net income/sales from x to $x + 1$, and

keeping all other variables at the same value c, the log odds ratio will increase by 0.493, i.e. according to equation (2) $\ln\frac{p_2}{1-p_2} - \ln\frac{p_1}{1-p_1} = c + 0.493(x+1) - (c + 0.493x) = 0.493$. Equivalent to that, odds ratio will increase as well, but the initial value should be multiplied by $e^{0.493} = 1.637$ to calculate the new, i.e. it's an increase by $\frac{p_2}{1-p_2} - \frac{p_1}{1-p_1} = e^{c+0.493(x+1)} - e^{c+0.493x} = (e^{0.493} - 1)e^{c+0.493x} = 0.637 * e^{c+0.493x}$.For net income/sales it corresponds to an increase in the odds of a company becoming high growth from 0.5 to 0.818, or from 2 to 3.274. Regarding the rest of the numerical variables, by increasing intangible assets/total assets by 1, the log odds ratio will increase by 5.887. Log odds ratio will also be increased by 0.022 with the increase of total debt over equity ratio by 1. As opposed to that, by increasing sales/number of employees, domestic sales/total sales, bank loan/total assets and retained earnings/total assets by 1, the log odds ratios will be decreased as follow by 0.0001, 0.672, 0.936 and 0.229.

In the case of the categorical variable:

$$\frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{e^{\beta_0+\beta_{1a}}}{e^{\beta_0+\beta_{1b}}} = e^{\beta_0+\beta_{1a}-(\beta_0+\beta_{1b})} = e^{\beta_{1a}-\beta_{1b}} \tag{5}$$

where β_{1a} and β_{1b} are the regression coefficients of the categories *a* and *b* for a categorical variable. Hence, for our model with all variables fixed except for the variable *size*, if a company changes its status from a small company to a micro company it will increase its log odds ratio by 1.832-1.247=0.585, or increase its odds ratio by $e^{0.585} = 1.795$, which corresponds to an increase in the odds of a company becoming high growth from 0.5 to 0.8975, or from 2 to 3.589.

If a categorical variable has only two categories, the odds ratio for a change from the base category to the remaining category is $e^{\beta_0+\beta_{1a}-\beta_0} = e^{\beta_{1a}}$, with all other variables being fixed. In our model 'Age' and 'Hihg-tech" are that type of categorical variables. So, the log odds ratio for a company to become high growth that is 7 years old, or older, as composed to a younger company is lower by 0.554, or the odds ratio is lower by the multiplier $e^{-0.554} = 0.575$. This corresponds to a decrease in the odds of a company becoming high growth from 0.5 to 0.2875, or from 2 to 1.15. Equivalent to that, the log odds ratio of a company to become high growth if it's in a high tech industry, as composed to a company that is not, and every other variable having the same value, is higher by 0.224, or the odds ratio is higher for the multiplier $e^{0.224} = 1.251$. Accordingly, this coincides with the increase from 0.5 to 0.6255, or from 2 to 2.502.

Odds ratios are equal to e^{β_1}, and the obtained regression coefficient β_1 corresponds to the log odds ratio. Or more intuitively, the log odds ratio gives the additive effect on the logit, while the odds ratio gives the multiplicative effect. (Mood, 2010; Gelman and Hill, 2007).

3.5. Usage of ROC curve and confusion matrix

Once the model has been developed, it should be tested. Usually, the entire dataset is divided into two subsets: the train sample and the test sample. The train sample is used to develop the model and the test sample to test how well the model works. Some standard measures used in testing logistic regression models are KS statistic, ROC curve and confusion matrix.

To calculate KS statistics, the following notation is used: m_1 is the number of high-growth companies, m_2 is the number of non-high-growth companies, I is the indicator function (1 if all its conditions are met, and 0 otherwise) and s_i is score of the $i-th$ client.

$F_{m_2,BAD}(a)$ and $F_{m_1,GOOD}(a)$ are defined as:

$$F_{m_1,GOOD}(a) = \frac{1}{m_1}\sum_{i=1}^{m_1} I(s_i \leq a \wedge y_i = 1) \tag{6}$$

$$F_{m_2,BAD}(a) = \frac{1}{m_2}\sum_{i=1}^{m_2} I(s_i \leq a \wedge y_i = 0) \tag{7}$$

Intuitively $F_{m_1,GOOD}(a)$ is a function that divides the number of correctly predicted high-growth companies by the number of all truly high-growth companies.

Equivalent to that $F_{m_2BAD}(a)$ is a function that divides the number of correctly predicted non-high-growth companies by the number of all truly non-high-growth companies.

The KS statistic has the following shape:

$$KS = \max_{a\in[L,H]} \left|F_{m_2,BAD}(a) - F_{m_1,GOOD}(a)\right| \tag{8}$$

where L and H are the minimum and maximum score values of the observed model, respectively.

The confusion matrix is a two-by-two matrix with the categories: (i) true positives (TP) - entries correctly labeled as positives, (ii) false positives (FP) - negative entries incorrectly labeled as positive (iii) true negatives (TN) – are negatives correctly labeled as negative, (iv) false negatives (FN) refer to positive examples incorrectly labeled as negative. *tp rate* and *fp rate* (11) is usually calculated from the confusion matrix.

The ROC curve or receiver operating characteristic curve is one of the most popular measures of the quality of a model. It is a visual measure, so more often the area under the ROC curve (AUC) will be used. Its value ranges from 0.5 to 1. The higher the value, the better the model, an AUC of 1 indicates a perfect

model, one that has classified every entry correctly. The ROC curve is based on a measure of the true positive rate and the false positive rate:

$$tp\ rate = \frac{\text{Positives correctly classified}}{\text{Total positives}} \tag{9}$$

$$fp\ rate = \frac{\text{Negatives incorrectly classified}}{\text{Total negatives}} \tag{10}$$

for all possible cut-offs. The curve is obtained by plotting $tp\ rate$ on the y axis and $fp\ rate$ on the x axis. The more the curve is concave, the better model, with the area under the ROC curve ranging from 0.5 to 1 (Fawcett, 2006).

ROC curve for our model is shown in Figure 1.

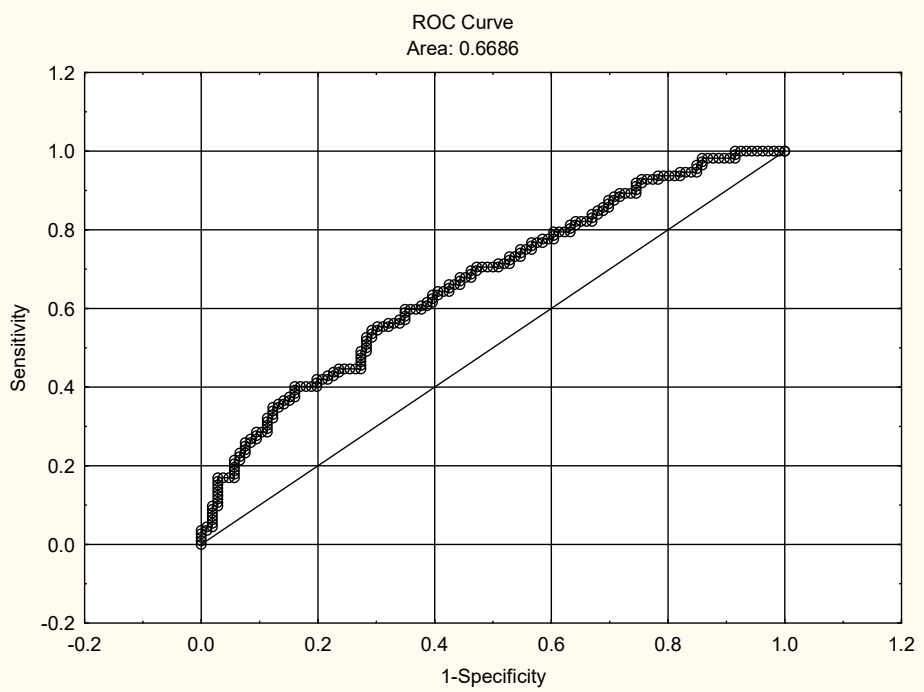

Figure 1: *ROC curve of the logistic regression model for high growth prediction*

As it is shown in Figure 1., area under ROC curve is 0.6686 indicating good classification quality.

It is quite common that researchers make mistakes in setting the cut-off value when applying logistic regression in growth modelling. When a dependent variable is binary, the baseline for setting the cut-off is not 1/L, where L is the number of

levels of the dependent variable, but the proportion of category 1 in the whole dataset. In our research, the number of high-growth companies in the sample is 616, so the cut-off is set to 0.494 (616/1248).
The confusion matrix for our model is presented in the Table 4. Because of missing values it occurs that for some entities it is not possible to predict if they will become high-growth companies, so the total number of predicted values may vary depending on variables which consist the model.

Actual	Predicted		Total
	High growth	Non-high growth	
High growth	70	42	112
Non high growth	42	64	106
Total	112	106	218

Table 4: *Confusion matrix*

Hit rates can be calculated from the confusion matrix: high-growth hit rate = 62.5%, non-high-growth hit rate = 60.38% and total hit rate=61.47%.

4. Discussion and conclusion

A company's growth is generally considered a key driver of economic development, and as such a dominant factor in the creation of new jobs. Company growth, especially those with high growth are elitist and minority and is a dominant factor in creating new jobs. In the UK, 4% of the fastest growing companies generated 50% of the new jobs, and in the USA, 3% are responsible for 70% of the new jobs (Lilischkis, 2011). These are the reasons why growth modelling has recently become so important. If we know what the growth determinants are, we can influence them and facilitate growth. Growth modelling is not an easy task and if done in the wrong way, it may lead to false conclusions. This was the motivation for our paper - describing the steps needed to develop and test a growth prediction model based on logistic regression with the special attention on common pitfalls and methodological errors when developing the model. The most common mistakes in applying logistic regression for growth modelling relate to the selection of dependent variable where the theoretical justification for the selection is not given, this is followed by including and selecting independent variables where usually many variables are inserted into the model without adhering to growth theory. It leads to other two mistakes - multicollinearity and overfitting. Furthermore, using samples that are not large enough for applying logistic regression leads to a prediction that is probably biased. Setting the cut-off is another common mistake which is usually set to 0.5, no matter of the sample structure. Finally, interpretation of the logistic regression coefficients is not straightforward

as is the case with linear regression, and this then leads to mistakes. The paper also provides a logistic regression model for predicting high growth in small and medium-sized companies in Croatia. It has been shown that growth determinants for Croatian companies are: bank loans over total assets, retained earnings over total assets, revenue per employee - the lower the better, as well as profit margin, share of intangible assets in total assets, total debt over total assets - the higher the better. Moreover, higher probability for is attributed to companies that export goods and services, younger, smaller and highly technological companies. The paper shows that if company high-growth modelling is done with good theoretical knowledge of growth and statistical methodology with taking care of multicolinearity, overfitting and underfitting on a large data set where error conditions are independent then high-growth model has good performance quality. As a guideline for further research, we suggest describing procedures and pitfalls for other methods used in growth modelling such as linear regression, panel data regression, neural networks and support vector machines.

Acknowledgement

This study is funded by Croatian Science Foundation under Grant No.3933 "Development and application of growth potential prediction models for SMEs in Croatia".

References

[1] Almus, M. (2002). What characterizes a fast-growing firm?. Applied Economics, 34(12), 1497–1508.

[2] Arrighetti, A., and Lasagni, A. (2013). Assessing the determinants of high-growth manufacturing firms in Italy. International Journal of the Economics of Business, 20(2), 245–267.

[3] Barringer, B. R., Jones, F. F., and Neubaum, D. O. (2005). A quantitative content analysis of the characteristics of rapid-growth firms and their founders. Journal of Business Venturing, 20(5), 663–687.

[4] Barkham, R., Gudgin, G., and Hanvey, E. (2002). Determinants of small firm growth: An inter-regional study in the United Kingdom 1986-90 (Vol. 12). Psychology Press, London.

[5] Bilandžić, A., Jeger, M. and Šarlija, N. (2015). How to Improve Interpretability of the Logistic Regression Model. Proceedings of the 13th International Symposium on Operational Research SOR '15, Bled, Slovenia, 279–284.

[6] Bursac, Z., Gauss, C. H., Williams, D. K. and Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine, 17(3), 1–8.

[7] Cassia, L., Cogliati, G. M. and Paleari, S. (2009). Hyper-Growth Among European SMEs: An Explorative Study. Available at: https://ssrn.com/abstract=1389521 [Accessed 12/05/16]
[8] Czepiel, S. A. (2002). Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. Available at: http://ww.saedsayad.com/docs/mlelr.pdf [Accessed 23/03/16]
[9] Davidsson, P., Achtenhagen, L., and Naldi, L. (2010). Small firm growth. Foundations and Trends in Entrepreneurship, 6(2), 69–166.
[10] Davidsson, P., & Wiklund, J. (2006). Conceptual and empirical challenges in the study of firm growth. Entrepreneurship and the Growth of Firms, 1(1), 39–61.
[11] Davidsson, P., & Wiklund, J. (2013). New Perspectives on Firm Growth, Edward Elgar, Cheltenham, Northapton.
[12] Delmar, F. (2006). Measuring growth: methodological considerations and empirical results. Entrepreneurship and the Growth of Firms, 1(1), 62–84.
[13] Delmar, F., Davidsson, P. and Gartner, W. B. (2003). Arriving at the high-growth firm. Journal of business venturing, 18(2), 189–216.
[14] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.
[15] Freel, M. S., and Robson, P. J. (2004). Small firm innovation, growth and performance: Evidence from Scotland and Northern England. International Small Business Journal, 22(6), 561–575.
[16] Garson, G. D. (2016). Logistic Regression: Binomial and Multinomial, Edition. Asheboro, NC: Statistical Associates Publishers.
[17] Gelman, A. and Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York.
[18] Geroski, P. A. (2005). Understanding the implications of empirical work on corporate growth rates. Managerial and Decision Economics, 26(2), 129–138.
[19] Heimonen, T. (2012). What are the factors that affect innovation in growing SMEs?. European Journal of Innovation Management, 15(1), 122–144.
[20] Henrekson, M., & Johansson, D. (2010). Gazelles as job creators: a survey and interpretation of the evidence. Small Business Economics, 35(2), 227–244.
[21] Hölzl, W. (2009). Is the R&D behaviour of fast-growing SMEs different? Evidence from CIS III data for 16 countries. Small Business Economics, 33(1), 59–75.
[22] Janssen, F. (2009). The conceptualisation of growth are employment and turnover interchangeable criteria?. Journal of Entrepreneurship, 18(1), 21–45.
[23] Jobson, J. (2012). Applied multivariate data analysis: volume II: Categorical and Multivariate Methods. Springer Science & Business Media, New York.
[24] Kolvereid, L. and Bullvag, E. (1996). Growth intentions and actual growth: The impact of entrepreneurial choice. Journal of Enterprising Culture, 4(01), 1–17.

[25] Kutner, M. H., Nachtsheim, C., & Neter, J. (2004). Applied Linear Regression Models. McGraw-Hill/Irwin.
[26] Lilischkis, S. (2011). Policies in Support of High Growth Innovative SMEs: Inno-grips Policy Brief No. 2. European Comission. Enterprise and Industry.
[27] Mason, C. H., & Perreault Jr, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. Journal of Marketing Research, 268–280.
[28] Mateev, M., & Anastasov, Y. (2010). Determinants of small and medium sized fast growing enterprises in Central and Eastern Europe: a panel data analysis. Financial Theory and Practice, 34(3), 269–295.
[29] Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. European Sociological Review, 26(1), 67–82.
[30] Moreno, A. M., & Casillas, J. C. (2007). High-growth SMEs versus non-high-growth SMEs: a discriminant analysis. Entrepreneurship and Regional Development, 19(1), 69–88.
[31] Organisation for Economic Co-operation and Development. (2010). High-Growth Enterprises: What Governments Can Do to Make a Difference, OECD Studies on SMEs and Entrepreneurship, OECD Publishing. Available at: http://dx.doi.org/10.1787/9789264048782-en
[32] Sampagnaro, G., & Lubrano Lavadera, G. (2013). Identifying High Growth SMEs Through Balance Sheet Ratios. Available at: http://ssrn.com.secure.sci-hub.cc/abstract=2207550 [Accessed 26/04/16]
[33] Shepherd, D., and Wiklund, J. (2009). Are we comparing apples with apples or apples with oranges? Appropriateness of knowledge accumulation across growth studies. Entrepreneurship Theory and Practice, 33(1), 105–123.
[34] Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1), 12–18.
[35] Storey, D. (1994) Understanding the Small Firm Sector. Routiedge, London.
[36] Šarlija, N., Pfeifer, S., Jeger, M. and Bilandžić, A. (2016). Measuring enterprise growth: pitfalls and implications. World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, 10(6), 1773–1780.
[37] Tu, Y. K., Kellett, M., Clerehugh, V., & Gilthorpe, M. S. (2005). Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. British Dental Journal, 199(7), 457–461.
[38] Weinzimmer, L. G., Nystrom, P. C. and Freeman, S. J. (1998). Measuring organizational growth: Issues, consequences and guidelines. Journal of Management, 24(2), 235–262.